THE UNIVERSITY *of* EDINBURGH
School of Physics
and Astronomy

Reducing the Biases in Machine Learning Algorithms for Higgs Physics

MPhys Project Report

Stefan Katsarov

Submitted for the 40pt MPhys Project course PHYS11016
August 6, 2023

Abstract

This report examines the reduction of classification uncertainty for Higgs bosons produced in vector boson fusion (VBF) which decay via the diphoton channel. ATLAS reports a high uncertainty in the measured standard model (SM) Higgs to vector boson coupling strengths. An adversarial neural network (ANN) is developed to address VBF Higgs background modelling uncertainty. The ANN achieves higher signal purity over the ATLAS implementation through a 30% decrease in background efficiency. Systematic uncertainties between different simulation parameters are studied. The sole classification between VBF and gluon-gluon fusion Higgs production modes is excluded as the direct source of this uncertainty. These results provide a means for reducing the Higgs to vector boson coupling uncertainty, allowing for a more stringent test of the SM.

Supervisor: Dr Liza Mijovic

Personal statement

In the first two weeks of the project, I read background material to familiarise myself with the current Higgs boson classification procedure taken by ATLAS and the biases which are present. My focus was placed on the classification between the different Higgs boson production modes. I read the most recent collaboration publications and supplemented this with doctoral publications containing more detailed explanations.

In the following month, once I began to understand the key principles in the ATLAS $H \to \gamma\gamma$ analysis I started to get some hands-on experience. I was provided with simulated detector data from Higgs boson events produced via the gluon-gluon fusion (ggF) and vector boson fusion (VBF) production modes. I picked out the most useful variables and made manual cuts to separate ggF and VBF events. I used this to achieve a baseline performance which I referred to while implementing a neural network to perform the same task. This allowed me to get a basic classification procedure going and gave me useful insights.

Throughout the first semester, along with the hands-on work, I was reading about how to implement the same classification features and training procedure as the original ATLAS analysis. I also looked at how the systematic uncertainty from simulation parameters is determined based on different parton shower models. I got more familiar with the PYTHIA and HERWIG shower models and the physics between the ggF and VBF production modes.

In the last month of semester one, I was focused on reproducing the systematic uncertainty reported by ATLAS. I sequentially generated the engineered event variables used in the ALTAS analysis, starting with the most simple first to get preliminary results. More complex features required the use of a 4-vector computation package called pylorentz. These features took significantly longer time to generate and generating them on my laptop became unfeasible. To bypass the overhead required with remote computing resources, only 10% of the dataset was initially used to obtain quicker results.

During the winter break, I was keen on seeing the results which would be produced when all of the features were used in training. I obtained results by testing between different samples showered with PYTHIA and HERWIG to determine the systematic uncertainty. However, these results could not reproduce the ATLAS uncertainty. Given this, I invested some time working with the remote computing resources to generate the full training set. However, using the full training set still did not reproduce the uncertainty reported by ATLAS.

In the first two weeks of semester two, I wanted to make my results more conclusive. I looked to see if there were trends in systematic uncertainty with classification performance, however, none were found. This allowed me to conclude this study and turn my focus to another related problem. I began to look into the cause of a bias in VBF classification that arises during background rejection.

In the next two weeks, I read through previous studies which used an adversarial neural network (ANN) to address a biased selection of background events which sculpts the background distribution to look like signal. I could make use of the code in my previous study

to generate the new training features which allowed me to make quick progress. I was also provided with the ANN code used in the previous studies to help with the implementation. After reading how ANNs are used I adapted the previous code and applied it to my study to get promising results.

In the last month, I placed most of my focus on writing the report and spent my remaining time developing a hyperparameter optimisation procedure for the ANN. This procedure gave clear results displaying the ANN's superior performance over the current ATLAS results. The hyperparameter optimisation provided strong indications for additional improvements which I would be very keen to delve into if more time were available.

Acknowledgments

I want to give my sincere thanks to Dr Liza Mijovic for her thorough and constant support throughout the project. I am grateful for all her invaluable guidance and to have been provided with this amazing opportunity. I really enjoyed all of the stimulating discussions we had and learned a lot throughout.

Contents

1 Introduction

1.1 Motivation

The discovery of the Higgs boson at the ATLAS and CMS experiments in 2012 [1,2] marked an important milestone for the standard model (SM) of particle physics. The Higgs boson was the last confirmation of the series of predictions made by the SM where it consolidated the validity of our most accomplished theory of particle physics.

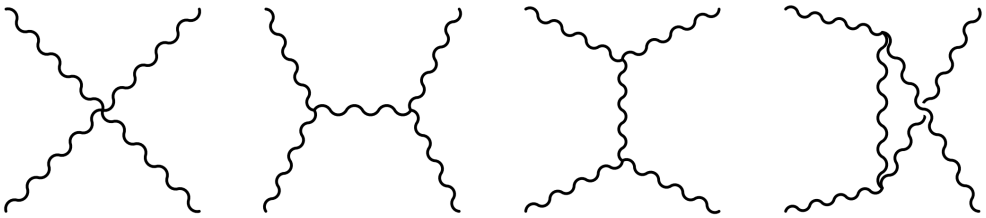
The Higgs boson was initially introduced to the SM as a means to solve the problem of the missing W and Z boson masses [3–5]. Without the Higgs boson, the SM has a symmetry in the electroweak force where all of the electroweak gauge bosons (photon, W and Z bosons) are predicted to be massless particles. The Higgs boson is introduced to the SM through coupling to the weak vector bosons (W and Z bosons) to achieve spontaneous electroweak symmetry breaking (EWSB) via the Higgs mechanism.

The introduction of the Higgs boson further leads to solving a number of other major problems within the SM. These problems include providing the masses of fermions through Yukawa couplings and fixing the breaking of unitarity in vector boson scattering at high energies [6, 7]. Unitarity being broken in this case refers to predicting a scattering probability greater than 1 and hence violating probability conservation.

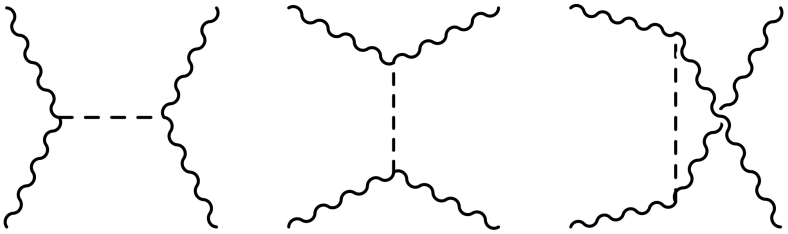
We can consider why this happens. Vector boson scattering amplitude is determined by adding the amplitudes of all the scattering processes containing triple and quartic gauge boson vertices (which can be seen in Figure 1a) [8]. At high energies, the $WW \to WW$ scattering cross section becomes dominated by longitudinally polarised gauge bosons. Without the Higgs boson, in the high energy limit, the total amplitude of the scattering of two longitudinally polarised W bosons takes the form [9]:

$$\mathcal{M}_{NoHiggs}\left(W_L^+W_L^+ \to W_L^+W_L^+\right) = -g^2\frac{S}{4M_W^2} \tag{1}$$

where g is the weak coupling strength, S is the square of centre-of-mass energy and M_W is the mass of the W boson.



(a) Contributions from electroweak gauge boson interactions.



(b) Higgs exchange contributions.

Figure 1. Vector boson scattering diagrams at tree-level [10]. Dashes represent Higgs boson propagation.

This amplitude is directly proportional to the centre-of-mass collision energy, S, and therefore means the scattering cross-section would diverge with energy. This implies that at a particular energy, the scattering probability must become greater than 1 and hence break unitarity [7]. If a Higgs interaction is included in the form of a Higgs exchange (vector bosons fuse into Higgs and Higgs decays back into vector bosons), seen in Figure 1b, the scattering amplitude of this process would take the form [9]:

$$\mathcal{M}_{Higgs}\left(W_L^+W_L^+ \to W_L^+W_L^+\right) = g_{HWW}^2 \frac{S}{M_W^4} \tag{2}$$

where g_{HWW} is the Higgs coupling strength to W bosons and is equivalent to $\frac{gM_W}{2}$.

Therefore, in the presence of a Higgs, the two contributions would be added together to obtain [9]:

$$\mathcal{M}_{NoHiggs} + \mathcal{M}_{Higgs} = g^2 \frac{M_H^2}{4M_W^2} \tag{3}$$

where M_H is the Higgs mass.

This produces a constant term which does not increase with energy and hence does not predict diverging scattering amplitudes. This allows the scattering probability to be conserved and does not violate unitarity. However, this only holds if the Higgs boson behaves

as predicted by the SM with SM Higgs couplings to the weak vector bosons. Therefore, determining the precise Higgs vector boson coupling strength can act as a strong test for the SM [11]. Any deviations from the SM predictions would give rise to anomalous effects that would lead to new physics.

Direct observations of the Higgs vector boson couplings have been carried out by the ATLAS experiment through the production of a Higgs boson via Vector-Boson Fusion (VBF) followed by the subsequent decay of the Higgs boson into two photons ($H \to \gamma\gamma$) [12]. These measurements can be seen in Figure 2. To express the experimentally measured event rates, modifiers can be applied to the SM Higgs couplings. For the Higgs couplings to the W and Z boson, these modifiers are represented by κ_W and κ_Z. In the SM these modifiers are by default 1. Any deviations in the observed measurements from the expected SM rates will change these modifiers away from 1. In the ATLAS results, the W boson coupling strength measurement is shown by the green circle and solid line. From the figure, we can see that the κ_W modifier is in tension with the SM result but due to the large uncertainty, this cannot be resolved.

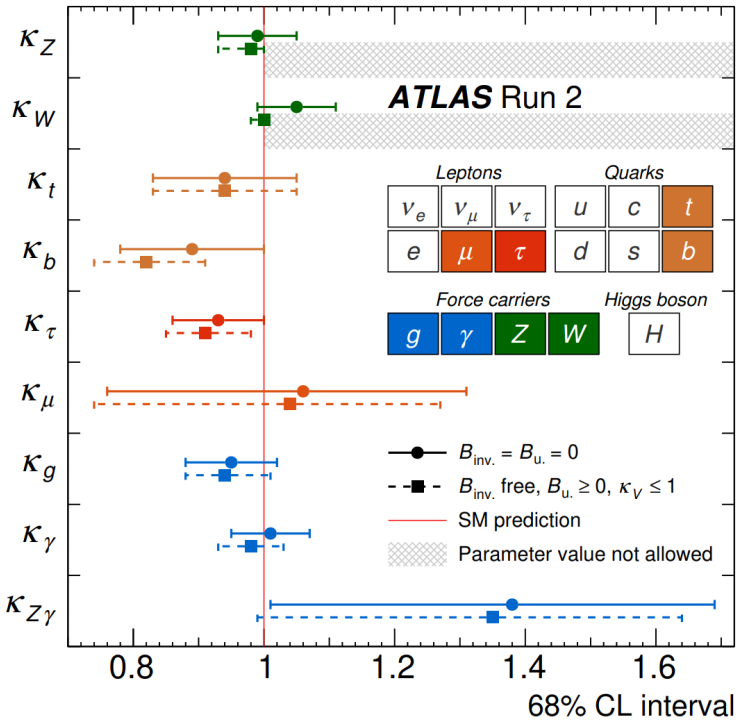


Figure 2. Coupling strengths measured by ATLAS for Higgs couplings between different SM particles [12].

The results obtained in ALTAS were achieved through the use of a neural network classification algorithm which is used to select events that look like VBF Higgs decaying into two photons [13]. This algorithm is trained on simulated data, and if the algorithm is not trained to be agnostic to the tuned simulation parameters the number of VBF events classified will be biased between different simulation samples. This introduces a systematic uncertainty

based on the theoretical simulation parameters. Furthermore, if the training data used by the algorithm is not carefully selected, the algorithm can develop a bias in background event selection. This bias is in the form of the classifier selecting more background events in the signal region, hence shaping the background to look like signal. This introduces a background modelling uncertainty that propagates to uncertainty in the number of fitted events in the signal region around 125 GeV. The combination of these two biases directly contributes to increasing the uncertainty in the measurement of the Higgs to vector boson coupling strengths.

1.2 Aim and outline

The aim of this project is to address the two types of bias present in VBF classification and hence provide solutions to reduce the measurement uncertainty in the Higgs to vector boson coupling strengths. This will be done by dedicating separate studies to each bias.

This report begins with section 2 containing the background material for the two studies. In section 2.1, the ATLAS detector and its relevant components for $H \to \gamma\gamma$ events are described. In section 2.2, the Higgs boson is introduced, where the relevance of the diphoton channel is explored and the key production modes are outlined. Section 2.3 details the LHC run-2 collision environment, the simulated event samples and the Higgs boson background sources. The reconstruction and selection of photons for candidate $H \to \gamma\gamma$ events are explained in section 2.4. In section 2.5, the ATLAS classification procedure is described, followed by an introduction to neural networks as the machine learning algorithms which will be used in this report.

The first study, in section 3 of this report, focuses on reducing the background modelling uncertainty of VBF events. This begins with posing the problem of biased mass acceptance in non-resonant background events in section 3.1. Section 3.2 outlines the current solution taken in the ATLAS $H \to \gamma\gamma$ classification to reduce the background modelling uncertainty. In section 3.3, an improved solution is presented through the use of an adversarial neural network (ANN). Details of the training features followed by definitions of the machine learning models and the training procedure are provided in sections 3.4 and 3.5, respectively. The model performance evaluation metrics are introduced in section 3.6. In section 3.7, the performance of the ATLAS solution is compared with the ANN solution introduced in this report.

The second study, in section 4, looks at reducing the systematic uncertainty arising in predicted VBF events from training on different simulated event samples. The problem of systematic uncertainty in the final states of simulated ggF and VBF events is introduced in section 4.1. Section 4.2 outlines the PYTHIA and HERWIG parton shower models that are used to generate the simulated events. The procedure of creating a reproducible benchmark ATLAS uncertainty and an independent analysis uncertainty is explained in the methods, in section 4.3. Section 4.4 defines the machine learning model that will be used. Finally, in section 4.5, the benchmark ATLAS uncertainty is determined and compared to the uncertainty obtained from the independent analysis.

