Optimizing a classification neural network for selecting events from Higgs-boson production in vector-boson fusion in the decay mode $H \to \tau\tau \to e\mu 4\nu$ at the ATLAS experiment

Ye Joon Kim

Supervised by:

Prof. Dr. Markus Schumacher

A Thesis Presented for the Degree of
Bachelor of Science

Institute of Physics
Albert-Ludwigs-Universität Freiburg
20.08.2020

Contents

1 Introduction

The Standard Model of particle physics is a theory that describes forces and particles on a microscopic scale [1][2][3]. A major part of the Standard Model is the Higgs mechanism, which allows for a consistent description of massive particles [4][5][6]. A consequence of the Higgs mechanism is the prediction of the existence of the Higgs-boson.

In 2012, the CMS and ATLAS collaborations detected a particle with a mass of approximately 125 GeV. In 2013 it was confirmed that this particle had properties consistent to that of the Higgs-boson predicted by the Standard Model [7].

However a question arose on whether one could more effectively differentiate signals from the Higgs-boson from other background processes that may leave similar signatures in the detectors. The precision of this measurement depends on the signal-to-background ratio s/b and the significance $s/\sqrt{s+b}$, where s and b are the number of signal events and background events classified as signal events, respectively. Methods in optimizing these values came with the increase in computing power of modern computers. The use of machine learning has especially proven to be a powerful tool in solving a wide range of problems.

An approach in machine learning is the use of artificial neural networks (NN). Artificial neural networks were initially developed to mirror the behavior of biological neural networks in animal brains. Although the function of modern neural networks differ significantly from those biological, the common foundation is that the system learns from examples over time, and uses this information to make predictions.

This thesis explores the use of artificial neural networks to discriminate VBF$\rightarrow H \rightarrow \tau\tau \rightarrow e\mu 4\nu$ decay signals from background processes. The influence of different hyperparameters such as the learning rate, L2 parameter, batch size, network architecture and input variables on the performance of the neural network are investigated.

Possible further investigations utilizing results of this thesis may be precise measurements of the cross section of VBF Higgs production or testing CP invariance in this sector using $H \rightarrow \tau\tau$ events.

2 Theoretical background

2.1 The Standard Model

The Standard Model of particle physics is a theory that describes three of the four fundamental forces of nature: electromagnetism, strong and weak forces, as well as fundamental particles and interactions between them. Particles are classified into four categories: leptons, quarks, gauge bosons and the Higgs-boson. Since its formulation in the 1970's it has seen exceptional success in tests and predictions, predicting the existence of particles such as the top quark [8] and tau neutrino [9], which were experimentally confirmed many years later.

2.2 The Higgs-boson

The Higgs mechanism allows a consistent description of massive particles. The Higgs mechanism postulates the existence of a Higgs-boson, whose mass is not predicted by the Standard Model. The mass of the Higgs-boson is therefore a free parameter of the Standard Model which had to be determined experimentally. The Higgs-boson was predicted to have an even parity and spin 0.

The energy of the LHC allowed the production of Higgs-bosons, and in 2011, CERN observed a particle with a mass of 125 GeV, which had properties consistent with the predictions of the Standard Model. In 2012 it was confirmed that CERN had indeed observed the Higgs-boson [7]. The mass of the Higgs-boson was measured to be 125.18 ± 0.16 GeV [10], with a spin of 0, neutral electric charge, and even parity, consistent with the predictions of the Standard Model.

2.3 Higgs-boson production modes

The four leading production modes of the Higgs-boson in pp-collisions at the LHC are gluon fusion (ggF), vector boson fusion (VBF), Higgs-Strahlung (VH), and Higgs production in association with a pair of top quarks ($\bar{t}tH$). The theoretical Higgs-boson production cross sections as a function of the mass of the Higgs-boson for different production modes can be seen in figure 1. Example leading order Feynman-Diagrams of these processes can be seen in figure 2. The predicted production cross section for each of these production modes for a center of mass energy of $\sqrt{s} = 13$ TeV and a Higgs-mass of 125.09 GeV can be found in table 1.

In this thesis, VBF production is considered as the signal process, where a quark or an anti-quark scatters with another quark or an anti-quark by emitting a W or a Z boson each, which emits a Higgs-boson. This is characterized by two jets in the forward and backwards region [13].

