School of Physical and Chemical Sciences
Queen Mary University of London

Search for a heavy CP-odd Higgs boson decaying to Zh with the ATLAS detector using a parameterised neural network

Ross Herencia (170301673)

April 8, 2022

Supervisor: Dr Ulla Blumenschein

SPA6776 Extended Independent Project
30 Credit Units

Submitted in partial fulfilment of the requirements for the degree of
BSc Physics from Queen Mary University of London

Declaration

I hereby certify that this project report, which is approximately 10,000 words in length, has been written by me at the School of Physical and Chemical Sciences, Queen Mary University of London, that all material in this dissertation which is not my own work has been properly acknowledged, and that it has not been submitted in any previous application for a degree.

Ross Herencia (170301673)

Acknowledgements

I would like thank Dr Ulla Blumenschein and Tong Qiu for their patience, knowledge and support throughout the project. I would also like to thank Hamza Khan for his many positive contributions to the project and for allowing me to include a discussion with his results.

Abstract

A machine learning-based approach is applied to the search for a neutral CP-odd Higgs boson A decaying into a standard model Higgs boson and a Z boson, with b quark-antiquark and lepton-antilepton final states, respectively. The search is conducted using a parameterised neural network trained to classify signals originating from A bosons in the mass range 300 GeV to 2000 GeV, against a background. Training data is generated from Monte Carlo simulations of the ATLAS detector for pp collisions at centre of mass energy $\sqrt{s} = 13$ TeV and an integrated luminosity of $139\,\text{fb}^{-1}$. The mass m_A is included as a feature to parameterise the neural network and its inclusion is shown to result in a model with the ability to interpolate between values. Furthermore, the separation of signal and background is shown to improve upon a boosted decision tree approach at values of m_A with a lack of data. 95% CL expected upper limits of 0.31 ± 0.03 pb to 8.40 fb are placed on the $gg \to A$ production cross section times the branching ratios of $A \to Zh$ and $h \to b\bar{b}$, with an average improvement of 25% over previously calculated limits in the range 420 GeV to 1000 GeV.

Contents

List of Figures

List of Tables

1 Introduction

This report is a contribution to the search for a CP-odd heavy Higgs boson A, predicted as one of the five Higgs bosons in the two Higgs doublet models (2HDM) [1]. The SM Higgs boson h - discovered by the ATLAS [2] and CMS [3] collaborations at the Large Hadron Collider (LHC) in 2012 - is the only one of the five Higgs bosons to have been observed, with a mass of 125 GeV [4]. The remaining four include a neutral CP-even (H), two charged ($H^{\pm}$) and a neutral CP-odd (A).

Extensions to the SM are currently being explored because they can provide solutions to some of the open questions left by the SM. In particular, 2HDMs can introduce additional sources of CP violation [1], to the extent that the matter-antimatter asymmetry of the universe [5] can be accounted for. Moreover, 2HDMs are able to resolve the naturalness problem [6] surrounding the SM Higgs boson mass.

This search uses a (mass-)parameterised neural network (PNN) classifier [7, 8], trained on Monte Carlo simulation data of the ATLAS detector [9], with the goal of separating signal and background events. Signal events are generated in the $gg \to A \to Zh \to l^+l^-b\bar{b}$ channel and the dominant backgrounds considered here are $t\bar{t}$ and Z+jets. The search is performed for A bosons generated in the mass range $300\,\text{GeV} \leq m_A \leq 2000\,\text{GeV}$, and m_A is included as a feature to parameterise the neural network. Using the classification outputs of the PNN, signal regions (SR) and control regions (CR) are defined to distinguish signal-like from background-like events, and the number of background events in the SR are used to place expected upper limits on the A production cross section $\sigma(gg \to A \to Zh(\to b\bar{b}))$ at 95% confidence level (CL). Related studies have previously been made on the search for A using a binned maximum-likelihood fit [10], and the classification task described here has simultaneously been carried out using boosted decision trees (BDT) [11, 12]. Results will be compared to both the fit-based and BDT approaches.

The remainder of the report consists of five sections. Section 2 provides a brief overview of the SM, including the Higgs mechanism, and a discussion of 2HDMs. Section 3 describes the signal and background processes and includes an explanation of the dataset. An introduction to neural networks and how they are applied to

this search is given in section 4. Results derived from the parameterised neural network are explored in section 5, and comparisons are made to the BDT and binned maximum-likelihood fit approaches. Conclusions are summarised in section 6.

2 Theory

This section provides an overview of the Standard Model (SM) including the introduction of the SM Higgs field through the use of the Higgs Mechanism. Some motivations for extending the SM are provided and a brief introduction to two-Higgs-doublet models is given.

2.1 The Standard Model

The SM is a framework that describes the known fundamental particles and their interactions. The fundamental particles are grouped as fermions and bosons, and the fermions are divided into quarks and leptons, of which there are three generations, as shown in table 2.1. There are six flavours of quarks occurring in *up-type down-type* pairs, and three flavours of lepton lepton-neutrino pairs. For each fundamental fermion there exists an anti-fermion which differs from its counterpart in the values of certain quantum numbers, most notably by having opposite charge. Fermions interact through exchanges of bosons which mediate the fundamental forces: the photon γ mediates electromagnetic (EM) interactions, the gluon g mediates the strong force, and the $W^{\pm}$ and Z bosons mediate the charged and neutral currents of the weak force. More details can be found in [13].

Table 2.1: The fundamental fermions grouped by type and generation. The possible interactions for each type of fermion is also given.

Type	Generation			Charge	Interactions
	1	2	3	Q	
Quarks	u	c	t	+2/3	Strong, Weak
	d	s	b	-1/3	& EM
Leptons	e^-	μ^-	τ^-	-1	Weak & EM
	ν_e	ν_μ	ν_τ	0	Weak

2.1.1 The Higgs Mechanism

The SM is described in relativistic quantum field theory, where the aforementioned particles are fields and the interactions arise as a result of requiring local gauge invariance of the fields' Lagrangian under transformations of the $SU(3) \times SU(2) \times U(1)$ symmetry group [14]. Gauge invariance is achieved by introducing massless 'gauge fields' into the Lagrangian, such that the extraneous terms brought in by the transformation cancel-out. Whilst this works for the electromagnetic and strong forces which have massless gauge bosons, it does not reproduce the massive bosons of the weak force.

The Higgs mechanism combines the concept of local gauge invariance with the idea of spontaneous symmetry breaking [13] to recover the massive weak bosons. As a simple example, consider the complex scalar field

$$\phi = \phi_1 + i\phi_2 \tag{2.1}$$

and the Lagrangian describing ϕ

$$\mathcal{L} = \frac{1}{2}(\partial_\mu \phi)^*(\partial^\mu \phi) - V(\phi^*\phi). \tag{2.2}$$

One choice of the potential $V(\phi^*\phi)$ is

$$V(\phi^*\phi) = -\frac{1}{2}\mu^2(\phi^*\phi) + \frac{1}{4}\lambda^2(\phi^*\phi)^2, \tag{2.3}$$

with $\mu, \lambda \in \mathbb{R}^+$. In this form the Lagrangian is invariant under rotations in ϕ_1, ϕ_2 space but the term of order $\phi^*\phi$ is positive, which would result in a negative Klein-Gordon mass [13]. This is resolved by rewriting the Lagrangian in terms of fluctuations about the minimum of the potential, of which there are an infinite set of solutions defined by

$$\phi_1^2 + \phi_2^2 = \mu^2/\lambda^2. \tag{2.4}$$

This requires choosing a vacuum state, such as $(\phi_{1_{\min}}, \phi_{2_{\min}}) = (\mu/\lambda, 0)$, and defining new fields η and ξ such that

$$\begin{aligned} \eta &= \phi_1 - \phi_{1_{\min}} = \phi_1 - \mu/\lambda, \\ \xi &= \phi_2 - \phi_{2_{\min}} = \phi_2. \end{aligned} \tag{2.5}$$

Substituting the new fields into Eq. (2.2) gives

$$\begin{aligned}\mathcal{L} = \frac{1}{2}\left[(\partial_\mu\eta)(\partial^\mu\eta) - (\mu\eta)^2 + (\partial_\mu\xi)(\partial^\mu\xi)\right] \\ - \left[\mu\lambda(\eta^3 + \eta\xi^2) + \frac{\lambda^2}{4}(\eta^4 + \xi^4 + 2\eta^2\xi^2)\right] + \frac{\mu^4}{4\lambda^2}.\end{aligned} \tag{2.6}$$

The term of order η^2 is the positive mass term $m_\eta = \sqrt{2}\mu\hbar/c$ that has been recovered. This has been acquired at the cost of the rotational symmetry in Eq. (2.2) being broken by the choice of vacuum state, which is referred to as spontaneous symmetry breaking.

Requiring the system to be invariant under a local $U(1)$ transformation

$$\phi \to e^{i\theta(x)}\phi, \tag{2.7}$$

is achieved by introducing a gauge field A^μ in the covariant derivative

$$\partial_\mu \to D_\mu = \partial_\mu + i\frac{q}{\hbar c}A_\mu. \tag{2.8}$$

The resulting Lagrangian describing the fields η, A^μ and ξ is

$$\mathcal{L} = \frac{1}{2}(\partial_\mu\eta)(\partial^\mu\eta) - (\mu\eta)^2 + \frac{1}{2}\left(\frac{q\mu}{\hbar c\lambda}\right)^2 A_\mu A^\mu + \ldots \tag{2.9}$$

where the first two terms describe the scalar field η, which represents the Higgs boson of this toy model, and the third term shows the gauge field A^μ is not massless but has a (Proca [13]) mass term. Additional terms including higher order couplings and terms involving the massless field ξ have been ignored. In fact, the gauge invariance may be exploited to 'rotate away' ξ by choosing θ such that ϕ is real.