2 Background

2.1 The ATLAS detector

ATLAS is one of the four main experiments in the Large Hadron Collider (LHC) at CERN, located in Geneva, Switzerland [14]. The LHC is a 27km long circular proton-proton collider which accelerates protons in opposite directions to reach the centre of mass collision energy of 13 TeV. The ATLAS experiment consists of a general-purpose detector which is forward-backwards symmetric with a cylindrical geometry measuring 25m in diameter and 44m in length, weighing 7000 tonnes. Along the central axis of the detector are located the LHC beam pipes carrying high-energy protons. At the centre of the detector, the opposing beams of protons are deflected via an electromagnetic field, focusing them onto the interaction point where they collide. At these high-energy collisions, a large fraction of the centre of mass energy is converted into new particles which shower throughout the detector.

The detector consists of four main cylindrical sub-layers that are consecutively wrapped around each other. These layers correspond to an inner tracking detector surrounded by a superconducting solenoid, electromagnetic and hadronic calorimeters, and a muon spectrometer. A two-level trigger is then used to select interesting events to be recorded based on detection criteria [15].

The ATLAS experiment uses a right-handed coordinate system with the interaction point as its origin [13]. The z-axis points along the beam direction, the x-axis points from the interaction point to the centre of the LHC ring and the y-axis points directly upwards from the interaction point. Cylindrical coordinates (r, ϕ) are used to orient in the transverse plane, with ϕ corresponding to the azimuthal angle around the beam direction. The pseudorapidity, measuring post-collision particle deflection relative to the beam axis, is defined in terms of the polar angle as $\eta = -\ln\tan(\theta/2)$. The distance ΔR is used to define conic regions originating from the interaction point and is defined in the $\eta - \phi$ space as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$.

The inner tracking detector is closest to the interaction point and measures the trajectories and momenta of charged particles [14]. This detector consists of a silicon pixel detector followed by a silicon microstrip detector which are both made up of solid-state sensors and cover the range $\eta < 2.5$. These detectors are complemented by the transition radiation tracker which enables radially extended track reconstruction and covers a range up to $|\eta| = 2.0$. Charged particles are registered as hits in these detectors which can then be fitted into a trajectory, referred to as a track.

The electromagnetic calorimeter is composed of a barrel region in the range $\eta < 1.475$ and two end-cap regions in the range $1.375 < \eta < 3.2$ [14]. These regions consist of lead/liquid-argon (LAr) layers in an accordion structure with liquid argon as the active material and lead as the absorber [16]. This calorimeter is where most fundamental electromagnetically interacting particles, such as photons and electrons, deposit their energy in the form of EM showers. These showers are usually completely contained within the calorimeter.

The hadronic calorimeter is composed of three regions [14]. At $\eta < 1.7$ is a barrel region made of steel/scintillator tile layers where steel is the absorber, and the scintillating tiles are the active medium [16]. At larger η are two LAr end-caps covering $1.5 < \eta < 3.2$ and two forward LAr calorimeters extending coverage between $3.1 < \eta < 4.9$. In these calorimeter regions, mesons in the form of jets deposit most of their energy through hadronic showers.

These calorimeter designs were driven by the potential production of the Higgs boson, with the $H \rightarrow \gamma\gamma$ decay channel specifically used as one of the benchmark channels for the electromagnetic calorimeter design [16]. The combination of all of these detector systems provides precise charged particle and photon measurements in the range $\eta < 2.5$ [13]. This is the η range which is considered in ATLAS Higgs measurements and also the range considered in this analysis.

2.2 The Higgs Boson

The Higgs Boson was postulated in 1964 when it was introduced to the standard model as a means to give the W and Z vector bosons their masses [3–5]. The problem at the time was that due to a symmetry between the photon and the W and Z bosons, the W and Z bosons were predicted to be massless particles. However, from experiments, the weak vector bosons were known to be very massive particles. To fix this the Higgs mechanism introduces a scalar field with a non-zero expectation value which spontaneously breaks this electroweak symmetry to give the weak vector bosons their masses.

As a further success, a Higgs boson coupling to the standard model fermions was then introduced through Yukawa interactions [6]. After electroweak symmetry breaking, these interactions were found to be responsible for providing the quarks and charged leptons with their masses.

Through the culmination of experimental data collected in LHC proton-proton collisions, in 2012 the joint discovery of the Higgs boson was announced by the ATLAS [1] and CMS [2] collaborations. The announcement placed the Higgs Boson mass at 125 GeV, which makes it the heaviest particle that can currently be produced.

2.2.1 The diphoton decay channel

Some of the first evidence of the Higgs boson was observed through the diphoton decay channel, where the Higgs boson decays into a pair of high-energy photons [17]. This can be seen through the Feynman diagram in Figure 3. The $H \rightarrow \gamma\gamma$ decay has a small branching ratio of $(0.227 \pm 0.007)\%$ [13]. However, setting this channel apart is its clean final state topology of two well-reconstructed photons with accurately measurable kinematic variables. When two photons are detected, their invariant mass can be calculated and used to determine the mass of the parent particle which produced them. In the case of Higgs events, these photons can therefore be used to directly calculate the invariant mass of the Higgs boson.

This leads to the formation of a narrow peak at the Higgs mass in the diphoton invariant mass distribution that can be effectively distinguished from a smoothly falling background.To obtain the exact mass and event rate from the Higgs boson decaying into two photons, a fit to the invariant mass distribution can be performed. Furthermore, the diphoton channel provides a versatile overview of Higgs events as it is sensitive to all dominant Higgs boson production modes. The combination of these features makes the $H \to \gamma\gamma$ channel one of the most important channels for precision measurements of Higgs boson properties.

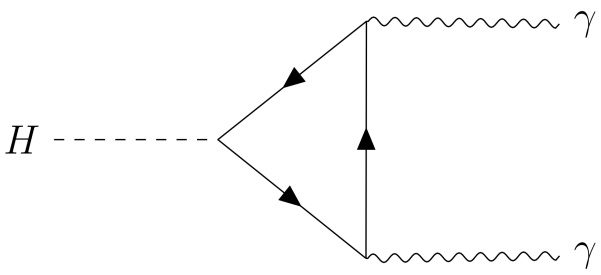


Figure 3. Decay of the Higgs boson with two photons in the final state.

Over the course of LHC run-2, the integrated luminosity of data collected from proton-proton collisions at 13 TeV has amounted to 139 fb^{-1} [13]. This has enabled the diphoton channel to achieve significant improvements in measurement sensitivity, which has allowed it to remain one of the most important channels for Higgs precision measurements [12]. The current data in the diphoton channel, collected during run-2 of the LHC can be seen in Figure 4 below.

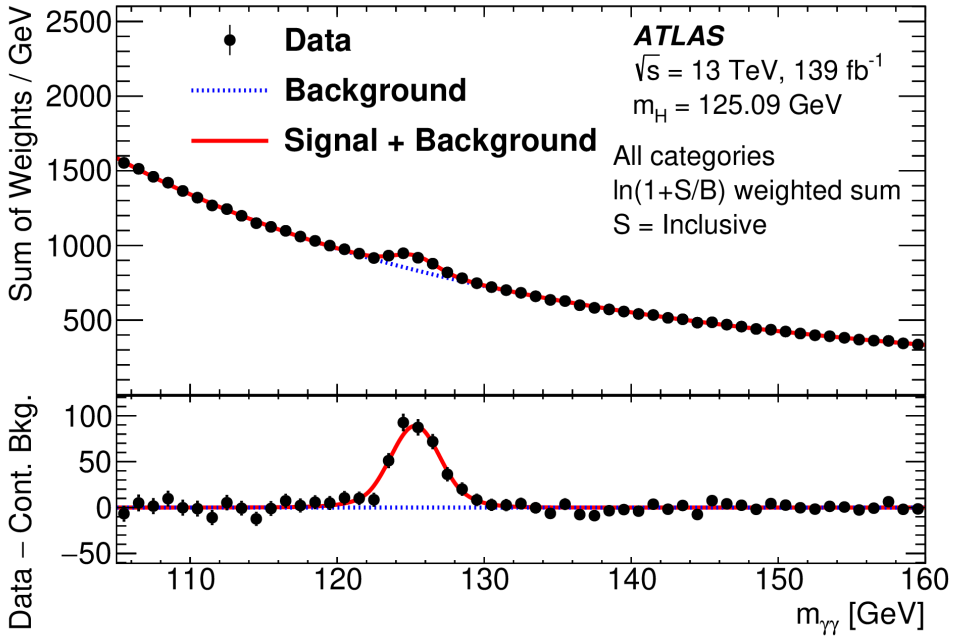
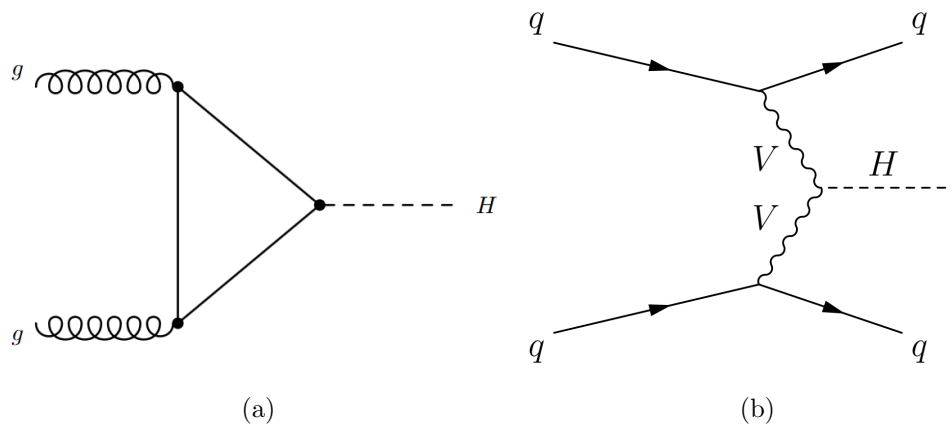


Figure 4. The data collected in the diphoton channel during run-2 of the LHC [13].

The figure shows the diphoton invariant mass distribution from all events which have been selected in classification. The main feature of the figure is the bump that rises above the smoothly falling background which signifies a resonance near the mass of 125 GeV. The excess of events near this mass corresponds to the narrow invariant mass of photons which the Higgs boson decays into.

2.2.2 The ggF and VBF production modes

The Higgs Boson can be produced through a variety of production processes [13]. However, this analysis will only focus on the two most dominant ones. The most frequent Higgs production mode is when colliding protons radiate a pair of high-energy gluons that fuse into the Higgs Boson. This mode is referred to as gluon-gluon fusion (ggF) and happens 87% of the time. The next most frequent mode is when quarks within the colliding protons radiate a pair of W or Z bosons that then fuse into the Higgs boson. This mode is referred to as vector-boson fusion (VBF) and occurs only 7% of the time. The Feynman diagrams of these respective production modes can be seen in Figure 5 below.



(a) (b)

Figure 5. The (a) ggF [18] and (b) VBF [19] Higgs boson production modes. In (b) V represents the W and Z vector bosons.

These two production modes can be characterized by their particular features. In VBF production, once the quarks radiate the vector bosons, they go on to hadronize and create jets. Therefore, to leading order the VBF process produces a Higgs boson accompanied by two jets. On the other hand, to leading order in the ggF process, there are no other objects associated with the event apart from the Higgs boson.

2.3 Collision data and simulation samples

2.3.1 Collision data

The datasets used in this analysis are based on proton-proton collisions at $\sqrt{s} = 13$ TeV that were recorded by the ATLAS detector in the period between 2015 and 2018, referred to as run-2 of the LHC [13]. Selection criteria are applied to the collision events to satisfy data quality requirements. Following this selection, the total recorded dataset amounts to an integrated luminosity of 139.0 ± 2.4 fb^{-1}. In this data, the mean number of interactions per proton bunch crossing averages to $\langle\mu\rangle = 33.7$.

The events selected from the collisions pass through diphoton or single-photon trigger requirements [13]. However, for this analysis, only events that pass the diphoton trigger requirements are considered. For two potential photons, the diphoton trigger has transverse energy thresholds of 35 GeV and 25 GeV for the leading ($p_T^{\gamma_1}$) and sub-leading ($p_T^{\gamma_2}$) photon candidates. These requirements are applied to any two high-p_T tracks associated with the collision vertex. Further photon identification selections are carried out using the calorimeter shower shape variables. The trigger efficiency for events passing the diphoton event selection is on average greater than 98%.

2.3.2 Simulation samples

The ggF and VBF production processes considered in this study were generated using POWEHEG BOX v2 [13]. In this simulation, the ggF process achieves next-to-next-to-leading-order (NNLO) accuracy while the VBF process was simulated at next-to-leading-order (NLO) accuracy. In collisions, protons are complex objects with an internal structure which can be approximately modelled by parton distribution functions (PDFs). To model the parton distributions during these simulations, the PDF4LHC15 set [20] of parton distribution functions is used, with the NNLO set used for ggF and NLO set for VBF.

All generated ggF and VBF events were then interfaced to PYTHIA 8.2 [21] to model the parton showering, hadronization and the underlying event. To determine the systematic uncertainties associated with the shower model an alternative dataset is produced where HERWIG 7 [22] is used for parton showering. These datasets are summarised in Table 1. To achieve the most accurate results, the generator parameters are tuned to produce simulated data which closely matches the available recorded data. In this simulation, PYTHIA 8.2 uses the AZNLOCTEQ6 tune and HERWIG7 uses the H7p1 tune.

All Higgs boson events were generated with a Higgs boson mass (m_H) of 125 GeV where the mass has an intrinsic width (Γ_H) of 4.07 MeV [13]. The cross-sections used for the ggF and VBF processes are taken at the centre-of-mass energy of $\sqrt{s} = 13$ TeV and for a Higgs boson with mass $m_H = 125.09$ GeV. These cross-sections can be seen in Table 1 below and can be combined with the branching ratio of $H \to \gamma\gamma$ to normalise the simulated signal events.