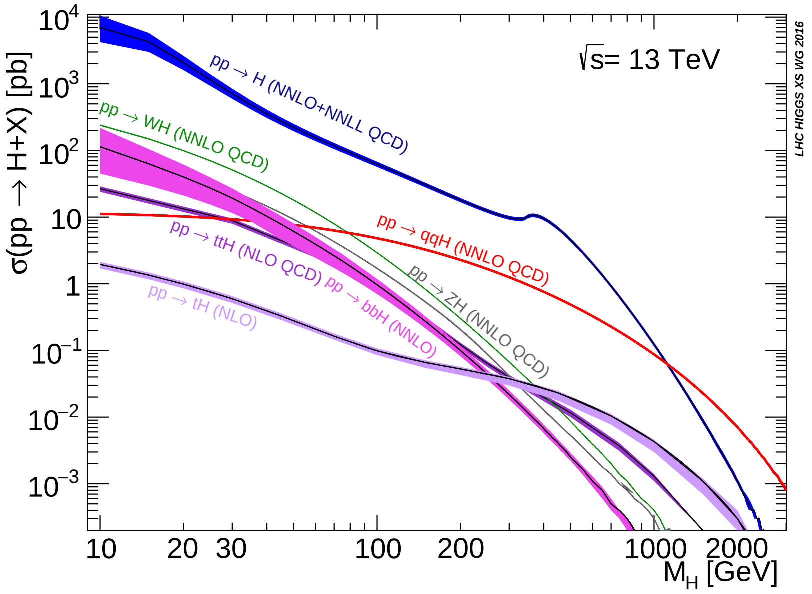


Figure 1: The predicted production cross section of the Higgs-boson as a function of the mass of the Higgs-boson for different production modes [11].

Process		H-Production Cross Section [pb]
VBF		$3.766\ ^{+0.45\%}_{-0.33\%}$(scale) $\pm 2.1\%$(PDF+α_s) $\times 10^6$
VH	W^-H	$0.527\ ^{+0.59\%}_{-0.63\%}$(scale) $\pm 2.03\%$(PDF+ α_s)
	W^+H	$0.831\ ^{+0.74\%}_{-0.73\%}$(scale) $\pm 1.79\%$(PDF+ α_s)
	ZH	$0.880\ ^{+3.50\%}_{-2.68\%}$(scale) $\pm 1.65\%$(PDF+ α_s)
ggH		$48.61\ ^{+4.27\%}_{-6.48\%}$(theory) ± 1.85 (PDF) $^{+2.59\%}_{-2.62\%}$(α_s)
$\bar{t}tH$		$0.507\ ^{+5.8\%}_{-9.2\%}$(scale) $\pm 3.6\%$(PDF+α_s)