2.1.2 The Higgs Field

The Higgs Mechanism may be applied to the SM Higgs field [14], which is a complex doublet with four components

$$\Phi = \begin{pmatrix}\phi^+ \\ \phi^0\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}\phi_1 + i\phi_2 \\ \phi_3 + i\phi_4\end{pmatrix}. \tag{2.10}$$

The Lagrangian describing Φ is

$$\mathcal{L}_{\text{Higgs}} = (D_\mu\Phi)^\dagger(D^\mu\Phi) - V(\Phi^\dagger\Phi), \tag{2.11}$$

where

$$V(\Phi^\dagger\Phi) = \lambda(\Phi^\dagger\Phi)^2 - \mu^2(\Phi^\dagger\Phi) \tag{2.12}$$

and D_μ is the covariant derivative of the electroweak symmetry $SU(2) \times U(1)$. The minimum of the potential

$$(\Phi^\dagger\Phi)_0 = \frac{\nu^2}{2},\ \nu = \frac{\mu}{\sqrt{\lambda}} \tag{2.13}$$

may be chosen such that the vacuum state is

$$\Phi_0 = \begin{pmatrix} 0 \\ \nu/\sqrt{2} \end{pmatrix}. \tag{2.14}$$

Expanding about the vacuum state and using the unitary gauge gives

$$\Phi = \begin{pmatrix} 0 \\ \frac{\nu+h}{\sqrt{2}} \end{pmatrix} \tag{2.15}$$

where h is the Higgs boson.

Using the Higgs Mechanism outlined above, the mass terms of $h, W^\pm$ and Z are found through the $SU(2) \times U(1)$ electroweak symmetry breaking. Moreover, the mass terms of quarks and leptons are subsequently derived from Yukawa couplings [14] to the Higgs field, and it can be shown that the coupling strength is proportional to fermion mass. As a result, Higgs decays to heavier particles are more likely.

2.2 Extensions to the Standard Model

2.2.1 Motivations to Extend the Standard Model

There are many reasons to investigate extensions to the standard model. Providing an explanation to the matter-antimatter asymmetry is one example and a necessary component of this is CP violation [5]. Currently, the known examples of CP violation are not able to account for the asymmetry but, by extending the SM, additional sources of CP violation can be introduced. A second reason is that of resolving the naturalness problem [6]: accounting for the apparent discrepancy over many orders of magnitude between the SM Higgs bare mass and its quantum correction, such that the sum of the two gives the measured mass.

2.2.2 Two-Higgs-Doublet Models

The two-Higgs-doublet models (2HDM) [1] are extensions to the SM Higgs field where there is an additional doublet. Generically, this results in eight degrees of freedom: three relating to the mass terms of the weak bosons and five relating to the massive scalar Higgs fields. Of these five Higgs fields, there are two charged ($H^{\pm}$), two neutral CP-even (H, h) and one neutral CP-odd (A).

2HDMs have a much larger number of free parameters than the SM Higgs field and it is possible to make simplifying assumptions to reduce the amount [15]. This can include assuming that the the Higgs sector is CP-conserving; these models are unable to contribute to the matter-antimatter asymmetry but are still important benchmarks worth investigating. This report focuses on the search for A in a generic CP-conserving 2HDM.

3 The Dataset

The data provided for this search was generated to reflect the interactions and measurements that are expected to be obtained from the ATLAS detector of the Large Hadron Collider at CERN. As a result, this section describes the ATLAS detector, as well as how the interactions which constitute the signal and background were generated. Following a discussion of the event selection process, the features used for training the parameterised neural network are presented.

3.1 The ATLAS Experiment

The Large Hadron Collider (LHC) [16] is a 27 km two-ring synchrotron designed to facilitate proton-proton (pp) beam collisions. The ATLAS detector [9] is a general-purpose detector located at one of four beam-crossing points of the LHC. A right-handed coordinate system is chosen such that the origin is located at the centre of the detector (the nominal interaction point), the z-axis points along the beam path, the positive y-axis points upwards perpendicular to the beam path, and the positive x-axis points towards the centre of the LHC ring. Using this coordinate system, the pseudorapidity η is defined as

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right), \tag{3.1}$$

where θ is the polar angle subtending the z-axis. η provides a Lorentz-invariant measure of a particles position. The angular distance between two particles ΔR^2 is defined as

$$\Delta R^2 = \Delta\eta^2 + \Delta\phi^2, \tag{3.2}$$

where ϕ is the azimuthal angle about the z-axis. This measure of distance is independent of the Lorentz boosts of the particles.

Bunches of protons collide at the nominal interaction point located at the centre of the ATLAS detector with a centre of mass energy $\sqrt{s} = 13\,\text{TeV}$, which is the energy available to create new particles in a collision. Particles which are produced

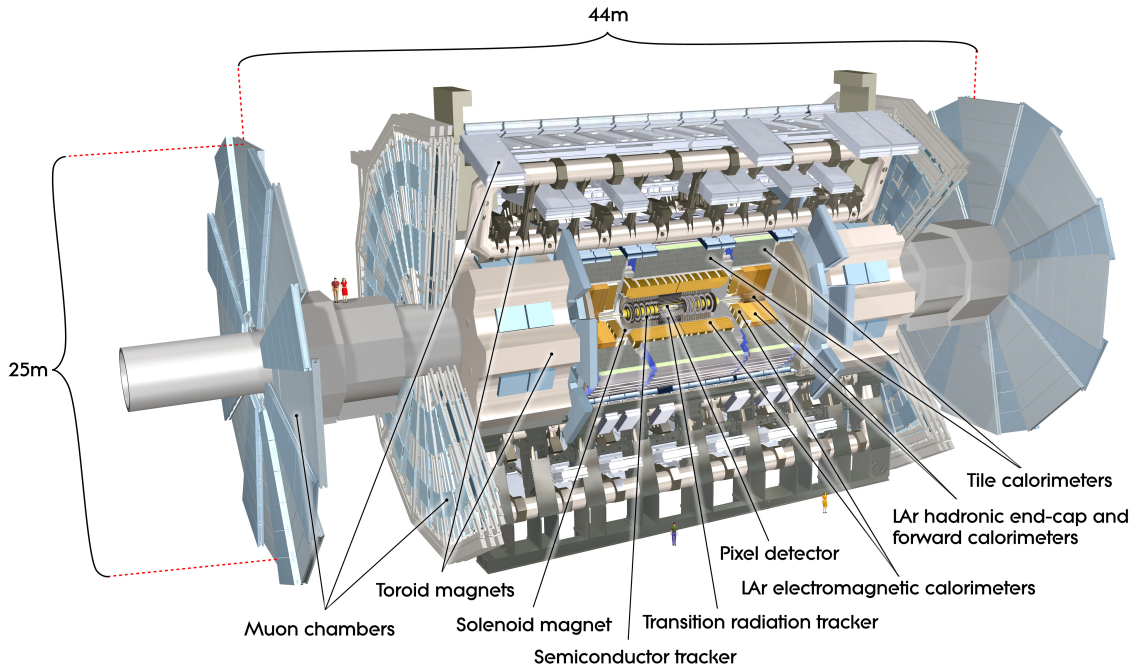


Figure 3.1: The ATLAS detector [9].

by collisions pass through a series of detector subsystems, arranged cylindrically around the nominal interaction point, including: an inner detector for tracking charged particles, electromagnetic and hadronic calorimeters for measuring energy deposits of particle showers, and muon chambers to track muons.

The inner detector is immersed in a 2 T magnetic field (which points along the beam line) and provides tracking for charged particles with $|\eta| < 2.5$. The curvature of a charged particles' trajectory in the magnetic field is proportional to its transverse momentum [17], so high energy charged particles are deflected less by the magnetic field. The electromagnetic calorimeter primarily measures the the energy deposits of electrons and photons which produce EM showers within the calorimeter. Similarly, the hadronic calorimeter measures the energy clusters of hadronic jets (small ΔR^2 cones of hadrons) which produce showers in the calorimeter. Both calorimeters cover the range $|\eta| < 3.2$.

The muon chambers encase the calorimeters and extend pass the end caps of the main body of the detector. Tracking in the muon chambers is provided for muons with $|\eta| < 2.7$.

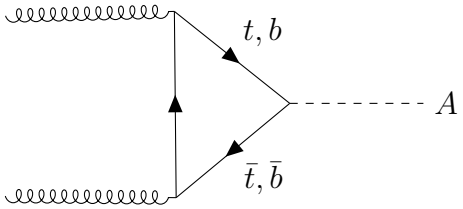


Figure 3.2: A Feynman diagram of the production of the A boson via gluon-gluon fusion.

3.2 Monte Carlo Simulation Data

The dataset provided for this search was generated using Monte Carlo (MC) simulations [18] of the signal and background at the ATLAS detector. The background component of the dataset was generated separately from the signal, which was produced for a discrete set of mass points of A in the range 300 GeV to 2000 GeV. In addition to the columns of measured and reconstructed quantities (referred to as features), each row (called an instance) in the dataset had an associated MC importance weighting which, in general, could be positive or negative. The sum of MC weights for each part of the dataset was normalised to the expected number of events in the 139 fb^{-1} of data collected by ATLAS during Run-2 [19] of the LHC. A signal cross section of 0.05 pb was used for all simulated mass points of A.