Process	Generator	Showering	PDF set	σ[pb] $\sqrt{s} = 13$TeV
ggF	POWEHEG BOX + NNLOPS	PYTHIA 8.2	PDF4LHC15	48.5
VBF	POWHEG BOX	PYTHIA 8.2	PDF4LHC15	3.78
ggF_{alt}	POWEHEG BOX + NNLOPS	HERWIG 7	PDF4LHC15	48.5
VBF_{alt}	POWHEG BOX	HERWIG 7	PDF4LHC15	3.78
$\gamma\gamma$	SHERPA	SHERPA	NNPDF3.0NNLO	

Table 1. Nominal and alternative signal samples and nominal background sample [13]

Background events which correspond to prompt diphoton production ($\gamma\gamma$) were simulated with the SHERPA 2.2.4 [23] generator. These samples were generated with the $\text{NNPDF3.0}_{\text{NNLO}}$ PDF set [24], where the parton showering was tuned to the default SHERPA parameters. The effects of collision pile-up during events were modelled by overlaying the original hard-scattering collision with simulated in-elastic proton-proton events.

Once the ggF and VBF samples were prepared the response of the detector was then modelled through a detailed simulation of the ATLAS detector based on GEANT4 [25]. However, for the prompt diphoton sample, a fast simulation of the ATLAS detector is used instead, where a parametrization of the calorimeter response replaces the full simulation [13].

2.3.3 Higgs background sources

The main background sources for Higgs events in the diphoton channel are processes that can imitate the two photons which the Higgs decays into and any of the signatures that particular production modes produce [13]. This criterion encompasses many different processes which can take place in proton-proton collisions represented generally by $pp \to \gamma\gamma$. However, apart from the continuum background of two photons, the selected background sample also consists of photon-jet pairs and jet-jet pairs where one or more jets in the event are misidentified as photons.

Processes which produce non-resonant background events that contaminate the Higgs events in the diphoton channel are not very well understood and are hence inaccurately modelled by the same simulation procedure used to generate the signal [26]. This means that background events need to be modelled through a coarse-grained approach which does not take into account the individual background processes and the exact physics behind them. Instead, the background diphoton invariant mass distribution is assumed to be a smoothly falling spectrum which can be accurately fit by only using the non-signal sidebands of the distribution (also referred to as the control regions) [13]. The fit of the distribution can then be interpolated over the signal region to also obtain the number of background events which contaminate the signal.

2.4 Event reconstruction and selection

This section focuses on the key steps taken to reconstruct and select $H \to \gamma\gamma$ events to be used later for further analysis. In the events, reconstructed photon candidates are selected by first requiring them to pass the pre-selection criteria outlined in section 2.3.1. The two highest-p_{T} photons that pass the pre-selection are then used as the basis of the diphoton system with their origin defining the event's primary vertex [13]. During reconstruction, photons must satisfy showering criteria and additional isolation requirements.

2.4.1 Photon reconstruction and selection

Photon candidates can be classified based on whether or not they decay into an electron-positron pair before entering the calorimeter, corresponding to the converted and non-converted categories [13]. Converted photons will have tracks associated with them in the inner detector while non-converted photons will not. Once photons or electrons make it to the EM calorimeter, they interact with the absorber and active material to produce regions of activated cells. Photons are then reconstructed through a topological cell-clustering algorithm using dynamic, variable-size energy clusters in the EM calorimeter [27]. The energy deposits are then calibrated to associate the photon candidate with physical energy. Energy clusters determined in this way can be associated with tracks that are reconstructed in the inner detector if photon conversion vertices are identified [13]. Photons are then classified as converted if calorimeter energy clusters are matched to two tracks that originate from a conversion vertex or a single electron track with no hits in the innermost tracking layer. If a photon does not satisfy these criteria then it is classified as unconverted.

The quality of photon reconstruction largely depends on the trajectory that photons take through the detector components. In this analysis, photons are required to pass through the first finely segmented region in the EM calorimeter that lies in $\eta < 2.37$ and must be outside of the transition region between the barrel and end-cap EM calorimeters that lies in $1.37 < \eta < 1.52$.

In the process of photon reconstruction, the diphoton primary vertex is determined through the use of a neural network algorithm that is trained on simulated data. To identify the best primary vertex the neural network is given the reconstructed vertices and trajectories of the two photons as inputs. This approach is superior to default primary vertex selection and achieves 8% improvement in the mass resolution for inclusive Higgs boson production.

To reduce background from neutral mesons decaying into photon pairs and charged particles from collision pile-up, photon isolation requirements are applied in both the calorimeter and the tracker [28]. The photon isolation region corresponds to a cone of size $\Delta R = 0.2$ that originates from the diphoton primary vertex and is centred around the calorimeter energy deposits [13]. Calorimeter energy deposits inside the isolation cone, but outside the photon shower window, must be less than 6.5% of the photon transverse energy. The tracker isolation only considers tracks with $p_{\mathrm{T}} > 1$ GeV within $\Delta R < 0.2$ that originate from the diphoton

primary vertex. The scalar sum of track p_{T} within the isolation region must be less than 5% of the photon transverse energy.

The final selection from the photon candidates requires the highest-p_{T} and second-highest-p_{T} photons to satisfy $\frac{p_{\mathrm{T}}}{m_{\gamma\gamma}} > 0.35$ and 0.25, respectively. Events that pass the trigger and the photon selection requirements are then taken to be used for further analysis of Higgs boson properties. On the other hand, events that pass the trigger but fail the photon selection requirements are instead used as a control sample to estimate and model the background. The standard model Higgs boson efficiency achieved through this process is predicted to be 39% for $|y_H| < 2.5$, where y is the rapidity.

2.5 Classification using Machine learning

Currently, the classification of $H \to \gamma\gamma$ events in ATLAS is carried out through the use of a Boosted decision tree (BDT) machine learning algorithm [13]. This model is based on an ensemble of weaker decision tree algorithms that are combined through a boosting procedure to enhance their collective performance. This algorithm is first trained on simulated data and then applied to classify unseen events.

The classification procedure in ATLAS is split up into two stages. First, a multi-class BDT algorithm is applied to classify Higgs events based on their production mode. To do this, production modes are split up into classification regions based on their kinematic properties. To achieve a finer picture within each production mode, the inclusive kinematic regions are further split up into subcategories. The inclusive kinematic regions and their subcategories are defined by mutually exclusive simplified template cross-section (STXS) regions. In total there are 28 STXS regions considered. During classification, these STXS regions are further split up into 45 analysis categories which Higgs events with different kinematic properties can be classified into. In the second stage, a binary BDT algorithm is applied separately to each STXS region. This algorithm is used to classify events that belong inside a given STXS region and events that do not belong inside, such as non-resonant background and signal from other STXS regions.

Although BDTs are powerful machine-learning models, they do not show the best scalability in performance with increased training data [29]. Neural network algorithms have been shown to scale better in performance as they are given more training data since they can learn more complex functions. This has made neural networks the increasingly favoured model as larger amounts of collision data are being collected.

2.5.1 Neural networks

Neural networks are a type of machine-learning model which is inspired by the way neurons in the human brain function [30]. A neural network represents a series of neurons (nodes) which are inter-connected, with each connection having an associated weight. An effective

arrangement of the nodes is achieved when they are structured into layers, corresponding to an input layer, a series of hidden layers and an output layer. This arrangement can be seen in Figure 3. For each node to be optimally used, nodes in successive layers are connected to every node in the previous layer.

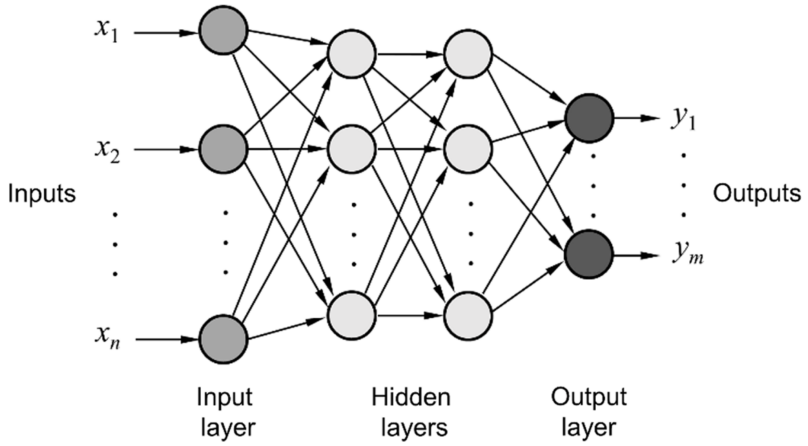


Figure 6. The structure of a neural network [31]

In this analysis, neural networks will be trained through a supervised approach where input training samples are provided with their corresponding labels [30]. In the first phase of training, unlabelled input data is fed through the input layer nodes, where each node takes on one of the input data values. The size of each input node value is then referred to as its activation. Following this, nodes in the following layers are sequentially updated with information being passed in one direction, referred to as forward propagation. To do this the neurons in the first hidden layer are updated through a weighted sum of the neuron activations in the input layer. A constant bias term is added to the sum at each layer to help regulate the updates. For a given node this takes the form:

$$b + \sum_{i=1}^{n} x_i w_i \tag{4}$$

where, b is the bias term, x_i are the previous layer node activations and w_i are the corresponding weights between nodes.

The result of the summation is then transformed through a non-linear activation function such as the sigmoid or rectified linear unit (ReLU) functions given by [32]:

$$S(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

and

$$RELU(x) = \max(0, x) \tag{6}$$

respectively.

This procedure obtains the activation for a given node and is applied separately to each node in the layer [30]. The procedure is repeated for all hidden layers up to the output layer. The output layer consists of either one node for binary classification or multiple nodes for multi-class classification. Once the network provides an output classification, this is then compared to the true classification label. This is done through a loss function which computes the effective classification performance of the network.

Following this, the second phase of training takes place where a gradient is calculated based on whether the loss function increased or decreased. This gradient is then used to sequentially update the node activations of each layer starting from the last hidden layer. This is referred to as backpropagation. Depending on whether the network performance improved or worsened after the update, node weights are either reinforced or penalised proportionally to the size of their change.

During the two phases of training the neural network picks up patterns within the input data which allow it more accurately predict the labels of samples. Once a neural network has been trained it can then be applied to unseen data to predict its labels.

3 Study 1: Classification between VBF Higgs and non-resonant background events

This study focuses on the binary classification between VBF Higgs and non-resonant background events, corresponding to the second stage of the ATLAS $H \to \gamma\gamma$ classification procedure. The effects of mass correlation are studied to reduce distortions in the shape of the background $m_{\gamma\gamma}$ distribution following classification.

3.1 Biased mass acceptance of non-resonant background

If a neural network is indiscriminately trained on all event variables and is used to predict $H \to \gamma\gamma$ events, it will be able to achieve very high discriminating power. Such a network would correctly classify almost all signal events and keep background acceptance to a minimum. However, this performance is not reflected in the diphoton invariant mass distribution. The diphoton invariant mass distribution of such a classifier can be seen in Figure 7 below. From this figure, it can be determined that the network achieves high accuracy by rejecting all background events outside of the signal region and filtering the remainder. This results in a significant distortion of the accepted background distribution which is referred to as background sculpting. Due to this distortion, reliable background extrapolation from the sidebands is no longer possible [33]. This propagates to an uncertainty in the number of fitted resonant events in the signal region. As a result, any measurements of Higgs properties using the events classified through this approach would produce large uncertainties.

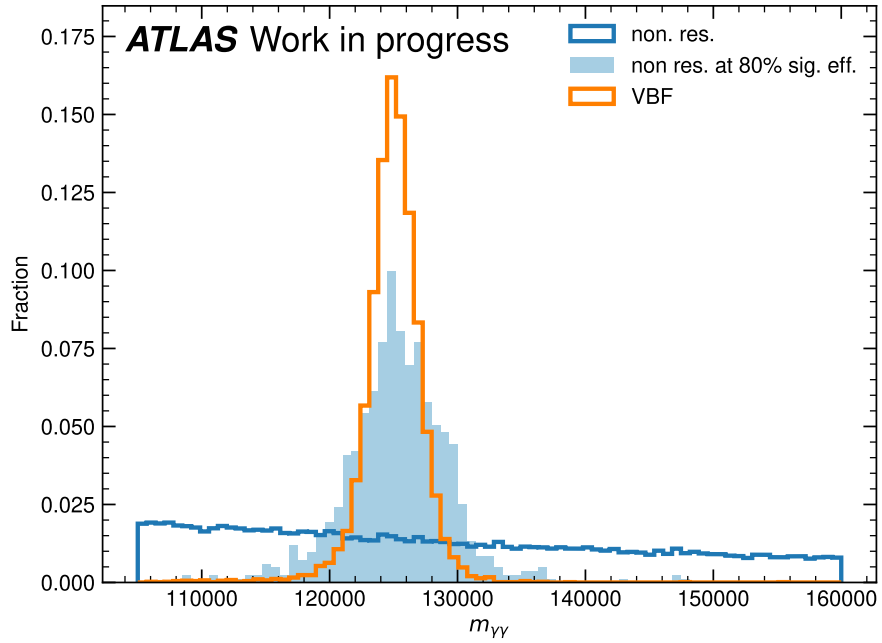


Figure 7. The effects of classifying background events based on training with variables highly correlated with the mass. The background distribution is significantly sculpted to look like the signal distribution.

The behaviour of the classifier can be understood by considering its training objectives and input features. In training, many features are provided but the diphoton invariant mass is omitted. The goal of the classifier is to find patterns in these features and obtain the best classification accuracy regardless of the method used to achieve it. The problem with this training objective is that many of the background features are correlated with the diphoton invariant mass. Therefore, as the network trains it can indirectly learn the invariant mass and find that this variable provides a powerful separation between signal and background. Since the background is close to uniformly distributed and the signal is a sharp peak, the network achieves high accuracy by rejecting all events outside the signal peak.

3.2 Mass decorrelation through feature selection

To remedy the problem of background sculpting, mass decorrelation needs to be performed [33]. The current solution implemented by the ATLAS analysis is to reduce mass correlation by removing event variables which have $m_{\gamma\gamma}$ correlation amongst the background that is above a certain threshold [13]. Removing variables that are highly correlated with $m_{\gamma\gamma}$ reduces the information training variables can give away about the $m_{\gamma\gamma}$ of each event. This significantly reduces the ability of the network to learn the signal and background $m_{\gamma\gamma}$ distributions and hence removes the background sculpting after classification. The current threshold of background correlation with $m_{\gamma\gamma}$ is placed at 5%.

The drawback of this method is that imposing selection criteria on the training features significantly reduces the number of training features available. By reducing the number of training features, the network has less information to pick out patterns and hence this reduces its discriminating power. This approach, therefore, creates a large tension between the objectives of high discriminating power and minimal background sculpting.

3.3 Mass decorrelation through an ANN

An alternative approach to mass decorrelation can instead be implemented with a focus on controlling the training objective of the classifier [33]. This approach allows a forgiving use of highly correlated variables with $m_{\gamma\gamma}$ and therefore removes the constraint on input features, significantly increasing the viable set of training data. This alleviates the trade-off between high discriminating power and background sculpting by decoupling their relationship and arriving at a more optimal solution.

In practice, this can be achieved through the use of an adversarial network which is introduced to train along with the classifier. This is implemented by linking the two networks in series, as seen in Figure 8 below.

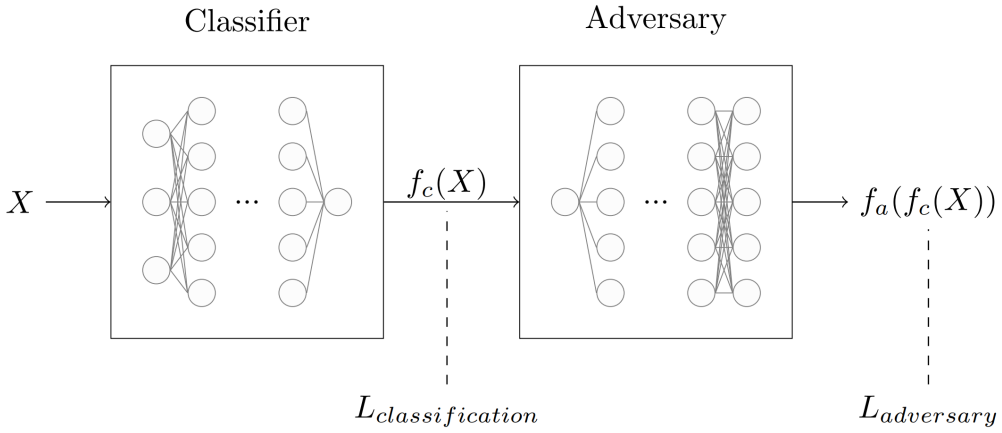


Figure 8. The combined classifier and adversary network in series [26].

The classifier loss function can be combined with the adversary loss function to obtain a modified loss function [33]:

$$L_{modified} = L_{clf} - \lambda L_{adv} \tag{7}$$

where $L_{modified}$ is the modified loss of the classifier, L_{clf} is the stand-alone classifier loss, L_{adv} is the adversary loss and λ is the regularisation strength (also referred to as the gradient reversal strength in this report). The objective of the classifier remains to still distinguish between signal and background events but now it does this by minimising the modified loss

function instead. The objective of the adversary is to minimise its own loss function, which is then used to update the modified loss function.

The adversary network has the goal of attempting to predict the invariant mass ($m_{\gamma\gamma}$) of signal events based on the output from the classifier (z). It does this by fitting a gaussian mixture model (GMM) to the accepted background distribution. This represents varying the parameters of a user-defined number of gaussian distributions to try and obtain an accurate representation of the signal mass distribution prior to classification.

If the adversary is able to get a good fit of the original mass distribution (maximising $p(m_{\gamma\gamma}|z)$) this results in minimising its loss function. However, when the adversary loss function decreases, due to the negative sign in front of the adversary term, the modified loss function increases. This means if the adversary is able to predict the prior signal mass distribution during training, the classifier is penalised thereby forcing it to achieve good signal and background separation through a more optimal approach.

In practice, this behaviour is implemented through a gradient reversal operation [34] at the connection between the two networks. Once the adversary fits the GMMs and happens to reduce its loss function, this results in a negative loss gradient [33]. Through gradient reversal, the negative gradient is turned positive and is backpropagated through the output of the classifier to penalise weight updates that learned the $m_{\gamma\gamma}$ distribution.

The gradient reversal strength, λ, plays an essential role in maintaining the right balance between training objectives. The classifier could either not be penalised enough and allow significant background sculpting or it could be over-penalised where small unrelated improvements in adversary loss could begin to distort classifier weight updates and stunt the training process. Therefore, λ serves as a gauge for the trade-off between classification performance and background sculpting.

3.4 Training features

Well-selected training features are a crucial component in allowing neural networks to achieve a powerful separation between signal and background events. Due to the similarity in studies, the already optimised feature definitions in [13] were used as the basis for the training inputs of this study. Using a subset of these definitions relevant to the physics objects considered in this study, low-level detector variables were combined to create highly engineered, physics-motivated input features.

Event selection cuts are applied to both the signal and background distributions where only events satisfying the criteria njets ≥ 2 and $m_{jj} \geq 300$ MeV are kept. Here njets is the number of jets in a given event and m_{jj} is the invariant jet mass of the two highest-p_T jets. The event samples were then split via random sampling into training (64%), validation (16%) and testing (20%) subsets, where the validation subset was used for hyperparameter tuning. The mean and standard deviation of the training set were then used to standardise all subsets. Following the training and hyperparameter optimisation of a model, the testing set is then used as an unseen dataset to evaluate how well a model generalises to new data.

3.5 Implementation

As a benchmark, the current ATLAS procedure of mass de-correlation [13] will first be implemented by solely training a deep neural network (classifier) on the selection of input features where the background features have a correlation with $m_{yy} < 5\%$. Then to demonstrate classification through a modified training objective, the classifier will be connected with an adversary network to form a single feed-forward architecture as introduced in Figure 8 above. Both network models are constructed in KERAS 2.9.0 using the TensorFlow 2.9.1 backend.

3.5.1 Defining the machine learning models

The benchmark classifier in this analysis consisted of 12 input nodes (one for each input feature), three hidden layers with 48, 32, and 16 nodes respectively, and a single output node. The hidden layers use a rectified linear unit (ReLU) activation function while the output node uses a sigmoid activation function. The sigmoid function provides an output in the range [0,1] which makes it suitable as a likelihood metric, with output values closer to 1 indicating a higher signal likelihood. The classifier is trained to perform binary classification between the event labels based on the input features by minimising the binary cross-entropy loss [35]:

$$L_{\text{clf}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i) \tag{8}$$

where $\hat{y}_i$ are the predicted outputs from the last layer node (in the range [0,1]) and y_i are the binary labels for the events, 0 = background, 1 = signal. The sum is carried out over all events and is then averaged.

To keep the results consistent, classification through a modified training objective will be performed by connecting an adversary network to the existing benchmark classifier network. In this role, the adversary will be constructed to fit the coefficients, means and widths of the components that make up a GMM which parametrises the background distribution.

The adversarial network consisted of one input node (the output of the classifier), 3 hidden layers of 200, 100 and 50 nodes respectively, and 3 GMM component fitting layers (1 layer each for gaussian coefficients, means and widths) that are combined into one posterior layer. The posterior layer produces a single output corresponding to the posterior probability. At the GMM fitting layers, the prior mass distribution is scaled in the range [0,1] before being given to the adversary as this makes fitting easier. The hidden layers use a ReLU activation function while the GMM layers use a softmax activation for the coefficients, a sigmoid activation for the means and a softplus activation for the widths. The softmax function makes sure the gaussian coefficients all add up to unity, to ensure the normalisation of the GMM. The sigmoid function makes sure the gaussian means are in the range [0,1] whereas the softplus function makes sure the gaussian widths remain positive.

The adversary will be tasked with de-correlating the classifier output from the signal $m_{\gamma\gamma}$ distribution by regularising its training through the addition of the adversary loss [33]:

$$L_{\text{adv}} = -\frac{1}{N}\sum_{i=1}^{N}\log(p_{\text{adv}}(m_i \mid z_i)) \tag{9}$$

where m_i is the true invariant mass of an event, z_i is the classifier output for the event (in the range [0,1]) and $p_{\text{adv}}(m_i \mid z_i)$ is the metric of how well each invariant mass is modelled by the GMM model constructed by the adversary.

The adversary loss is only computed for background events as we want to only penalise sculpting of the background mass distribution. To do this, in the adversary loss signal events are given a weight of zero while background events are given a weight of one. The background weights are then rescaled to sum to the total number of signal and background events. The combined network is then trained to minimise the joint objective which is given by equation 7.

3.5.2 Training the models

Training of both models was done through a selection of subsets from the pre-processed training features. To obtain the benchmark training set, the linear correlation between all input features and $m_{\gamma\gamma}$ is determined. The subset of features satisfying m_{yy} corr. $< 5\%$ is the set of features which are used to train the stand-alone classifier to determine the benchmark performance.

Obtaining the combined model training set is more involved. The training variables used for the combined model are largely dependent on the impact of the adversary on the classifier's training. The stronger the adversary penalises the classifier the greater becomes the range of training variables which the classifier can use without producing background sculpting. This means, to achieve optimal performance, both the gradient reversal strength and the threshold of $m_{\gamma\gamma}$ correlation within the training set need to be tuned together. The most optimal training set for the combined network is determined in the hyperparameter optimisation section.

To train the benchmark and combined models, training parameters motivated by [33] were used. The configuration used in this study corresponded to a training batch size of 5000 events with each model trained for 100 epochs using the ADAM optimiser. This batch size was found to provide a good balance between classification performance, training times and reproducibility of results. Furthermore, 100 epochs were found to be large enough for each model to reach a sufficient saturation in training performance while reducing time spent in a training plateau.

3.6 Evaluation metrics

To effectively evaluate the performance of the models, a set of evaluation metrics can be devised based on the training objectives of high discriminating power and minimal background sculpting.

3.6.1 Classification efficiency

A natural metric for discriminating power is classification accuracy. In this study, we are interested in the fraction of events that are correctly classified as signal and the fraction of events that are incorrectly classified as signal, referred to as signal efficiency and background efficiency.

These efficiencies are defined as:

$$\varepsilon_{\text{sig}} = \frac{N_{\text{sig}}^{\text{classified}}}{N_{\text{sig}}^{\text{total}}}, \quad \varepsilon_{\text{bkg}} = \frac{N_{\text{bkg}}^{\text{classified}}}{N_{\text{bkg}}^{\text{total}}} \tag{10}$$

respectively. Where N^{total} refers to the total number of events that belong to a given class which pass through the trigger selection requirements and $N^{\text{classified}}$ refers to the number of events passing through both the trigger and classification requirements. To make a clear summary comparison between models with different performance, the background efficiency will be reported at a fixed signal efficiency of 80%.

These efficiencies can be combined to produce a receiver operating characteristic (ROC) curve by varying the selection threshold which defines signal and background events in the discriminant score distribution. To produce the curve, the signal efficiency is plotted on the y-axis against the background efficiency on the x-axis. In this plot, it is desirable for the signal efficiency to remain as high as possible while background efficiency is kept to a minimum. This represents a higher signal purity, with a larger fraction of signal events being selected out of the total. To provide a quantifying metric based on the ROC the area under the curve (AUC) can be reported. The ROC AUC can vary between the range [0,1], with greater values indicating higher discriminating power.

3.6.2 Jenson-Shannon divergence

Determining the best metric for background sculpting is less straightforward as there are many methods of measuring the difference between two distributions. The strongest motivated method, also used in [33], is the Jenson-Shannon divergence (JSD) which is based on the Kullback-Leibler divergence. For the discrete probability distributions P and Q, the Kullback-Leibler divergence is defined as [36]:

$$\mathrm{KL}(P\|Q) = -\sum_i P_i \log\left(\frac{Q_i}{P_i}\right) \tag{11}$$

where the subscript i denotes each of the discrete distribution bins. This metric can be used to measure and quantify the relative entropy between the probability distributions P and Q.

As a symmetrised and finite valued version of the Kullback-Liebler divergence, the JSD can be introduced as [37]:

$$\mathrm{JSD}(P\|Q) = \frac{1}{2}\mathrm{KL}(P\|M) + \frac{1}{2}\mathrm{KL}(Q\|M) \tag{12}$$

where $M = \frac{1}{2}(P+Q)$.

In this study the JSD metric will be used to quantify the difference between the normalised background mass distributions before and after a selection cut is applied to the background events. When comparing the two distributions, a smaller JSD score indicates that the background has little distortion while a larger JSD score indicates the presence of background sculpting. To summarise the level of background sculpting between different models, the JSD score will be reported at 80% signal efficiency.

3.7 Results and analysis

In the following analysis, the stand-alone classifier will be referred to as the NN and the combination of the classifier and the adversary will be referred to as the ANN. In this section, the performance of each model will be analysed. To perform the analysis the NN with the architecture described in section 3.5.1 is trained on the benchmark training set outlined in section 3.5.2. This corresponds to the current ATLAS implementation to remove background mass sculpting and will be the benchmark result. Following this, the best ANN model will be determined and then its performance will be compared to the benchmark model.

3.7.1 Benchmark performance of the stand-alone classifier

Once the NN classifier is trained its performance is then tested using the unseen testing dataset. The NN classifier is used to give each unlabelled event in the testing set a discriminant score (D-score) that represents the likelihood that a given event is signal. The events in the D-score distribution are then given back their labels to produce the distribution in Figure 9 below. Using this distribution, a D-score threshold can then be placed to define all of the events above a certain D-score as signal. This then allows to calculate the signal and background efficiencies at the given threshold. For this analysis, a D-score threshold is

chosen such that 80% of the signal events are correctly classified as signal, equivalent to 80% signal efficiency.