3.2.1 The Signal

The signal data was generated in the $gg \to A \to Zh \to l^+l^-b\bar{b}$ search channel. The production of A via gluon-gluon fusion is shown in figure 3.2*. MadGraph5 [21] was used to model the generation of A bosons in the mass range 300 GeV to 2000 GeV.

The decay channel $A \to Zh \to l^+l^-b\bar{b}$ is shown in figure 3.3 and consists of two components: a Z boson decaying into a lepton-antilepton pair, and a SM h boson decaying into a b quark-antiquark pair. Both the z and h decays were generated using PYTHIA8 [22].

3.2.2 The Background

The background consists of dominant processes which have overlapping final states with the signal. For this search, the Z+jets [23] and $t\bar{t}$ [24] backgrounds are of most importance. At the LHC, Z bosons are predominantly produced via the Drell-Yan process [25] with the same lepton-antilepton final states as the signal Z decay. b quark-antiquark pairs are produced from gluons radiated by the initial state quarks.

*All Feynman diagrams have been reproduced using the TikZ-Feynman package [20].

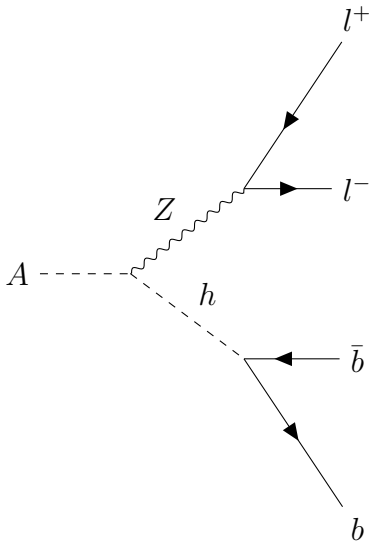


Figure 3.3: A Feynman diagram of A decaying to Z and h, with lepton-antilepton and b quark-antiquark final states, respectively.

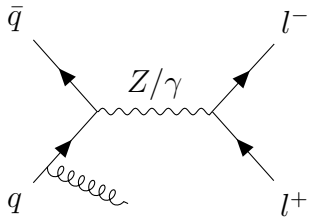


Figure 3.4: A Feynman diagram of the Z+jets background.

Hadronisation of the b quarks leads to the formation of b-jets which overlap with the final state of the SM Higgs decay $h \to b\bar{b}$. The Z+jets background shown in figure 3.4 was generated using the SHERPA generator [26].

$t\bar{t}$ events occur at the LHC through gluon fusion and present a significant background because top quarks can decay weakly to a b quark and a lepton neutrino pair, as shown in figure 3.5. These interactions of $t\bar{t}$ were generated using POWHEG-BOX [27] + PYTHIA8 [22].

3.2.3 Event Reconstruction and Selection

As part of the simulation process, events were selected by the trigger system of the ATLAS detector to obtain data from particular regions of the interest, and to reduce the event rate [9]. Events which pass the trigger system are reconstructed using reconstruction algorithms - details of which can be found in [10]. Electrons are reconstructed by matching tracks in the inner detector to energy deposits in

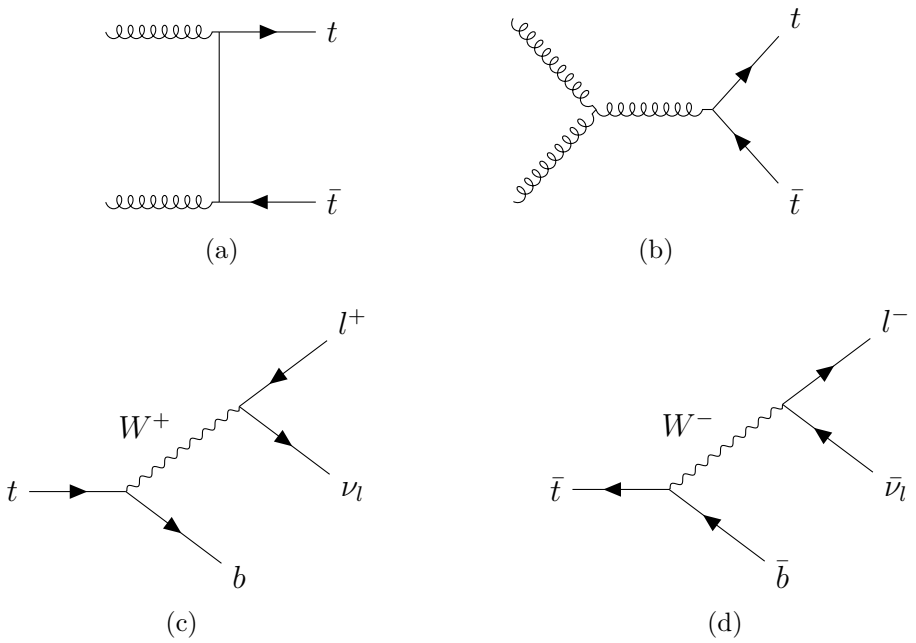


Figure 3.5: Feynman diagrams of the production of the $t\bar{t}$ background via gluon fusion (a) & (b), and the $t\bar{t}$ decays (c) & (d).

the EM calorimeter. Reconstructed muons are required to have tracks in the inner detector, energy deposits in the calorimeters and tracks in the muon chamber. In general, reconstructed lepton tracks are required to be isolated from jets which are identified by energy clusters in the hadronic calorimeter. The b-tagging algorithm MV2c10 [28] was used to distinguish b-jets from c-jets and light jets (which contain neither c-hadrons nor b-hadrons). Neutrinos do not interact with the detector, but their presence can be inferred from missing transverse energy E_T^{miss}, which is the energy required to set the total sum of transverse energy to zero.

Event selection criteria were imposed to further reduce the number of events to those which are most relevant for the search. A summary of the selection criteria is given in table 3.1 which has been reproduced from [10]. The data used for this search was pre-selected to only contain events with lepton-antilepton final states and two resolved b-tagged jets, as these would be expected from the search channel. Merged jets [10], produced by SM Higgs bosons with large momenta decaying to collimated b quarks, were not selected for. Missing energy event selection was carried out using the METHT [29] variable - which is defined as the missing transverse energy divided by the scalar sum in quadrature of electron, muon and jet transverse momenta $\sqrt{H_T}$ - because it provides better cuts than E_T^{miss} alone.

Table 3.1: Event selection criteria for the 2 lepton channel reproduced from [10].

Variable	Selection criteria
Number of jets	≥ 2
Leading jet $p_{T_{j_1}}$ [GeV]	> 45
Dijet invariant mass m_{jj} [GeV]	$100 < m_{jj} < 145$
Leading lepton $p_{T_{l_1}}$ [GeV]	> 27
Sub-leading lepton $p_{T_{l_2}}$ [GeV]	> 20
METHT $= E_T^{\text{miss}}/\sqrt{H_T}$ [$\sqrt{\text{GeV}}$]	$< 1.15 + 8 \times 10^{-3} \cdot m_{Zh}/(\text{GeV})$
Dilepton $p_{T_{ll}}$ [GeV]	$> 20 + 9\sqrt{m_{Zh} - 320/(\text{GeV})}$, for $m_{Zh} > 320$GeV
Dilepton invariant mass m_{ll} [GeV]	$\max[40, 87 - 0.030 \cdot m_{Zh}/(\text{GeV})] < m_{ll} < 97 + 0.013 \cdot m_{Zh}/(\text{GeV})$

3.3 The Features of the Dataset

The final dataset was comprised of 20 features and approximately 4×10^5 signal instances and 1.7×10^6 background instances. A full list of features is given in table 3.2. The inclusion of a third jet is done to account for additional jets, such as those produced by gluon radiation.

The distributions of four features are shown in figure 3.6 for three mass points of A against the background, all of which have been normalised to the full Run-2 integrated luminosity. Figures 3.6(a) and 3.6(b), show the reconstructed masses of A and Z, respectively. For both features the background overlaps with the signal across the entire mass range. The peak around the Z boson mass $m_Z \approx 91\,\text{GeV}$ [4] in (b) demonstrates the dominant presence of the Z+jets background. Figures 3.6(c) and 3.6(d) show the leading b-jet and leading lepton momenta. For these features the separation between signal and background can be seen to improve for higher values of m_A.

The background was found to overlap similarly for the remaining 16 features. The lack of any clear separation between signal and background suggests that a simple cut on one of the features would be insufficient. This is the main motivation behind exploring alternative methods. The fact that, for certain features, the background is less overlapping at higher masses, suggests that separation should improve with m_A.