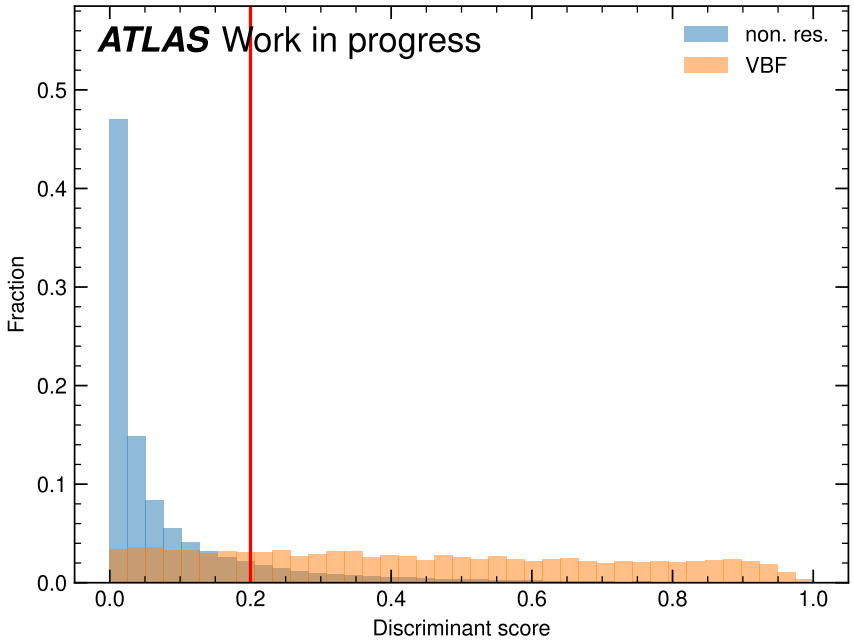


Figure 9. The D-score distribution of the NN benchmark network. The threshold is placed at a signal efficiency of 80% to signify the cut-off point where all events above the given discriminant score are chosen as signal.

To analyse the level of background sculpting, the background $m_{\gamma\gamma}$ distributions before and after the selection cut need to be compared in shape. The comparison of the two background distributions can be seen in Figure 10 below. This figure shows that through the use of training set variables with $m_{\gamma\gamma}$ corr.<5%, the background sculpting has been effectively removed. The background distribution has the same shape before and after the cut. The background distribution after the cut has much fewer events and is therefore more greatly influenced by statistical fluctuations. This explains why it is less consistent in shape over different bins. A clearer comparison between the background distributions can be observed through the ratio plot in the lower plot of Figure 10. In this plot, the two distributions closely follow each other and, within statistical uncertainty, there is no clear difference between the two.

To obtain an objective measure of the degree of similarity, the entropy between the distributions will be calculated through the JSD metric. However, to provide an effective handle for background sculpting, the JSD will be measured separately for the inclusive and Higgs mass regions. This is because although there may not be a clear overall shift in the inclusive $m_{\gamma\gamma}$ range, there could still be a localised shift in the Higgs region. In this analysis, the Higgs region will be defined as the region between 121 GeV $< m_{\gamma\gamma} <$ 129 GeV. As a measure of the

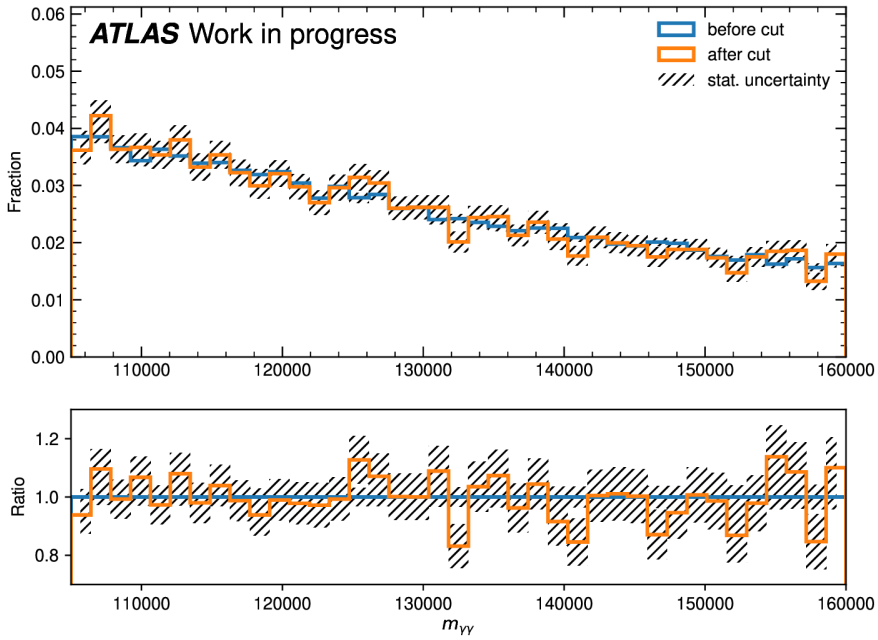


Figure 10. Background distribution ratio plot before and after selection cuts for training with $m_{\gamma\gamma}$ corr.<5%. No background sculpting can be seen.

discriminating power of the benchmark model, the background efficiency will be reported at 80% signal efficiency.

The benchmark model was trained on five separate occasions to obtain five independent JSD score measurements for both the inclusive and Higgs regions and five independent background efficiency measurements. The means were determined from these measurements to obtain accurate averages. However, the uncertainty of the JSD cannot be taken from the standard deviation of these scores as the events that are selected in each training are highly correlated. Instead, for each JSD score the background efficiency achieved by the model was used to select a random sample of events from the original background distribution with the same efficiency. The JSD score was then calculated between the sampled and original distribution. This pseudo-experiment was performed 1000 times, the JSD scores were ordered in the size (smallest to largest) and the JSD score at the 68th percentile was taken to represent the uncertainty of the measurement. As an approximation, this uncertainty will be reported as being symmetrical around the mean value. The benchmark JSD scores measured in this way correspond to $\text{JSD}_{Incl} = 7 \pm 4 \times 10^{-4}$ and $\text{JSD}_{Higgs} = 3 \pm 3 \times 10^{-4}$, while the background efficiency was measured to be 0.165 ± 0.002.

From these results it becomes apparent that the JSD score is greater for the inclusive region, suggesting that there should be greater distortion between the two distributions over the inclusive region compared to the Higgs region. However, this cannot be assumed as the JSD score depends on the number of bins and for each region a different number of bins is

considered. In this analysis, the Higgs region bins are just a subset of the inclusive region bins that lie within the Higgs mass range. In the inclusive region, the JSD is calculated over 40 bins while for the Higgs region, only 6 of these bins are used. Therefore, the summation of differences over the fewer bins in the Higgs region is likely to be smaller and hence produces a lower JSD.

Furthermore, the results also show that a greater variance in JSD scores is present for the Higgs region. The main reason for this is that each bin in the Higgs region contributes a larger fraction to the JSD score compared to the inclusive region. Therefore, for the same level of fluctuations in both regions, the Higgs region JSD will be more strongly affected. Given the complex relationship between the two JSD scores, they will not be directly compared to each other and will only be used as a comparison between different models.

3.7.2 Hyperparameter optimisation of the ANN

The optimal ANN values for the gradient reversal strength, λ, and the $m_{\gamma\gamma}$ correlation threshold for the input variables are chosen through a hyperparameter grid search. The aim of the grid search was to find the set of parameters which achieve the highest discriminating power at the same or lower level of background sculpting (JSD score) as the benchmark model. This ensures the improvement in performance does not come at a cost of higher mass correlation. The best hyperparameters are selected which satisfy JSD(ANN) $\leq$ JSD(NN) and minimise the background efficiency. The hyperparameters with the corresponding set of values used in the grid search can be seen in Table 2 below.

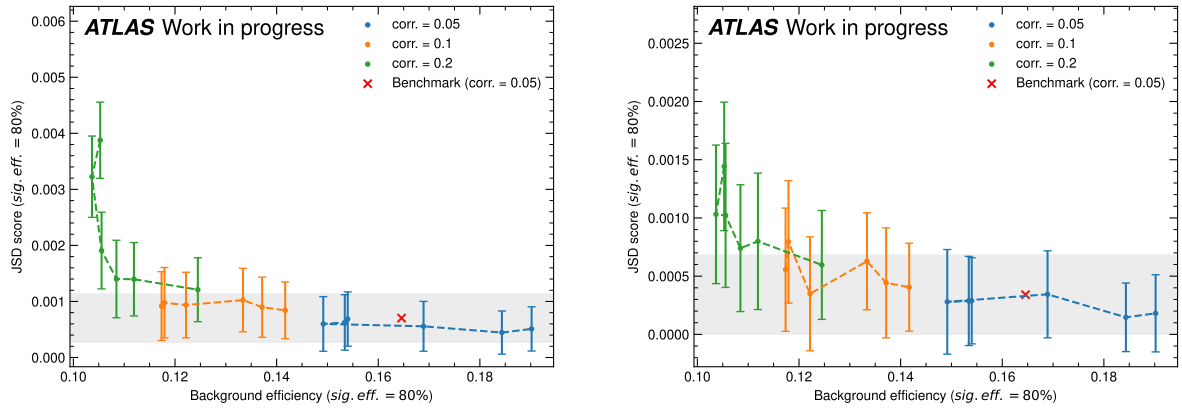
Parameter	Value	Selected value
Training set correlation	[0.05, 0.1, 0.2]	0.1
Regularisation strength	$[0.1875, 0.375, 0.75, 1.5, 3, 6]$	0.1875

Table 2. The ANN hyperparameters optimised via grid search, the values that each parameter was searched over and the corresponding chosen parameter values.

The position of each training set correlation threshold was carefully designed. The lower threshold was chosen to be identical to the benchmark threshold. At this threshold, the ANN performance is used for two purposes. The first purpose is to act as a control metric on whether the optimisation is working correctly, while the second purpose is to provide a direct comparison of how the performance can improve by increasing the dataset. The upper threshold was chosen as the most effective compromise between increasing the training set size and excluding the photon variables with the highest correlation. This is because the increase in information from these variables is significantly outweighed by their high potential for background sculpting. Finally, the middle threshold was chosen to achieve an equal increment in the training set size between the lower threshold, itself and the upper threshold. Given this, the number of variables used in the 0.05, 0.1 and 0.2 thresholds corresponds to 12, 19 and 28 respectively.

The gradient reversal strength values were chosen with considerations of efficiency and span. The thresholds were determined by training on the benchmark training set as the ANN performance can be directly compared with the benchmark values. The upper threshold was chosen at the point where the ANN $\mathrm{JSD}_{\mathrm{Incl}}$ fell below the benchmark score and where further increases in the threshold produced an insignificant reduction in $\mathrm{JSD}_{\mathrm{Incl}}$. The lower thresholds were then successively placed to decrease in size by a factor of 2. The final threshold was determined where decreasing the gradient reversal strength further did not provide a decrease in the background efficiency.

The grid search approach searches through all of the possible combinations of hyperparameters and records the corresponding JSD and background efficiency scores achieved by each model. The results of the hyperparameter optimisation can be seen in the JSD-background efficiency space shown in Figure 11. The results for the inclusive and Higgs regions JSD scores are shown separately in plots 11a and 11b respectively.



(a) (b)

Figure 11. The hyperparameter optimisation results. Each model's performance is shown in the JSD-background efficiency space for both the (a) inclusive and (b) Higgs JSD regions.

The results in these plots were obtained by training with each training set separately and varying the gradient reversal strength, starting from the highest strength and successively decreasing to the lowest. The individual sets of coloured points in the plots represent the performance of a model trained on the same training set. The points are connected by dashed lines to show how the model performance changes as smaller gradient reversal strengths are used in each iteration. The red cross represents the benchmark performance, while the grey band corresponds to the uncertainty on the benchmark JSD score. The red cross and the grey band are used as a reference to gauge the ANN performance.

In the plots, we can see that similar behaviour is exhibited by the models trained on each training set. For each training set model the highest background efficiency takes place at the highest gradient reversal strength and as the gradient reversal strength decreases each

set of models decreases its background efficiency. The behaviour from the 5% correlated set used to determine the gradient reversal strength boundaries can also be seen in the 10% and 20% sets. This is signified by the background efficiency continuously falling until the last point where the efficiency remains constant (10% corr.) or increases (20% corr.). This shows that the gradient reversal strength is directly correlated with the performance of the model. This is expected as lower gradient reversal strength means a weaker regularisation from the adversary and hence a smaller perturbation from the stand-alone classifier arrangement.

However, as the gradient reversal strength decreases, quite different behaviour is observed in the JSD score between the three training sets. For the models trained on the 5% and 10% correlated training sets, the JSD score is generally flat with decreasing gradient reversal strength. This indicates that as the performance increases through a reduction in regularisation, the adversary remains effective in preventing the classifier from learning the $m_{\gamma\gamma}$ distribution. In contrast, for the model trained on the 20% correlated training set, the JSD does not remain constant with decreasing regularisation but sharply increases. This is especially the case for the smallest regularisation values tested. This indicates that at a particular regularisation threshold, the adversary can no longer stop the classifier from learning the $m_{\gamma\gamma}$ distribution. Beyond this point, the performance remains constant while background sculpting increases.

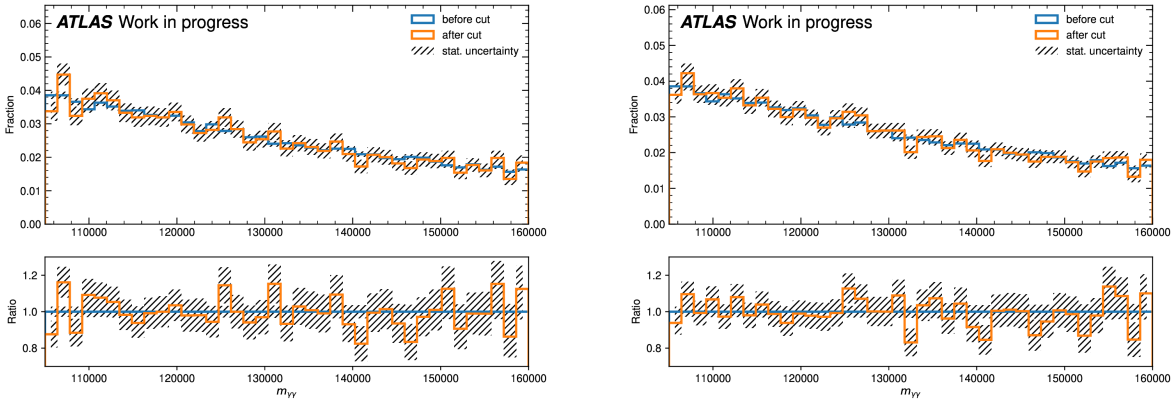
Another interesting result we can see in the plots is that when the ANN is applied to the benchmark training set it is able to achieve a significantly lower background efficiency at equal or lower JSD compared to the benchmark result. This indicates that the ANN is able to make better use of the same data and is hence a more efficient model.