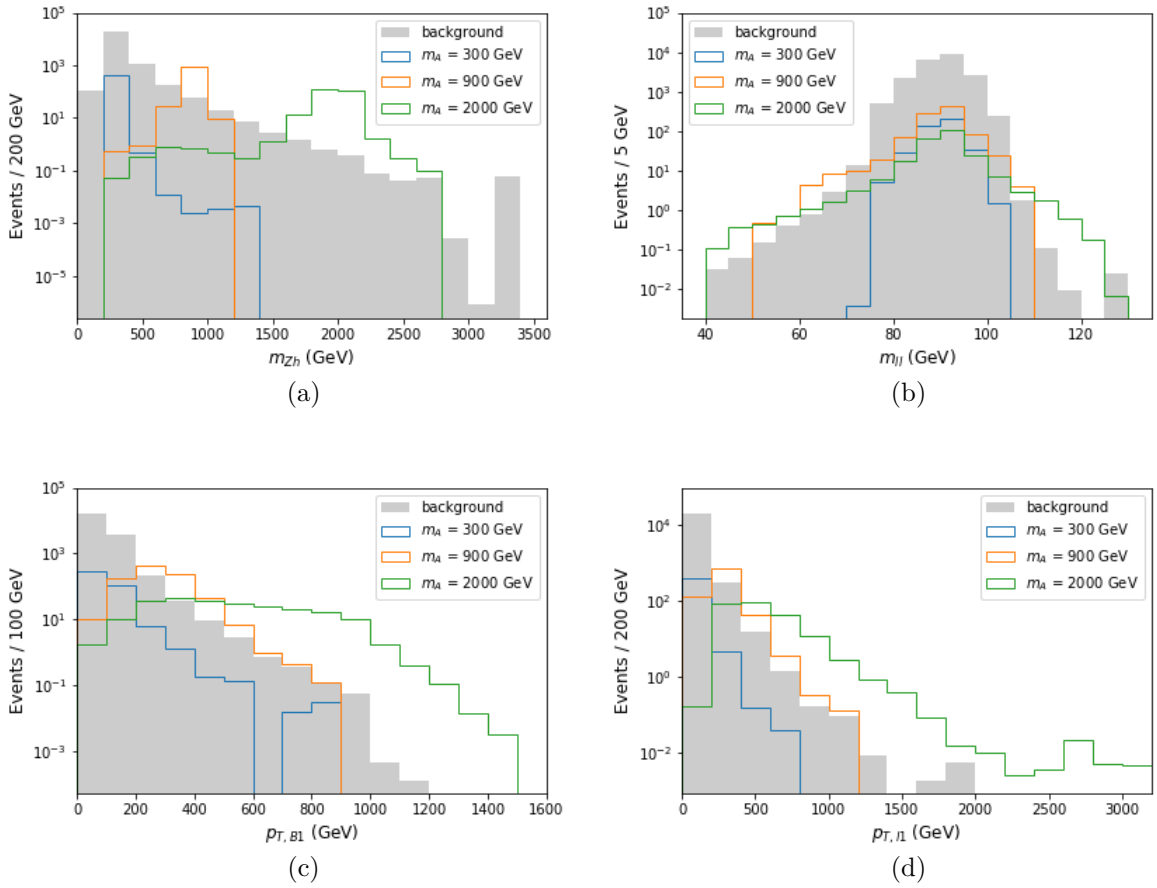


Figure 3.6: Distributions of the signal for $m_A = \{300, 900, 2000\}$ GeV against the background, for four features of the dataset. The distributions are normalised to the expected number of events from the full Run-2 luminosity of 139 fb^{-1} times a signal cross section of 0.05 pb. The features include the reconstructed mass of the A boson (a), the reconstructed Z mass (b), the leading b-jet momentum (c), and the leading lepton momentum (d).

Table 3.2: The 20 features of the dataset which have been used to train the parameterised neural network.

Input Feature	Description
m_{bb}, m_{Zh}, m_{ll}	Reconstructed masses of the lepton pair, b quark pair and Z, h mother particles.
$P_{T,l_1}, P_{T,l_2}, P_{T,B_1}, P_{T,B_2}, P_{T,h}$, $P_{T,Z}, P_{T,J_3}$	Measured transverse momenta of the leptons, B mesons and third jet; reconstructed transverse momenta of Higgs and Z bosons.
METHT	A measure of the missing transverse energy.
$\Delta\eta_{ll}, \Delta\phi_{ll}, \Delta R_{B_1,B_2}$, $\Delta R_{B_1,J_3}, \Delta R_{B_2,J_3}$	Pseudorapidity and azimuthal angle of lepton pair; angular distance between B mesons, and B mesons and the third jet.
MV2c10B1, MV2c10B2, MV2c10B3	B-tagging algorithm outputs.
m_A	Mass of the A boson.

4 Machine Learning

Machine learning (ML) [7] is an iterative process that enables a computer program to learn from data without intermediary user input. This section provides a brief description of neural networks, with a focus on binary classifiers, followed by discussions on parameterisation and the implementation of a parameterised neural network in this search.

4.1 Neural Networks

A neural network (NN) [7] is an example of a ML model which is highly customisable and able to learn complicated patterns in data. NNs are typically composed of a collection of artificial neurons which are Threshold Logic Units (TLU) [7]. A TLU performs a weighted sum of input data $\mathbf{x}$ and outputs a single value given by

$$h_{\mathbf{w}}(\mathbf{x}) = f(\mathbf{x}^T\mathbf{w}), \tag{4.1}$$

where the components of $\mathbf{w}$ are weights and f is a step function. A NN is simply an extension to this concept, where multiple TLUs are connected to the input data in parallel - creating a Perceptron [7] - and then layered in series, creating a Multilayer Perceptron (MLP) [7]. The output of a single layer is

$$h_{\mathbf{W},\mathbf{b}}(\mathbf{X}) = f(\mathbf{X}\mathbf{W} + \mathbf{b}), \tag{4.2}$$

where $\mathbf{X}$ and $\mathbf{W}$ are matrices. $\mathbf{X}$ has a column for each neuron of the previous layer and a row for each instance, and $\mathbf{W}$ has a column for each neuron in the layer and a row for each neuron in the previous layer. The bias term $\mathbf{b}$ is a vector of weights for a neuron that always outputs 1 and is necessary to shift the activation function f, which is a non-linear function so that complex patterns can be modelled.

A NN contains an input layer, an output layer, and one or more so-called hidden layers, each with a configurable number of neurons. The activation function is generally unchanged from one hidden layer to the next and some examples discussed

in this report include the ReLU [7] and SELU [30] functions

$$\mathrm{ReLU}(z) = \max(0, z), \tag{4.3}$$

$$\mathrm{SELU}(z) = \lambda(\max(0, z) + \min(0, \alpha e^z - \alpha)), \tag{4.4}$$

where λ, α are predefined constants.

The output layer activation function varies depending on the task and, since the search is concerned with binary classification, the example discussed here is the sigmoid function

$$f(z) = \frac{1}{(1 - e^z)}, \tag{4.5}$$

which outputs a value between 0 and 1; a value closer to 0 is background-like, and closer to 1 is signal-like.

Supervised NNs [7] are trained using labelled data. The label tells the NN the value it should have output for a given instance. Training consists of iteratively updating the model weights, such that the difference between an output x and its corresponding label y is minimised. In practice this is achieved using the backpropagation algorithm [31] and a loss function - in this case, the binary cross entropy (or log loss [7])

$$L(x, y) = -w_{MC}[y \log(x) + (1 - y) \log(1 - x)], \tag{4.6}$$

where, for the application discussed here, w_{MC} is taken as the absolute value of the MC weight for a particular training instance. The backpropagation algorithm works as follows. A forward pass is made through the NN, computing and storing the outputs of each layer, and the loss associated with each instance is calculated. Then, a backwards pass is made through the NN to identify which model weights to update in each layer in order to minimise the loss. Following this, another forward pass performs the necessary updates to the model weights via a gradient descent [7] step

$$\mathbf{W}_{r+1} = \mathbf{W}_r - \eta \nabla_{\mathbf{W}} L(\mathbf{W}) \tag{4.7}$$

for a given step in the training process r. The learning rate η [7] controls the size of the update to the model weights.

Datasets are typically split into training, validation and test sets, and the training set is further split into batches. Model weights are updated after every batch, and one pass through all batches is called an epoch. After each epoch the loss can be calculated for the validation set and compared to the training set. Ideally both

losses will decrease and be converging to a similar value. If the training loss decreases whilst the validation loss increases, this suggests that the model is overfitting the training set, which is referred to as overtraining [7]. Overtraining can in part be mitigated by randomly shuffling the training set so that it is less likely that the model will only 'see' one type of data (such as background) too often. Comparisons between candidate models are made using the validation set, so that the test set can be reserved for a final measure of a chosen models' performance, without introducing bias into the model selection.

4.2 Implementation of the Parameterised Neural Network

The implementation of the parameterised neural network was done in Python 3.7 using the PyTorch package [32]. All of the code produced for this search is available through GitHub: `github.com/Ross-Herencia/HiggsML`.

4.2.1 Improving Neural Network Performance

The PyTorch package includes additional tools and modifications which can be used to extend upon the main concepts of NNs discussed in section 4.1. Some of these have been included for this search, with the aim of improving performance and reducing the amount of time required to train a model.

Stochastic gradient descent (SGD) [7] was implemented with momentum [7] and Nesterov acceleration [33, 34], using the SGD optimiser from PyTorch. This is a modified version of Eq. (4.7) given by,

$$\boldsymbol{\theta}_{r+1} = \boldsymbol{\theta}_r + \mathbf{m}_r \tag{4.8}$$

where

$$\mathbf{m}_r = \beta \mathbf{m}_{r-1} - \eta \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_r + \beta \mathbf{m}_{r-1}). \tag{4.9}$$

$\boldsymbol{\theta}$ represents the model weights, and the momentum vector $\mathbf{m}$, along with the friction term β, control how quickly a model can reach a minimum of the loss function. The greatest benefit is gained from the momentum when the gradient is a constant d for

many consecutive steps because taking $\mathbf{m}_{r-1} = \mathbf{m}_0 = -\eta d$ gives

$$\begin{aligned} \mathbf{m}_1 &= \beta \mathbf{m}_0 - \eta d = -\beta \eta d - \eta d \\ \mathbf{m}_2 &= \beta \mathbf{m}_1 - \eta d = -\beta^2 \eta d - \beta \eta d - \eta d \\ &\dots \end{aligned} \tag{4.10}$$

which is a geometric series and, for $|\beta| < 1$, can be written as

$$\mathbf{m}_r = -\eta d \frac{1}{1-\beta}. \tag{4.11}$$

Therefore, a value of $\beta = 0.7$ implies the potential for a 3 times faster descent than would be obtained from Eq. (4.7). Nesterov acceleration shifts the point at which the gradient is calculated from the local position $\boldsymbol{\theta}$, to one which is ahead in the direction of the momentum, $\boldsymbol{\theta} + \beta \mathbf{m}$. As long as the momentum vector is pointing in the direction of the minimum, a better measure of the gradient is obtained, and this can also reduce the number of steps required for convergence. The stochastic approximation means that only one random instance from each batch is used to determine the weight update, which is essential to reduce training times with large datasets.