To determine the best ANN model only models whose JSD uncertainty overlaps with the benchmark JSD uncertainty will be considered. From the models that satisfy this criterion, for each training set only the candidate model which has the lowest background efficiency will be considered. Based on this selection the models from the training sets with 10% and 20% correlation show a much lower background efficiency compared to the 5% correlated training set. Finally, between the 10% and 20% correlated training sets, the 10% correlated set is the favoured choice as it shows a significantly more consistent relationship between regularisation strength and JSD score. The same conclusion can be taken from both the inclusive and Higgs JSD plots. Therefore the ANN model parameters chosen from the grid search correspond to a 10% correlation threshold and gradient reversal strength of 0.1875.

3.7.3 Performance of the stand-alone classifier vs the ANN

The chosen ANN from the hyperparameter search achieved a background efficiency of 0.1174. The uncertainty on this value can be assumed to be small as a consistent relationship in background efficiency was seen for the different gradient reversal strengths applied to the 10% correlated training set. Compared to the background efficiency score of 0.165 for the benchmark classifier, the ANN has been able to achieve a 30% reduction. The performance between the two networks over the full signal efficiency range can be summarised by the ROC AUC. The ANN achieves a ROC AUC of 0.9194, while the NN ROC AUC is 0.8937.

The ANN JSD scores measured from the hyperparameter search correspond to $\text{JSD}_{Incl} = 9 \pm 6 \times 10^{-4}$ and $\text{JSD}_{Higgs} = 6 \pm 5 \times 10^{-4}$. To obtain a clearer interpretation of the JSD scores the background distribution for the ANN can be compared before and after the classification. The ratio plot of the background distribution before and after the selection cut can be seen in Figure 12a for the ANN, this can be directly compared with the ratio plot of the NN in Figure 12b.



(a) (b)

Figure 12. The background distributions before and after classification for the ANN in (a) and the NN in (b). The background sculpting of the ANN can be seen to be equivalent to that of the NN.

From these ratio plots, there cannot be seen any sign of background sculpting in the background distribution of either network. The background distribution of the ANN has a less consistent shape but this is largely due to the ANN having lower background efficiency and hence being affected more by statistical fluctuations. Within the current statistical uncertainties, the background ratio plot confirms that the two networks have the same level of background sculpting.

These results confirm that the ANN allows the use of more correlated variables to increase the amount of information available in training while still retaining the level of background sculpting to that of using less correlated variables. This shows that by modifying the training of the benchmark ATLAS classifier through an adversarial network a higher signal purity can be achieved for VBF Higgs events than is possible with the current ATLAS solution.

4 Study 2: Classification between ggF and VBF Higgs events

This study focuses on the classification between the ggF and VBF Higgs production modes, corresponding to the first stage in the ATLAS $H \to \gamma\gamma$ classification procedure. The sys-

tematic uncertainty arising from training classifiers on events simulated through different parton showering generators is studied.

4.1 Systematic uncertainty in modelling of ggF and VBF final states

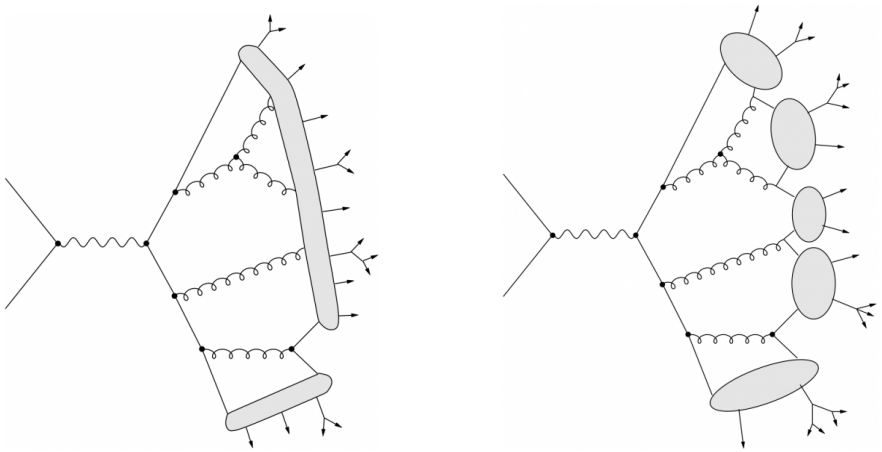
To leading order, in VBF Higgs production the two quarks which radiate the vector bosons go on to produce jets, leaving a Higgs + 2 jets final state [38]. This provides a clear detector signature. On the other hand, to leading order, in ggF Higgs production no other objects are produced in the event apart from the Higgs boson. This leaves the ggF process with no distinct features. However, since the ggF production mode heavily dominates the rate of Higgs production (84%), the frequency of NLO processes becomes significant. In NLO processes the gluons in ggF can radiate secondary gluons. The radiated gluons are individual coloured partons and so they go on to hadronise and produce jets. In many cases, two secondary gluons are radiated which then result in two jets, thereby imitating the detector signature of the VBF LO final state. This process, therefore, makes the distinction between the two production modes heavily reliant on their specific jet characteristics.

In terms of classification, this means that a classifier will need to make use of the fine details in jet kinematics between the two production modes in order to achieve high performance. However, this poses the problem of determining the correct simulation parameters that will be used to generate the samples as these will significantly influence the final jet kinematics. This is a problem because the exact simulation parameters are defined by theory and currently there is no accepted theory of modelling the QCD parton interactions which are responsible for forming jets [13]. Therefore, to determine the dependence in classification performance based on the simulation parameters, a classifier is tested on separate samples generated through different parton shower models.

In ATLAS, the systematic theory uncertainty from the underlying event and parton shower is determined based on testing between a nominal sample showered by PYTHIA and an alternative sample showered by HERWIG [13]. These samples are outlined in Table 1. Through this approach, in the ATLAS analysis, the systematic uncertainty in the total expected VBF cross-section is reported at $\pm$ 16%. This means that testing a classifier between the two samples produces a $\pm$ 16% uncertainty in the total number of VBF events which were accepted in each case. This shows that the small differences in the final state hadrons simulated between the two samples create a significant difference in the number of accepted VBF events. This is an undesirable result as the classification becomes highly dependent on arbitrary nuisance parameters [39].

4.2 The Pythia and Herwig shower models

The two shower models used in ATLAS are based on different methods of hadronization [13]. PYTHIA is based on the string model while HERWIG is based on the cluster model [40]. The representation of these models can be seen in Figure 13.



(a) (b)

Figure 13. The hadronization models used to simulate the training samples. The PYTHIA string model is shown in (a) while the HERWIG cluster model is shown in (b) [41].

The string model is experimentally based on the observation that quark and anti-quark potential rises linearly with the distance between quarks in a meson system [40]. As the quarks in a meson separate, the rising potential between them is converted into new quark-antiquark pairs which results in the formation of a string of hadrons in the angular region. On the other hand, the cluster hadronization model is based on the pre-confinement property of QCD [40]. Pre-confinement suggests that at particular points the parton shower forms colour singlet combinations of partons, called clusters. These clusters of single, coloured partons then go on to form colour-neutral hadrons.

4.3 Methods

The aim of this study was first to independently reproduce the original systematic uncertainty reported by ATLAS. If the uncertainty could be reproduced then the following goal was to determine an effective solution to reduce it.

4.3.1 Benchmark ATLAS uncertainty

As a first stage in this process, the ATLAS ggF and VBF classification results [13] were processed to create reproducible classification categories. In the original analysis, events classified into individual analysis categories are further separated into categories of signal purity based on different classification cuts. For this study, the sub-categories of purity were merged together for each analysis category.

The accepted bin counts for each merged category were then normalised separately for the PYTHIA and HERWIG results. This procedure makes sure that the total counts across the

categories add up to the experimentally measured counts in the $H \to \gamma\gamma$ decay channel. The counts were individually normalised for the ggF and VBF production modes. The experimentally measured counts were determined through the equation:

$$N = L \times \sigma \times BR \times eff, \tag{13}$$

where N is the measured count, L is the integrated luminosity, σ is the production mode cross-section, BR is the $H \to \gamma\gamma$ branching ratio and eff is the trigger efficiency.

The normalised counts in the merged categories were then compared between the PYTHIA and HERWIG models to determine the uncertainty. The uncertainty in bin counts between the two results was determined through the equation:

$$\text{Bin uncertainty} = \frac{\mathrm{N_{PYTHIA}} - \mathrm{N_{HERWIG}}}{\mathrm{N_{PYTHIA}}} \tag{14}$$

where $\mathrm{N_{PYTHIA}}$ are the number of bin counts given by the PYTHIA model and $\mathrm{N_{HERWIG}}$ are the number of counts given by the HERWIG model.

Once the uncertainty in the ATLAS results was determined through this procedure, the uncertainty in the VBF analysis bins was then used as the benchmark to compare with the VBF uncertainty in the independent analysis.

To determine the importance of each bin to the overall VBF uncertainty in counts, the significance of each bin was calculated. This calculation was based on the original ATLAS results which were achieved after considering both the production mode classification and background rejection stages. The significance is given by the equation [13]:

$$S = \sqrt{2 \cdot \left((s+b) \cdot \log\left(1 + \frac{s}{b}\right) - s \right)} \tag{15}$$

where s are the signal counts and b are the background counts for each analysis category respectively. The significance is reported in [13] for the separate analysis categories.The significance for the separate categories was combined in quadrature to obtain the significance for the merged categories that are used in this study.

4.3.2 Independent analysis procedure

The original ATLAS analysis was carried out considering all of the Higgs boson production modes. However, for this study, only the ggF and VBF Higgs production modes are considered which together account for approximately 94% of all Higgs production. No non-resonant background sources are considered throughout this study.

Physics-motivated engineered features from the original ATLAS analysis [13] that were relevant to the ggF and VBF production modes were generated from the event kinematics variables. The engineered features were generated separately for the nominal and alternative data samples. The list of engineered features used in this study can be seen in table 2. In total 23 engineered features are used.

The engineered features for the nominal and alternative samples were processed differently. The engineered PYTHIA features were split into training (56%), validation (14%) and testing (30%) subsets. The mean and standard deviation of the training set were used to standardise all subsets. The engineered HERWIG features were all kept as a testing set, where the set was standardised using its own mean and standard deviation. A network is then trained and validated using the PYTHIA sample. This network is then tested separately on the PYTHIA and HERWIG testing sets.

The predicted bin counts for each model were then processed in the same way as for the benchmark ATLAS results: The counts were normalised and the uncertainty between the counts for each model was determined. The uncertainty in bin counts from the independent analysis will then be compared with the benchmark ATLAS results.

To determine the relationship between performance and the uncertainty between the predictions of the two models, the analysis was performed multiple times for classifiers trained for different numbers of epochs.

4.4 Implementation

The independent analysis procedure is implemented through a multi-class neural network as the classifier. This network was constructed in KERAS 2.9.0. The neural network consisted of 23 input nodes (one for each input feature), one hidden layer of 100 nodes and 26 output nodes (one for each analysis category). The hidden layer uses a ReLU activation function while the output layer uses a softmax activation function. The softmax activation function allows the 26 analysis categories to obtain a probability which sums to 1. The classifier is trained to perform multi-class classification between the event labels by minimising the categorical cross-entropy loss [42]:

$$L_{\text{categorical}} = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{M} y_{ij}\log\left(p_{ij}\right) \tag{16}$$

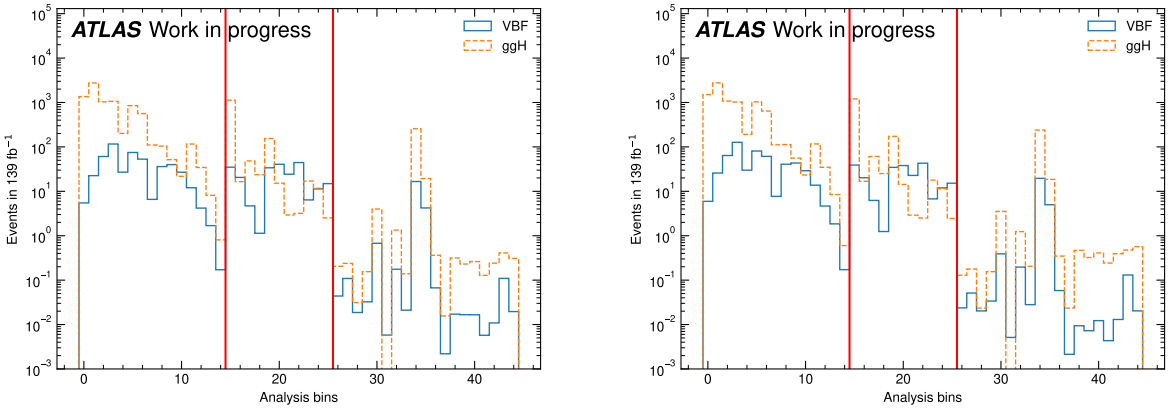
where p_{ij} is the softmax probability and y_{ij} is the one-hot-encoded vector element for the jth class and ith event respectively. The one hot encoded vector contains 1 at the true category position and zeros in all other positions. The sum of the loss function is carried out over all M categories and N events and is then averaged over the events.

The model was then trained with a batch size of 1024 events using the ADAM optimiser. The training epochs were varied to determine how the uncertainty in the predictions changes with different levels of classification performance.