The learning rate η can be varied during training by using a learning rate scheduler - a technique that can improve performance and convergence time [7]. The exponential scheduler from PyTorch multiplies η by a constant γ every epoch. Choosing $\gamma < 1$ reduces the learning rate so that later epochs have a smaller impact on the model. This type of scheduler was chosen for its simplicity and for its ease of optimisation (see sec. 4.2.3).

4.2.2 Parameterisation of a Neural Network

Parameterisation of a NN is the inclusion of additional parameters as features of the dataset and has been shown to improve a models' ability to interpolate [8]. The reproducibility of this technique is of particular interest because the mass of the A boson is unknown; since it would be implausible to train many models over a continuous mass range, the ability for a single model, trained over a discrete set of simulated mass points, to be able to classify events corresponding to intermediate masses equally well, is highly desirable. For this reason, the parameter m_A was included as one of the 20 features listed in table 3.2 and with this inclusion the NN is referred to as a (mass-)parameterised neural network (PNN).

For each part of the signal data - in the training, validation and test sets - the value of m_A was set to the corresponding mass that was used to generate the data, for all instances. For the background, m_A was assigned randomly to one of the signal masses, for each instance. validation and test sets were also produced where the background was paired with data for one signal mass point, and m_A was set to match the signal mass, for all instances. This avoids the scenario where the model simply cuts on differing values of m_A when testing the classifier output at a specific mass point. The $m_A = 1200\,\text{GeV}$ signal data was withheld from the training and validation sets in order to test the models' ability to interpolate, and will be referred to as the blind mass. Because of the wide variation in features, each was scaled to a normal distribution with a mean of zero and unit variance. A split size of $60 : 20 : 20$ was used for the training, validation and test sets, respectively.

4.2.3 Hyperparameter Optimisation and the Discovery Significance

Hyperparameters [7] are parameters of the ML algorithm which are set before training. For the PNN this includes the number of hidden layers, the number of neurons per layer, the batch size, η, γ, β and the activation function. Hyperparameters impact a models' performance because they control how it learns from data, and it is generally necessary to optimise hyperparameters.

The hyperparameters of the PNN were optimised manually by choosing a set of values for each parameter and trying combinations of them. Specific values or combinations of values which produced the best performing models on the validation set were identified, and subsequent sets of values with finer intervals were explored. Early stopping [7] was implemented to halt the training of models which did not decrease the validation set loss for a set number of consecutive epochs (typically $\sim 10\%$ of the maximum).

The batch size was kept constant at 500 instances per batch because it has been shown that the convergence of a NN is dependent on the learning rate to batch size ratio [35]. The computational time of an epoch is dependent on the batch size but not on the learning rate, therefore it makes sense to choose a batch size depending on computational limitations, and vary the learning rate freely. With 500 instances and a training split of 60%, approximately 2500 SGD steps were expected per epoch. The momentum was also kept constant at $\beta = 0.7$ to decrease the number of hyperparameter combinations and to avoid conflict with the learning rate. Results of the hyperparameter optimisation process are given in section 5.

Determining the best hyperparameters, and consequently the best model, requires a metric of a models performance. Since the classifier output of the PNN for each instance is a value between 0 and 1, the output can be binned and the sum of MC weights in the ith bin for both the signal (s_i) and background (b_i) can be calculated. Using this, it is possible to calculate the discovery significance [36]

$$\sigma = \sqrt{\sum_{i=1}^{N} 2\left[(s_i + b_i)\ln\left(1 + \frac{s_i}{b_i}\right) - s_i\right]}, \tag{4.12}$$

where N is the total number of bins. σ increases with the ratio s_i/b_i and therefore, bins with large overlap should not be expected to contribute much to the significance. As a result, the significance provides a measure of the separation between signal and background, whilst the loss measures the quality of the classification. The convention of summing over i inside the square root was an ad hoc addition made for this particular application.

Three edge cases had to be considered when using Eq. (4.12):

1. $b_i = 0$: The average background MC weight over all bins was used because it should be a small but representative value which preserves the separation.

2. $s_i/b_i = 0$: The contribution from the bin was set to zero.

3. $s_i/b_i < -1$: Since only the background contained negative MC weights, the absolute value of b_i was taken.

Case 3 is likely the most controversial and indeed, there are alternatives, such as taking absolute value of the background MC weights, or removing then entirely. Considering the $(s_i + b_i)$ term, it can be seen that case 3 gives the most conservative value because b_i still contains contributions from the negative weights. The alternative methods were tested with both the signal and background weights normalised to 1, and it was found that the increase in significance across all mass points was $\leq 4\%$, compared to case 3. Therefore, for this dataset and normalisation scheme, case 3 is valid.

4.3 Applications in High Energy Physics

ML has previously been used in high energy physics (HEP) across a variety of applications [37], including signal classification. The use of NNs for classification is a relatively recent technique and a common alternative is the boosted decision

tree (BDT). Decision trees use a series of nodes with binary decisions derived from features in the dataset to separate data into classes [7]. Boosting a decision tree is the iterative process of re-weighting the dataset, running it through the decision tree and taking an weighted average of all previous classifier outputs [38]. An example of a BDT used in HEP would be the aforementioned b-tagging algorithm MV2c10 [28] which classifies b, c and light jets.

The search for A was also performed [11] using the XGBoost [12] BDT algorithm. Comparisons between the BDT and PNN approaches are presented in section 5.3.

5 Results

In this section the optimisation and performance of the parameterised neural network is discussed and a comparison is made to a boosted decision tree approach on the same classification task and with the same dataset. Expected upper limits on the A production cross section are presented and compared to the boosted decision tree and binned maximum-likelihood approaches.

5.1 Hyperparameter Optimisation and Preliminary Results

The hyperparameter optimisation process is outlined briefly in table 5.1, with the ranges and sets of values tested, as well as what was found to be the optimal set. Although not every single combination was tested, a large majority of combinations were covered, and values which contributed to the largest significances were thoroughly investigated.

The results suggested that the number of hidden layers and the number of neurons per layer did not have a large impact on the significance, within the range tested. In general, larger models were more computational intensive and took longer to train. Therefore, smaller values for the number of neurons per layer and the number hidden layers were chosen preferentially to keep training times down. The learning rate η and decay factor γ were found to only affect the number of epochs required to converge and did not lead to a variety of different solutions. Figure 5.1 shows the loss as a function of epoch for three sets of hyperparameters, including what was taken to be the optimal set (c). The larger learning rates allow the model to achieve a smaller loss in fewer epochs, at the cost of validation loss becoming less stable. Models trained with $\eta = 0.1$ and $\eta = 1$ were found to achieve a significance 18% greater than models trained with $\eta = 0.01$. This motivated the training of models with $\eta \in [0.5, 1.5]$ and lead to the optimal learning rate given in table 5.1.

The ReLU activation function can result in dead neurons [7] when the weighted sum of inputs for a neuron becomes negative, meaning that only 0 is output. The

Table 5.1: The selection of hyperparameters considered in the optimisation of the parameterised neural network and the range of values tested. The optimal set of hyperparameters is also given, which all models presented in this report have been trained with.

Hyperparameter	Values	Optimal
Hidden layers	range = [1, 6], interval = 1	2
Neurons per layer	range = [50, 400], interval = 50	50
Learning rate η	set = {0.01, 0.1, 1} range = [0.5, 1.5], interval = 0.1	1.4
Decay factor γ	set = {0.9, 0.99, 0.999}	0.999
Activation function	ReLU, SELU, Leaky ReLU & PReLU	ReLU

lack of variation in the significances motivated a test for the presence of dead neurons in the PNN. Alternative activation functions including SELU (see Eq. (4.4)), Leaky ReLU [7] and Parametric Leaky ReLU (PReLU) [7] - which do not suffer from dead neurons - were used to train models using the optimal hyperparameters. The latter two activation functions are given by:

$$\text{Leaky/P ReLU}_{\alpha}(z) = \max(az, z), \tag{5.1}$$

where $\alpha = 0.01$ is the default leakage in PyTorch, and in PReLU, α is parameter which the model learns during training. The significance did not increase for any of alternative activation functions suggesting that either the PNN did not contain dead neurons or, that their presence was negligible.

Figure 5.2 shows the PNN classifier output of the test set using a model trained with the optimal set of hyperparameters. The output is divided into 20 bins and each bin contains s_i total MC weight for the signal and b_i total weight for the background. For the classifier outputs the MC weights were normalised such that the sum over s_i and b_i for all i was equal to 1, respectively. Moving from figure 5.2(a) to 5.2(d), separation of signal and background can be seen to improve with increasing mass. Except at the lowest mass point $m_A = 300\,\text{GeV}$, the quality of the classifier output for the signal is typically good, with the majority of signal weight being classified in or above the 0.5 bin, meaning more signal-like. To improve the separation further, it was considered necessary to improve the background classification.