4.5 Results and analysis

4.5.1 Benchmark ATLAS uncertainty

Based on the ATLAS results, the ggF and VBF trigger efficiencies were 0.6863 and 0.6715 respectively. Using Eq. 13, the trigger efficiencies can be combined with the cross sections in Table 1, $H \to \gamma\gamma$ branching ratio of 0.23% and total integrated luminosity of 139 fb^{-1} to give the experimentally measured ggF and VBF counts. This produces ggF and VBF counts of 10507 and 800 respectively, rounded down to the nearest integer. The normalised event counts predicted by models trained on PYTHIA and HERWIG can be seen in Figure 14a and 14b respectively. In these figures event counts are non-zero outside of the ggF and VBF regions as more production mode classification categories were present in the original analysis.



(a) (b)

Figure 14. Normalised analysis category bin counts for (a) the PYTHIA predictions and (b) the HERWIG predictions. Bins from 0 to the first red line correspond to ggF, bins between the red lines correspond to VBF and bins after the second red line correspond to other production modes.

The relative difference in the counts between the two models can then be determined through Eq. 14. This is only carried out over the VBF bins as this is the region of interest. The per bin uncertainties in the VBF region for the ATLAS results can be seen in Figure 15. In the kinematic cuts of the bins, p_{T}^{H} corresponds to the transverse momentum of the Higgs boson. Eq. 15 is used to determine the significance of each bin in the ATLAS analysis. The significance was determined for the merged bins in this analysis by combining the separate significance values in quadrature to give the results seen in Table 3.

From Table 3 it can be seen that one region of kinematic cuts based on the same variables has the highest significance. This means that this kinematic region has the highest impact on the overall number of VBF events that are classified.

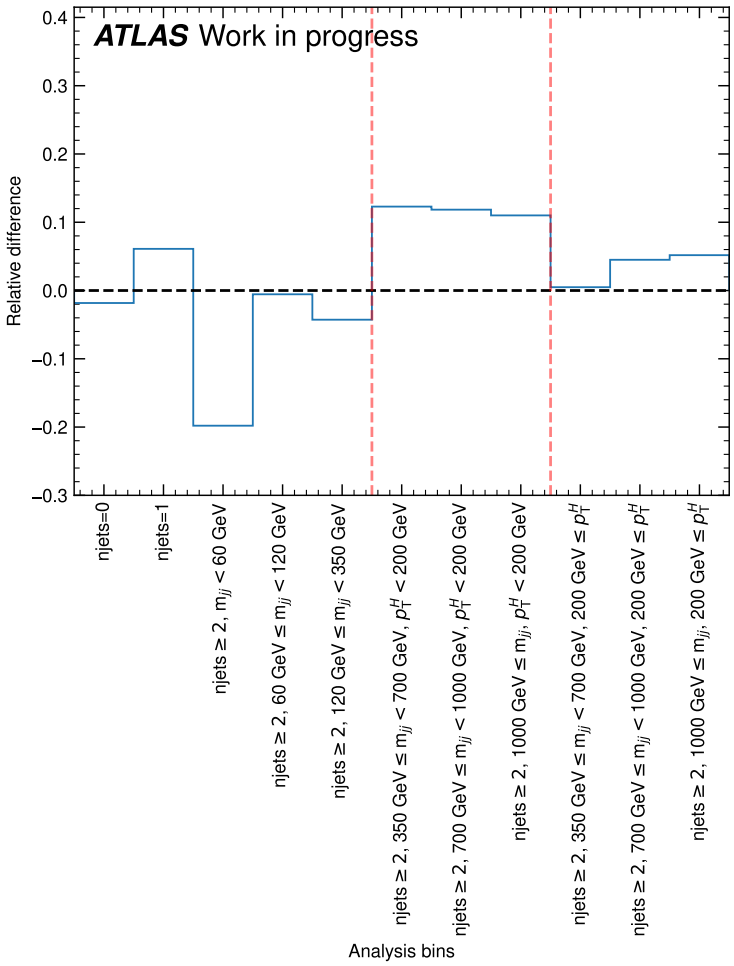


Figure 15. Per-bin relative difference between PYTHIA and HERWIG counts in the VBF analysis bins. The red dashed lines mark the region of the highest significance.

Analysis bin	Significance
njets=0	0.122
njets=1	0.937
njets$\geq$ 2, m$_{jj}$<60 GeV	0.307
njets$\geq$ 2, 60 GeV$\leq$m$_{jj}$<120 GeV	1.509
njets$\geq$ 2, 120 GeV$\leq$m$_{jj}$<350 GeV	0.792
njets$\geq$ 2, 350 GeV$\leq$m$_{jj}$<700 GeV, p_{T}^{H}<200 GeV	2.090
njets$\geq$ 2, 700 GeV$\leq$m$_{jj}$<1000 GeV, p_{T}^{H}<200 GeV	1.862
njets$\geq$ 2, 1000 GeV$\leq$m$_{jj}$, p_{T}^{H}<200 GeV	4.743
njets$\geq$ 2, 350 GeV$\leq$m$_{jj}$<700 GeV, 200 GeV$\leq p_{\mathrm{T}}^{H}$	0.937
njets$\geq$ 2, 700 GeV$\leq$m$_{jj}$<1000 GeV, 200 GeV$\leq p_{\mathrm{T}}^{H}$	1.309
njets$\geq$ 2, 1000 GeV$\leq$m$_{jj}$, 200 GeV$\leq p_{\mathrm{T}}^{H}$	3.450

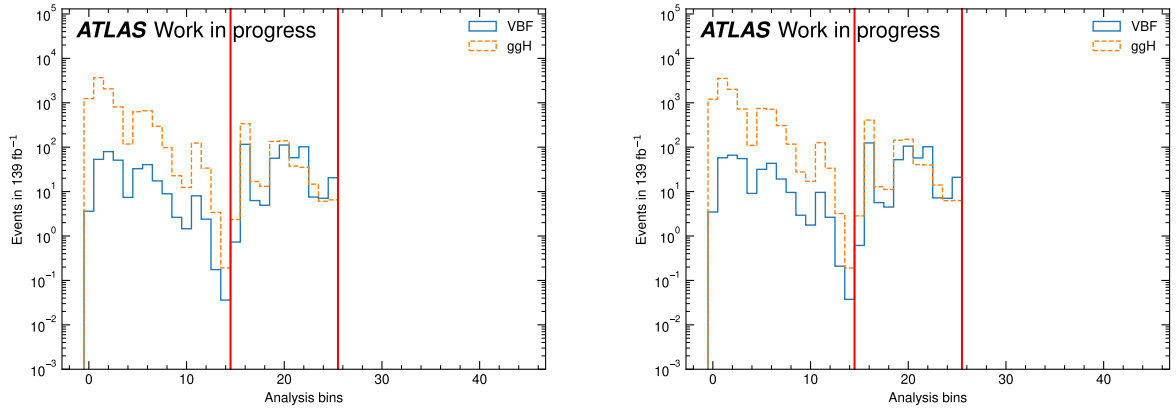
Table 3. The significance of the VBF analysis bins. The red box indicates the most significant kinematic region.

From these results, it can be seen that the bins with the 4 highest significance scores are all shifted in the same direction. Out of these 4 bins, 3 of them correspond to kinematic cuts in the same region. This means the classifier in ATLAS consistently accepts more events for the PYTHIA sample for these bins. Given that these 3 bins form the region of the highest significance and are all shifted positively at a 12% relative difference, this propagates to an approximate 12% uncertainty in the inclusive VBF counts. Therefore, from these results, it can be determined that the current data processing procedure produces results that are consistent with the original ATLAS results. These results can therefore stand as the benchmark to compare with the independent analysis.

4.5.2 Independent analysis uncertainty

A network was trained with the PYTHIA training and validation sets and was then separately tested on the PYTHIA and HERWIG testing sets. This network was trained for 50 epochs as this allowed it to achieve high classification accuracy (approx. 83%).

The normalised ggF and VBF predictions obtained can be seen for the PYTHIA and HERWIG samples in figure 16a and 16b respectively. As only ggF and VBF production modes are considered for the independent analysis, there are no predictions made for the categories outside of these production modes. This is why there are no events predicted for categories after the second red line.



(a) (b)

Figure 16. Normalised analysis category bin counts for (a) the PYTHIA predictions and (b) the HERWIG predictions of the independent analysis. Bins from 0 to the first red line correspond to ggF, bins between the red lines correspond to VBF and bins after the second red line correspond to other production modes.

The relative difference between predictions on the PYTHIA and HERWIG datasets was calculated for the VBF region through Eq. 14. These results can be seen in Figure 14.

In these results, the uncertainty in the most significant region has a similar shape to the result in the ATLAS benchmark. However, the size of the uncertainty for the independent results is on the order of 3%, which is significantly lower than the 12% in the benchmark result.

These preliminary results provide an indication that the independent analysis has not been able to reproduce the original uncertainty reported by ATLAS. However, to confirm this the performance of the classifier needs to be considered. If the original uncertainty were reproducible in the independent analysis, the uncertainty between the predictions on the PYTHIA and HERWIG samples should grow with classifier performance. This is because for the classifier to achieve better performance on the PYTHIA sample it will need to learn the finer details in the final state jet kinematics. This would make the classifier better adapted to the PYTHIA sample. Therefore when applied to classify between the PYTHIA and HERWIG samples, the classifier should show a larger difference in performance for kinematic regions that are simulated differently between the two samples. Therefore to be confident that the preliminary result is not just due to a poorly performing classifier, the relationship between the uncertainty and the classifier's performance is examined.

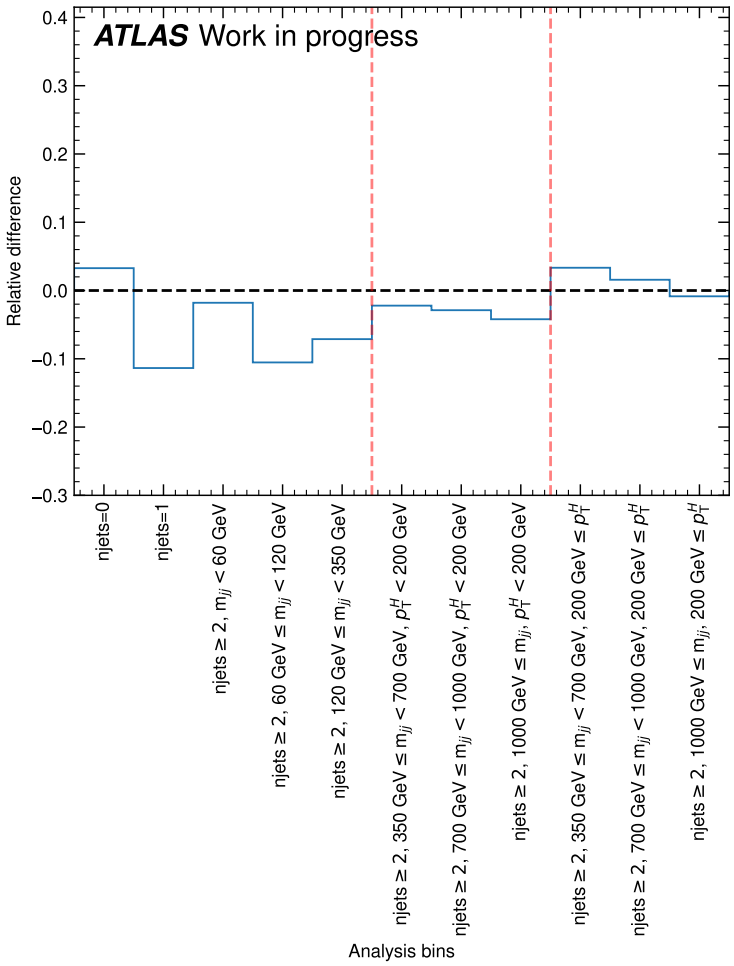


Figure 17. Per-bin relative difference between PYTHIA and HERWIG counts in the VBF region for the independent analysis. The red dashed lines mark the region of the highest significance.

To determine how the uncertainty changes with classification performance, networks were trained for different numbers of epochs. 3 networks were trained for 1, 10 and 100 epochs. The classification accuracy these networks achieved on the validation set was 0.7456, 0.8255 and 0.8386 respectively. The relative uncertainty that was calculated for these networks can be seen in Figure 18a, for the inclusive VBF region, and Figure 18b, for the significant region.

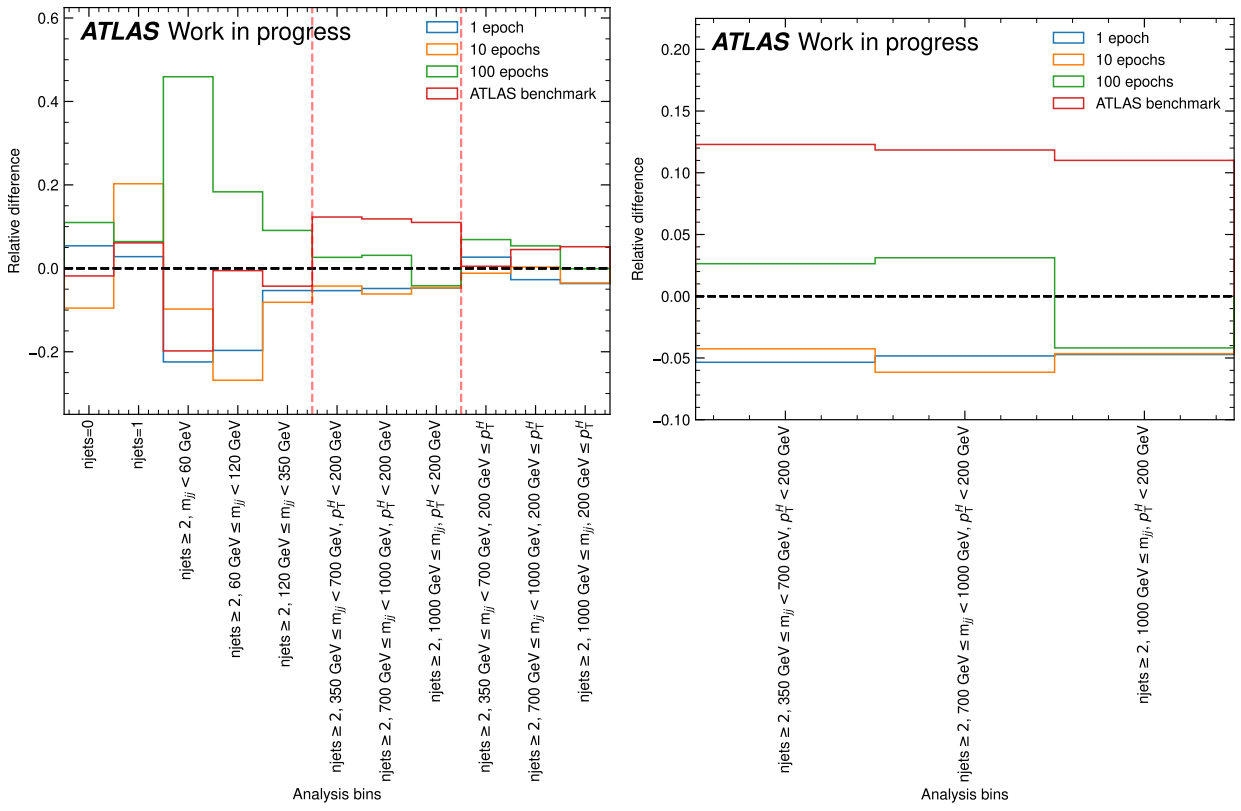


Figure 18. The uncertainty between the PYTHIA and HERWIG predictions with classifiers trained for different numbers of epochs. The inclusive VBF analysis bins can be seen in (a) with the red dashed line corresponding to the most significant region, while the bins only in the most significant region can be seen in (b).