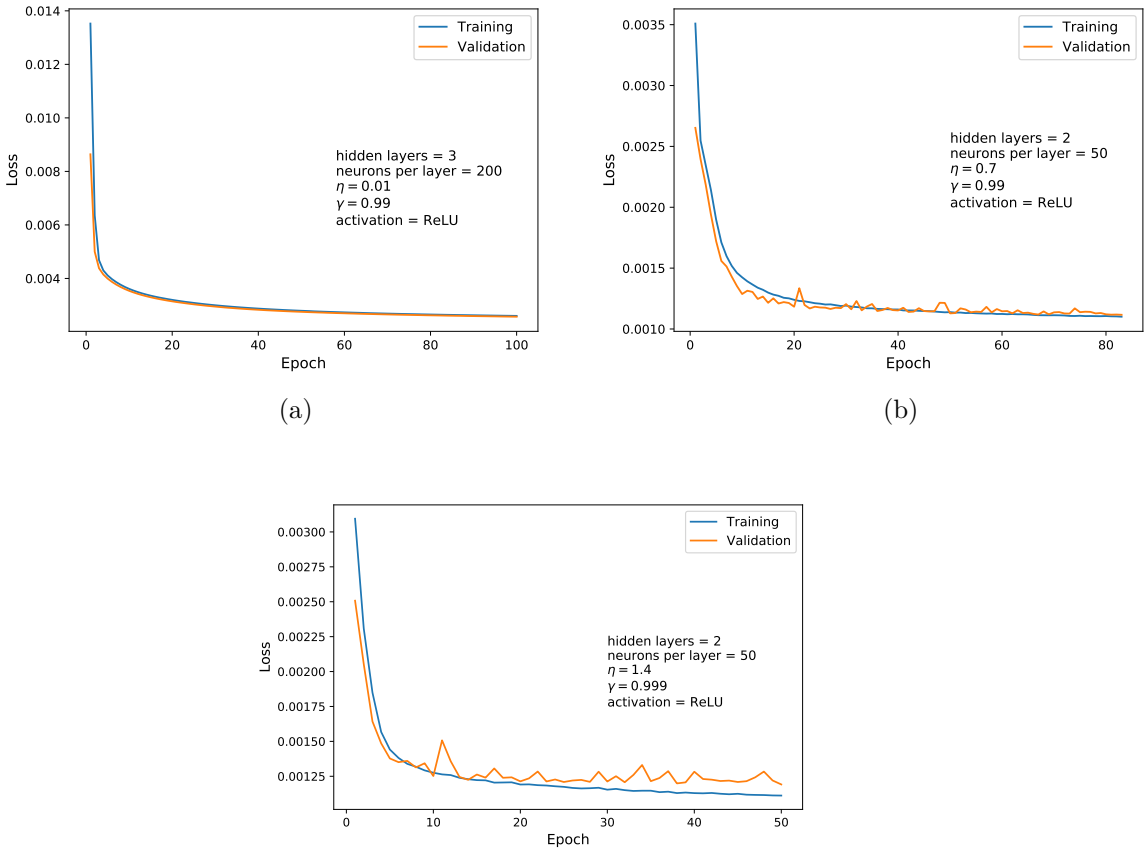
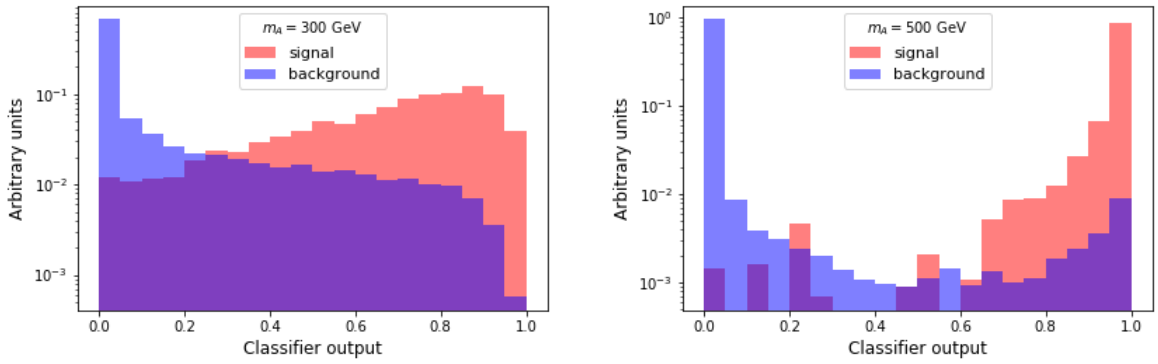
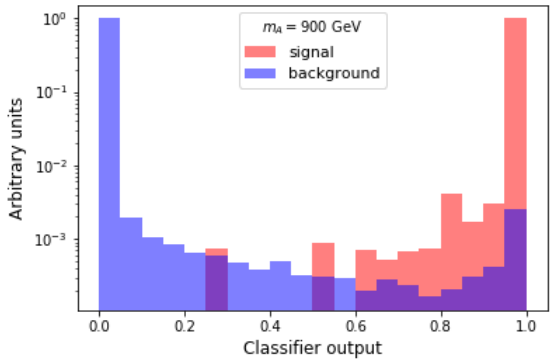


Figure 5.1: Training and validation loss as a function of epoch for three different sets of hyperparameters. Plot (c) shows the loss curves for the optimal set of hyperparameters.

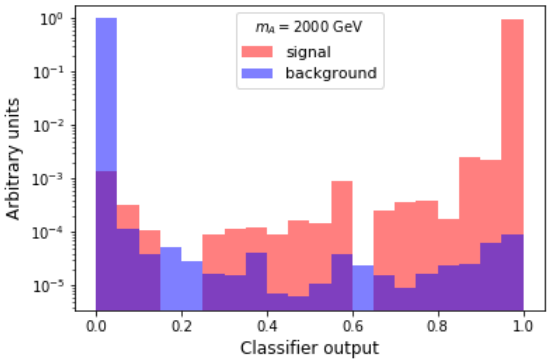


(a)

(b)



(c)



(d)

Figure 5.2: The classification output of the test set for a model trained with the optimal set of hyperparameters. The signal and background MC weights of the test set are both normalised to one.

5.2 A Rescaling of the Training MC Weights for Signal Data

An unconventional solution was identified to aid in the separation of signal and background. Increasing or decreasing the relative scaling between the sum of MC weights for the signal and the background in the training set, was found to impact the significance of the validation set. Considering the binary cross entropy loss function given in Eq. (4.6), two cases can be identified as

$$L = \begin{cases} -w_s \log(x), & \text{for signal events,} \\ -w_b \log(1-x), & \text{for background events,} \end{cases} \tag{5.2}$$

where w_s and w_b represent the MC weight of a signal and background instance in the training set. Ignoring the complication of momentum, the model weights of the PNN are be updated according to

$$\boldsymbol{\theta}_{r+1} = \begin{cases} \boldsymbol{\theta}_r + w_s \eta \nabla_{\boldsymbol{\theta}} L, & \text{for signal events,} \\ \boldsymbol{\theta}_r + w_b \eta \nabla_{\boldsymbol{\theta}} L. & \text{for background events.} \end{cases} \tag{5.3}$$

If on average $\bar{w_s} = k\bar{w_b}$, for $k > 0$, the model weight update size is scaled by a factor of k for signal instances, relative to background events. Therefore, the impact of a signal instance on the model is increased or decreased, depending on whether k is greater or less than 1.

Figure 5.3 shows the significance as a function of mass, derived from the validation set classifier outputs, for several relative scaling factors. A different model was trained for each value of k using the optimal set of hyperparameters. The output previously shown in figure 5.2 is equivalent to a model where $k = 1$ and was used as the benchmark for improvement. A factor of $k = 10$ resulted in a decrease in significance across all mass points whereas, $0.1 \leq k \leq 0.01$ increased the significance for $m_A \geq 400\,\text{GeV}$. $k = 0.001$ improves the significance at higher masses but underperforms at low masses. The model trained with $k = 0.01$ results in the largest consistent increase in significance for $m_A > 500\,\text{GeV}$ which was identified as a key area in this search. Consequently, this model was used for all of the following results.

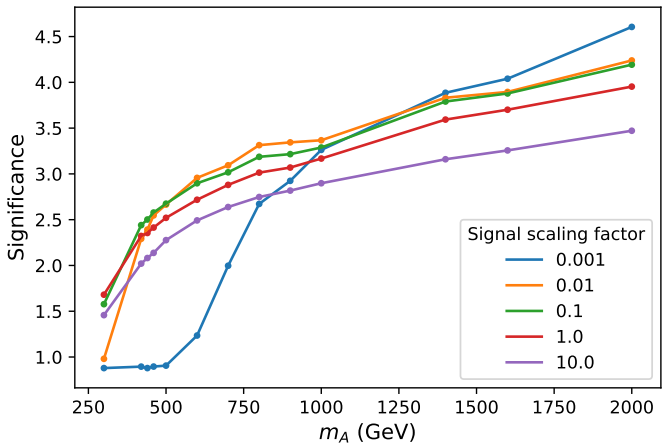


Figure 5.3: Significance as function of mass for scaling factors of training set signal MC weights relative to the training set background MC weights. Connecting lines are drawn as a guide to the eye for visualisation purposes.

5.3 PNN Results and Comparison to a BDT

The PNN classifier output of the model trained with $k = 0.01$ is shown in figure 5.4. Again, the sum of s_i and b_i over i is normalised to 1. Figures 5.4(a), 5.4(b) and 5.4(d) all show a reduction in background weight populating the higher bins by at least a factor of 10. This suggests that the background classification has improved. However, the improvement has come at the cost of the signal classification worsening and spreading out over more bins.

Figure 5.4(c) shows the classifier output for the blind mass point $m_A = 1200\,\text{GeV}$, for which all data was held-out until testing. The output is consistent with the trained mass points suggesting that the model has successfully interpolated between the $m_A = 1000\,\text{GeV}$ and $m_A = 1400\,\text{GeV}$ data.

Figure 5.5 shows that the significance as a function of mass for the test set, as well as the results obtained using a BDT approach [11]. The significance of the blind mass is approximately what would be expected of a trained mass, given the surrounding data, and matches the BDT which was trained on data for $m_A = 1200\,\text{GeV}$. Below the blind mass the significance is similar for the PNN and the BDT. The improvement of the PNN for higher masses is thought to be a result of PNNs ability to interpolate, whereas the BDT struggles because of the lack of data compared to low mass points.

In summary, the results produced by the PNN suggest that the mass parameterisation does improve a neural networks ability to interpolate, and is in agreement with what was previously published [8]. This implies that a PNN of the sort dis-

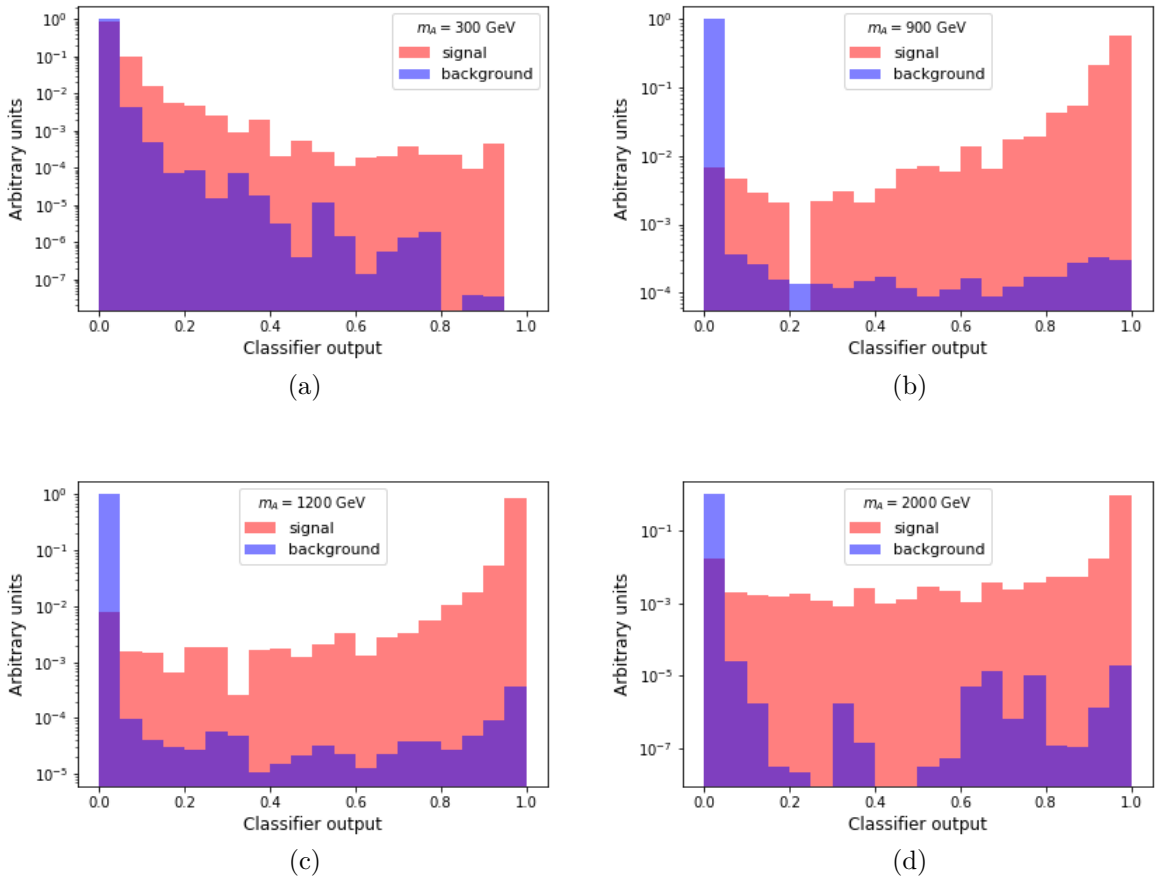


Figure 5.4: The classifier output of the test set for a model trained with the training set signal MC weights scaled by a factor of $k = 0.01$ relative to the training set background MC weights. The signal and background MC weights of the test are both normalised to one.

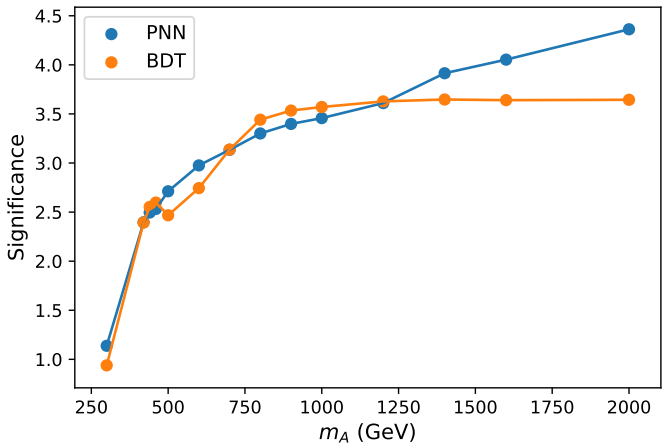


Figure 5.5: Significance as a function of m_A for the parameterised neural network (PNN) and a boosted decision tree (BDT) [11]. Lines connecting data points are drawn as a guide to the eye.

cussed here should generalise well when applied to real data. Whether the most optimal set of hyperparameters was found is unclear and it may be of interest to apply a more sophisticated process to the optimisation.

5.4 Expected Upper Limits

Expected upper limits on the on the A production cross section $\sigma(gg \to A \to Zh) \times BR(Zh \to b\bar{b})$ were derived from the classifier outputs using the CLs method [39]. The validation and test sets were combined and normalised to the full Run-2 luminosity of 139 fb^{-1}, and the classifier output was split into a signal region (SR) and control region (CR). The number of background events in the SR was used to determine an upper limit on the expected number of signal events in the SR. Given the number of simulated signal events in the SR, and using the simulated signal cross section of 0.05 pb for all masses, an expected upper limit on the signal cross section was determined. The SR was optimised in the range [0.1, 1.0] at each m_A by choosing the SR resulting in the smallest expected cross section. The optimal cuts for the beginning of the SR are given in table 5.2.

Figure 5.6 shows the expected upper limits on the A production cross section at 95% CL for the PNN, BDT and the binned maximum-likelihood fit performed by the ATLAS Collaboration [10]. The branching ratio BR($Z \to l^+l^-$) = 10.08% [4] was removed from the sum of MC weights in order to be comparable to the ATLAS result. An ideal consideration of the uncertainties would require accounting

Table 5.2: The optimal signal region cuts which minimised the central values of the expected upper limits on the A production cross section at 95% CL. The signal region is defined such that it contains all events classified with a value $\geq$ the cut.

m_A (GeV)	Optimal signal region cut
300	0.10
420	0.15
440	0.25
460	0.35
500	0.35
600	0.65
700	0.75
800	0.85
900	0.90
1000	0.95
1200	0.95
1400	0.85
1600	0.95
2000	0.10

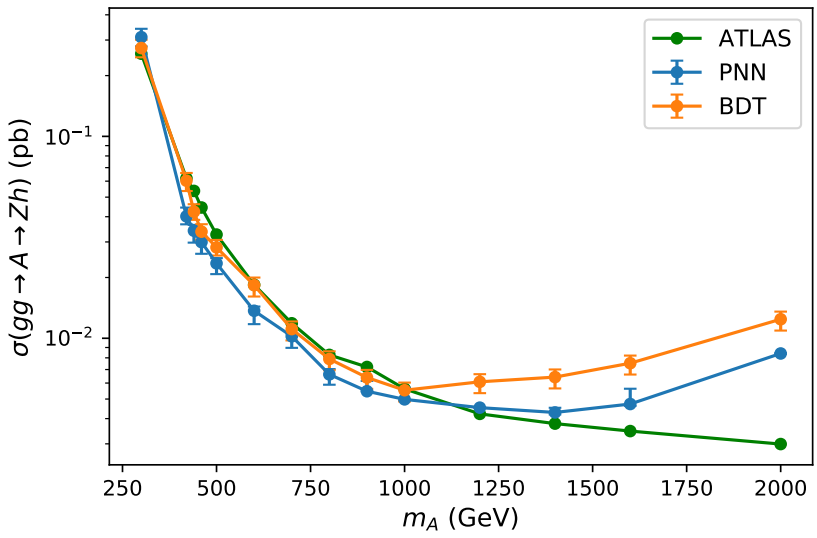


Figure 5.6: 95% CL expected upper limits on the production cross section $gg \to A$ times the branching ratios of $A \to Zh$ and $h \to b\bar{b}$. Results are shown for the PNN, BDT [11] and binned maximum-likelihood fit [10] performed by the ATLAS Collaboration.

for many systematic uncertainties [10]. For this search the main contributions to the uncertainty were taken as the two-b-tag efficiency ($\approx$15%) [40] and the Monte Carlo modelling uncertainty of the $t\bar{t}$ and Z+jets backgrounds ($\approx$15%) [23, 24]. Summation in quadrature provides an estimated uncertainty of 20% and this was accounted for by fluctuating the number of background events in the SR by $\pm$20%.