In these results, the relative difference in VBF bins significantly varies with the classification performance. The VBF bins for different performances vary by both size and direction. In these results, no clear trend can be seen to indicate that the uncertainty in the prediction increases with the performance of the classifier. The lack of a trend in combination with the low overall uncertainty confirms that the original benchmark uncertainty cannot be reproduced through the current independent analysis.

These results show that as the classifier becomes more adapted to the PYTHIA sample, it improves its classification between ggF and VBF events through differences in the events that are not exclusively based on their jet kinematics. Hence, in this process, the network does not show a consistent difference in performance between the PYTHIA and HERWIG samples for the differently simulated kinematic jet regions.

The likeliest reason for this is that the original study was simplified too far and the key classification steps which were responsible for the large uncertainty have not been taken into account. These results exclude the sole classification between ggF and VBF Higgs samples as the source of the VBF systematic uncertainty reported by ATLAS. This means that

the original uncertainty must arise by either including more production modes, introducing non-resonant background contamination, or following the original classification steps more closely.

5 Future work

In the first study, only 10% of the low-level detector variables were used to create the engineered features due to storage and computation restrictions. This limited the size of the training data available during the training of the stand-alone classifier and the ANN. The resolution of the results in this study would be significantly improved if the full dataset of engineered features is generated. This would allow for the final conclusions to be more accurate.

During hyperparameter optimisation, the results from training on the 10% correlated training set indicated that there was further room for improvement. Unlike the 5% and 20% correlated datasets, the 10% correlated dataset did not reach a turning point in the background efficiency for the smallest gradient reversal strength that was tested. This means that there is potential for the best ANN model to achieve lower background efficiency at the same level of background sculpting. Therefore, this motivates further hyperparameter optimisation over smaller gradient reversal strengths.

The network architecture and training hyperparameters in this study were largely based on previous studies applied to similar problems and were not explicitly optimised for the application in the study. Given that the performance of both the stand-alone classifier and the ANN significantly depends on these hyperparameters, their optimisation would lead to improvements in both networks. However, there is greater potential for the ANN to improve as it makes use of more hyperparameters. This suggests that optimisation of network architecture and training hyperparameters should further increase the gain in performance by the ANN at the same level of background sculpting.

Combining these implementations indicates that significant improvements in both the confidence of the results and the performance of the ANN are possible.

In the second study, the independent analysis procedure only considers network architectures with different hidden layers of 100 nodes. The hyperparameter search of this study can be extended to make use of hidden layers with varying dimensions. Improvements in the performance of the classifier through this implementation would allow for a more stringent probe of the relationship between classification performance and systematic uncertainty.

To create tighter bounds for the source of the VBF systematic uncertainty, more of the original ATLAS classification pipeline can be included in this study. This can be done by considering more of the Higgs production modes, adding non-resonant background events and introducing a background rejection stage in the classification. These steps can be sequentially added to the current study to closely match the ATLAS environment and determine the point at which the original uncertainty can be reproduced.

6 Conclusion

The first study in this report focused on reducing the modelling uncertainty of accepted non-resonant VBF Higgs events in the invariant diphoton mass distribution. A modelling uncertainty was shown to originate from a neural network exploiting a mass correlation between variables during training. This was shown to result in a distortion of the accepted background mass distribution, where the distribution becomes shaped to look like signal. The current solution for mass decorrelation in the ATLAS $H \to \gamma\gamma$ analysis was implemented, corresponding to training a classifier using only variables which have <5% correlation with the mass. The level of mass correlation exploited by the classifier was measured by comparing the original and accepted background distributions through the JSD metric. The performance of the classifier was measured through the background efficiency at 80% signal efficiency. The JSD and background efficiency obtained through the current ATLAS solution were used as a benchmark for comparison.

An improved solution for mass decorrelation was introduced by regularising the benchmark classifier's training objective through the use of an adversarial network. The training set $m_{\gamma\gamma}$ correlation and the regularisation strength were used as the ANN hyperparameters. These hyperparameters were optimised through grid search. The best hyperparameters were taken from the model which achieved the lowest background efficiency for the same JSD score, within uncertainty, as the benchmark model. Consistent performance within training sets was given priority in the selection. A training set with $m_{\gamma\gamma}$ corr. <10% and regularisation strength of 0.1875 produced the best performance. The ANN trained with these hyperparameters achieved a 30% decrease in background efficiency at the same level of background sculpting as the ATLAS implementation. The results obtained in this analysis provide strong motivation that additional hyperparameter optimisation and training data will produce further gains in performance over the current ALTAS solution.

The second study in this report looked to determine the source of a 16% systematic uncertainty reported by ATLAS in the number of VBF Higgs events classified based on the underlying event and parton shower model. The uncertainty was determined by training a classifier on a nominal dataset and then testing on a nominal and alternative dataset. The nominal and alternative datasets were generated with PYTHIA and HERWIG as the parton shower models respectively. The original ATLAS classification results were processed to create independently reproducible classification categories. The processed classification categories were shown to produce a VBF uncertainty consistent with the original ATLAS uncertainty. This uncertainty is used as a benchmark for the later analysis.

An independent analysis procedure was implemented to reproduce the original uncertainty and look for its source. The independent analysis only considered the classification between the ggF and VBF Higgs production modes which together make up 94% of Higgs production. Classification of the two production modes was carried out over the same classification categories as the ATLAS benchmark result. The systematic uncertainty of the classification was determined through the same procedure used in ATLAS. The results of the independent analysis showed that the sole classification between ggF and VBF events does not reproduce

the benchmark ATLAS uncertainty. This, therefore, excludes the baseline ggF-VBF classification as the source of the uncertainty and points towards the inclusion of more stages in the ATLAS classification pipeline as being responsible.

The combined results of this report, therefore, addressed both the bias of classifying different VBF events based on the simulation parameters and the bias of selecting more background events in the signal region. This directly corresponds to reducing the uncertainty in the first and second stages of the ATLAS $H \to \gamma\gamma$ classification procedure. These results take an important step towards reducing the overall uncertainty in the measurement of the Higgs to vector boson coupling strengths, allowing for a more stringent test of the SM.

References

[1] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B **716** (2012) 1, arXiv: 1207.7214 [hep-ex].

[2] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B **716** (2012) 30, arXiv: 1207.7235 [hep-ex].

[3] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13** (1964) 321.

[4] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13** (1964) 508.

[5] G. Guralnik, C. Hagen and T. Kibble, *Global Conservation Laws and Massless Particles*, Phys. Rev. Lett. **13** (1964) 585.

[6] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19** (1967) 1264–1266.

[7] B.W. Lee, C. Quigg and H.B. Thacker, *Weak Interactions at Very High-Energies: The Role of the Higgs Boson Mass*, Phys. Rev. D **16** (1977) 1519.

[8] V. Barger, K. Cheung. T. Han and R.J.N. Phillips, *Strong W+W+ scattering signals at pp supercolliders*, Phys. Rev. D **42** (1990) 3052.

[9] S. D. Rindani, *Strong gauge boson scattering at the LHC*, 2009, arXiv: 0910.5068 [hep-ph].

[10] U. Schnoor, "Vector Boson Scattering and Electroweak Production of Two Like-Charge W Bosons and Two Jets at the Current and Future ATLAS Detector", Presented 30 Jan 2015, PhD thesis: Dresden, Tech. U., 2014, URL: https://cds.cern.ch/record/2000941.

[11] CMS Collaboration, *Constraints on anomalous Higgs boson couplings to vector bosons and fermions in its production and decay using the four-lepton final state*, Phys. Rev. D **104** (2021) 052004, arXiv: 2104.12152 [hep-ex].

[12] ATLAS collaboration, *A detailed map of Higgs boson interactions by the ATLAS experiment ten years after the discovery*, Nature **607** (2022) 52, arXiv: 2207.00092 [hep-ex].

[13] ATLAS Collaboration, *Measurement of the properties of Higgs boson production at* $\sqrt{s}$ *= 13 TeV in the* $H \to \gamma\gamma$ *channel using 139* fb^{-1} *of pp collision data with the ATLAS experiment*, (2022), arXiv: 2207.00348 [hep-ex].

[14] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003.

[15] ATLAS Collaboration, *Operation of the ATLAS trigger system in Run 2*, JINST **15** (2020) P10004, arXiv: 2007.12539 [hep-ex].

[16] H. Wilkens, *The ATLAS Liquid Argon calorimeter: An overview*, J. Phys.: Conf. Ser. **160** (2009) 012043.

[17] ATLAS Collaboration, *Search for the Standard Model Higgs boson in the diphoton decay channel with 4.9 fb^{-1} of ATLAS data at $\sqrt{s}$ = 7 TeV*, 2011, URL: https://cds.cern.ch/record/1406356.

[18] CERN, *Gluon fusion (ggF)*, URL: https://atlas.cern/glossary/gluon-fusion (visited on 26/03/2023).

[19] CERN, *Vector boson fusion (VBF)*, URL: https://atlas.cern/glossary/vector-boson-fusion (visited on 26/03/2023).

[20] J. Butterworth et al., *PDF4LHC recommendations for LHC Run II*, J. Phys. G **43** (2016) 023001, arXiv: 1510.03865 [hep-ph].

[21] T. Sjöstrand et al., *An introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191** (2015) 159, arXiv: 1410.3012 [hep-ph].

[22] J. Bellm et al., *Herwig 7.1 Release Note*, (2017), arXiv: 1705.06919 [hep-ph].

[23] E. Bothmann et al., *Event generation with Sherpa 2.2*, SciPost Phys. **7** (2019) 034, arXiv: 1905.09127 [hep-ph].

[24] R. D. Ball et al., *Parton distributions for the LHC run II*, JHEP **04** (2015) 040, arXiv: 1410.8849 [hep-ph].

[25] GEANT4 Collaboration, S. Agostinelli et al., *Geant4 – a simulation toolkit*, Nucl. Instrum. Meth. A **506** (2003) 250.

[26] C. Shimmin et al., *Decorrelated Jet Substructure Tagging using Adversarial Neural Networks*, Phys. Rev. D **96** (2017) 074034, arXiv: 1703.03507 [hep-ex].

[27] ATLAS Collaboration, *Electron and photon performance measurements with the ATLAS detector using the 2015–2017 LHC proton-proton collision data*, JINST **14** (2019) P12006, arXiv: 1908.00005 [hep-ex].

[28] ATLAS Collaboration, *Measurement of the inclusive isolated prompt photon cross section in pp collisions at $\sqrt{s}$ = 7 TeV with the ATLAS detector*, Phys. Rev. D **83** (2011) 052005, arXiv: 1012.4389 [hep-ex].

[29] ATLAS Collaboration, *ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset*, (2022), arXiv: 2211.16345 [physics.data-an].

[30] I. Goodfellow, Y. Bengio and A. Courville, *Deep Learning*, MIT Press, 2016.

[31] A. Pouliakis et al., *Artificial Neural Networks as Decision Support Tools in Cytopathology: Past, Present, and Future*, Biomedical Engineering and Computational Biology **7** (2016) S31601.

[32] Keras, *Layer activation functions*, URL: https://keras.io/api/layers/activations/ (visited on 26/03/2023).

[33] ATLAS Collaboration, *Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS*, ATL-PHYS-PUB-2018-014, 2018, URL: https://cds.cern.ch/record/2630973.

[34] Y. Ganin et al., *Domain-Adversarial Training of Neural Networks*, Journal of Machine Learning Research **17** (2016) 1, arXiv: 1505.07818 [stat.ML].

[35] D. Godoy. *Understanding binary cross-entropy / log loss: a visual explanation*. URL: https://towardsdatascience.com/understanding-binary-cross-entropy-logloss-a-visual-explanation-a3ac6025181a (visited on 03/18/2022).

[36] S. Kullback and R. A. Leibler, *On Information and Sufficiency*, Ann. Math. Statist. **22** (1951) 79.

[37] J. Lin, *Divergence measures based on the Shannon entropy*, IEEE Transactions on Information Theory **37** (1991) 145.

[38] S. Gangal and F. J. Tackmann, *Next-to-leading-order uncertainties in Higgs+2 jets from gluon fusion*, Phys. Rev. D **87** (2013) 093008, arXiv: 1302.5437 [hep-ph].

[39] A. Ghosh, B. Nachman, D. Whiteson, *Uncertainty Aware Learning for High Energy Physics*, Phys. Rev. D **104** (2021) 056026, arXiv: 2105.08742 [physics.data-an].

[40] S. Höche, *Introduction to parton-shower event generators*, 2015, arXiv: 1411.4085 [hep-ph].

[41] B. Webber, *Parton shower Monte Carlo event generators*, Scholarpedia **6** (2011) 10662.

[42] S. Saxena, *Binary Cross Entropy/Log Loss for Binary Classification*, URL: https://www.analyticsvidhya.com/blog/2021/03/binary-cross-entropy-log-loss-for-binary-classification/ (visited on 26/03/2023).