For the PNN, the expected upper limits vary from 0.31 ± 0.03pb for $m_A = 300$GeV to 8.4fb for $m_A = 2000$GeV. The minimum occurs for $m_A = 1400$GeV with a cross section of $4.29^{+0.24}_{-0.09}$fb. The PNN is more sensitive than the BDT, except at $m_A = 300\,\text{GeV}$, which was understood to be a weak point of the model from figure 5.3. An improvement has been made over the ATLAS result from 420 GeV to 1000 GeV. This is in spite of only using a SR with two b-tagged jets, whereas ATLAS made use of one and two b-tagged jets. The reduction in sensitivity at higher masses for both the PNN and BDT is thought to be caused by a lack of data for merged jets, which were not included in the pre-selection (as discussed in section 3), but were included for ATLAS. With the inclusion of events with one b-tagged jet and events with merged jets, it is thought that the PNN should improve over previous results across the entire mass range.

6 Conclusions

A parameterised neural network has been applied to the search for a heavy CP-odd Higgs boson A in the Monte Carlo simulated search channel $gg \to A \to Zh \to l^+l^-b\bar{b}$, with A bosons generated in the mass range 300 GeV to 2000 GeV . The mass-parameterisation of the neural network using m_A was shown improve upon the separation of signal and background over a boosted decision tree, specifically at masses where there was a lack of data. The two approaches were found to perform equally well at a mass of A for which the parameterised model had not been trained on any corresponding data.

95% CL expected upper limits on the A production cross section $\sigma(gg \to A \to Zh) \times BR(Zh \to b\bar{b})$ have been determined from the classification outputs of the parameterised neural network. The expected limits were found to vary from 0.31 ± 0.03 pb for $m_A = 300$ GeV to 8.4 fb for $m_A = 2000$ GeV with a minimum of $4.29^{+0.24}_{-0.09}$ fb at $m_A = 1400$ GeV. This improves upon previous results using a binned maximum-likelihood fit [10] by an average of 25% in the mass range 420 GeV to 1000 GeV despite a lack of data for jets with one b-tag, or merged jets.

Bibliography

[1] Gustavo Castelo Branco et al. Theory and phenomenology of two-higgs-doublet models. *Physics reports*, 516(1-2):1–102, 2012.

[2] The 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*, 716:1–29, 2012. doi: 10.1016/j.physletb.2012.08.020.

[3] CMS Collaboration. Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC. *Phys. Lett. B*, 716:30–61, 2012. doi: 10.1016/j.physletb.2012.08.021.

[4] P. A. Zyla et al. Review of Particle Physics. *PTEP*, 2020(8):083C01, 2020. doi: 10.1093/ptep/ptaa104.

[5] A. D. Sakharov. Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe. *Pisma Zh. Eksp. Teor. Fiz.*, 5:32–35, 1967. doi: 10.1070/PU1991v034n05ABEH002497.

[6] A. Rodrigues Vieira et al. Naturalness and Theoretical Constraints on the Higgs Boson Mass. *Int. J. Theor. Phys.*, 52:3494–3503, 2013. doi: 10.1007/s10773-013-1652-x.

[7] Aurélien Géron. Hands-on machine learning with scikit-learn and tensorflow: Concepts, tools, and techniques to build intelligent systems, 2nd edition. O'Reilly Media, Inc., 2019. ISBN: 9781492032649.

[8] Pierre Baldi et al. Parameterized machine learning for high-energy physics. *arXiv preprint arXiv:1601.07913*, 2016.

[9] Georges Aad et al. The atlas experiment at the cern large hadron collider. *Journal of instrumentation*, 3(S08003), 2008.

[10] ATLAS Collaboration. Search for heavy resonances decaying into a Z boson and a Higgs boson in final states with leptons and b-jets in 139 fb^{-1} of pp collisions at $\sqrt{s} = 13TeV$ with the ATLAS detector. *ATLAS-CONF-2020-043*, 2020.

[11] Hamza Khan. Boosted decision trees in the search for a heavy CP-odd Higgs boson decaying to Zh in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Queen Mary University of London, 2022.

[12] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, p. 785–794, 2016.

[13] David Griffiths. Introduction to Elementary Particles 2nd Edition. Wiley, 2008. ISBN: 9783527406012.

[14] Michael E. Peskin and Daniel V. Schroeder. An Introduction to quantum field theory. Addison-Wesley, 1995. ISBN: 9780201503975.

[15] Hannah Arnold. Search for a CP-Odd Higgs Boson Decaying to Zh in pp Collisions at $\sqrt{s} = 13$ TeV and Development of a b-Jet Tagging Calibration Method for c Jets at the ATLAS Experiment. Fribourg U., 2018. URL http://cds.cern.ch/record/2692202.

[16] Lyndon Evans and Philip Bryant. LHC Machine. *JINST*, 3:S08001, 2008. doi: 10.1088/1748-0221/3/08/S08001.

[17] David Griffiths. Introduction to electrodynamics fourth edition. Cambridge University Press, 2017. ISBN: 9781108420419.

[18] Torbjörn Sjöstrand. The PYTHIA Event Generator: Past, Present and Future. *Comput. Phys. Commun.*, 246:106910, 2020. doi: 10.1016/j.cpc.2019.106910.

[19] J. T. Boyd. LHC Run-2 and Future Prospects. 2019 European School of High-Energy Physics, 1 2020.

[20] Joshua Ellis. TikZ-Feynman: Feynman diagrams with TikZ. *Comput. Phys. Commun.*, 210:103–123, 2017. doi: 10.1016/j.cpc.2016.08.019.

[21] J. Alwall et al. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014. doi: 10.1007/JHEP07(2014)079.

[22] Torbjorn Sjostrand et al. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008. doi: 10.1016/j.cpc.2008.01.036.

[23] ATLAS Collaboration. Modelling and computational improvements to the simulation of single vector-boson plus jet processes for the ATLAS experiment. 12 2021.

[24] Michał Czakon et al. Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_S^4)$. *Phys. Rev. Lett.*, 110:252004, 2013. doi: 10.1103/PhysRevLett.110.252004.

[25] Sidney D. Drell and Tung-Mow Yan. Massive lepton-pair production in hadron-hadron collisions at high energies. *Phys. Rev. Lett.*, 25:316–320, Aug 1970. doi: 10.1103/PhysRevLett.25.316. URL https://link.aps.org/doi/10.1103/PhysRevLett.25.316.

[26] Enrico Bothmann et al. Event Generation with Sherpa 2.2. *SciPost Phys.*, 7 (3):034, 2019. doi: 10.21468/SciPostPhys.7.3.034.

[27] Simone Alioli et al. A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX. *JHEP*, 06:043, 2010. doi: 10.1007/JHEP06(2010)043.

[28] ATLAS Collaboration. Measurements of b-jet tagging efficiency with the ATLAS detector using $t\bar{t}$ events at $\sqrt{s} = 13$ TeV. *JHEP*, 08:089, 2018. doi: 10.1007/JHEP08(2018)089.

[29] Jeffrey Wayne Hetherly. Using VH Associated Production to Search for the $b\bar{b}$ Decay of the Higgs Boson with Data from the ATLAS Detector at $\sqrt{s} =$ 13TeV. University of North Texas, 2017. URL https://cds.cern.ch/record/2313140.

[30] Andrinandrasana David Rasamoelina et al. A review of activation function for artificial neural network. 2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI) p. 281–286, 2020.

[31] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[32] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 32 p. 8024–8035, 2019. Curran Associates, Inc.

[33] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/k^ 2). Dokl. akad. nauk Sssr **269**, p. 543–547, 1983.

[34] Aleksandar Botev et al. Nesterov's accelerated gradient and momentum as approximations to regularised update descent. 2017 International Joint Conference on Neural Networks (IJCNN) p. 1899–1903, 2017.

[35] Stanislaw Jastrzębski et al. Width of minima reached by stochastic gradient descent is influenced by learning rate to batch size ratio. Artificial Neural Networks and Machine Learning – ICANN 2018, Springer International Publishing p. 392–402, 2018.

[36] Glen Cowan and Eilam Gross. Discovery significance with statistical uncertainty in the background estimate, 2008. URL http://www.pp.rhul.ac.uk/~{}cowan/stat/notes/SigCalcNote.pdf.

[37] Dimitri Bourilkov. Machine and deep learning applications in particle physics. *International Journal of Modern Physics A*, 34(35):1930019, 2019.

[38] Andreas Hoecker et al. Tmva-toolkit for multivariate data analysis. *arXiv preprint physics/0703039*, 2007.

[39] Alexander L. Read. Modified frequentist analysis of search results (The CL(s) method). Workshop on Confidence Limits p. 81–101, 2000. CERN-OPEN-2000-205.

[40] ATLAS Collaboration. Search for heavy resonances decaying into a vector boson and a Higgs boson in final states with leptons and b-jets in 139 fb^{-1} of $\sqrt{s} = 13TeV$ pp collisions with the ATLAS detector. *ATLAS-CONF-Note*, 2022.