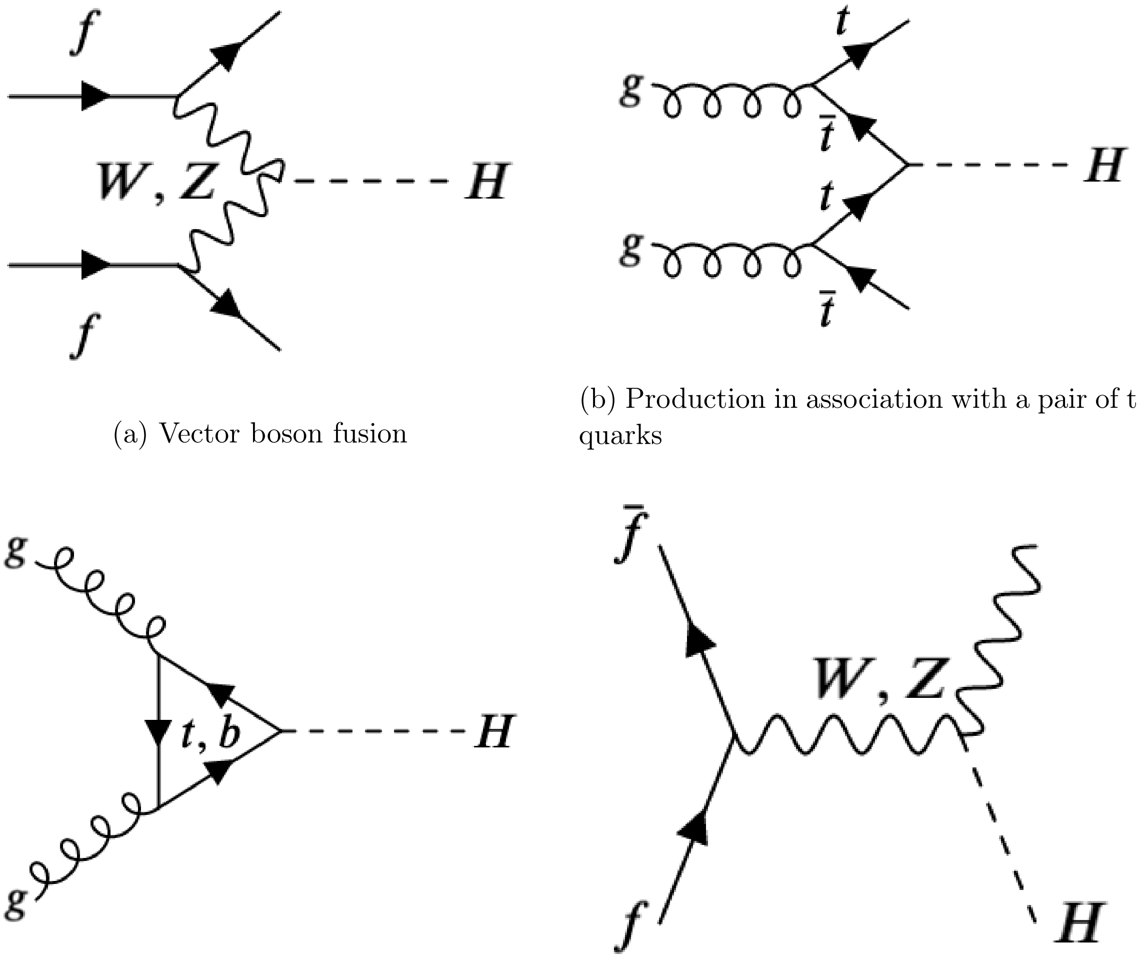
Table 1: The production cross section of the Higgs-boson as predicted by the Standard Model at a Higgs mass of 125.09 GeV at a center of mass energy of 13 TeV. Explanations for uncertainties are described in [12].

2.4 Decays of the Higgs-boson

This thesis considers the $H \to \tau\tau \to e\mu 4\nu$ process. However, since neutrinos are not detected by the ATLAS detector, this process is not distinguishable from the $H \to WW \to e\mu 2\nu$ process. Therefore this process is also classified as a signal process.

The different predicted decay branching ratios of the Higgs-boson depend on the mass of the Higgs-boson. The branching ratio of each possible Higgs decay as a function of the mass of the Higgs-boson can be seen in figure 3. The predicted branching ratios of the relevant decays $H \to \tau\tau$ and $H \to WW$ for a Higgs mass of $m_H = 125.36$ GeV can be seen in table 2.

The τ-lepton has a very short mean lifetime of $290.3 \pm 0.5 \times 10^{-15}$ s and therefore only its decay products can be directly measured. The τ-lepton may decay hadronically or leptonically. It decays into $\mu + \bar{\nu_\mu} + \nu_\tau$ in $17.39 \pm 0.04\%$ of cases and into $e + \bar{\nu_e} + \nu_\tau$ in $17.82 \pm 0.04\%$ of cases [10], where the e and μ are the visible parts of the decay, which are



(a) Vector boson fusion

(b) Production in association with a pair of top quarks

(c) Gluon-Fusion

(d) Higgs-Strahlung

Figure 2: Example Feynman diagrams of the different production modes of the Higgs-boson: (a) Vector boson fusion, (b) Production in association with a pair of top quarks (c) Gluon-Fusion (d) Higgs-Strahlung.

detected by the detectors, and the neutrinos the invisible parts of the decay, which are not detected by the ATLAS detectors. Therefore, a di-τ system decays into $e + \mu + 4\nu$ in 6.20±0.03% of all cases. A W boson decays into a $e + \nu_e$ and $\mu + \nu_\mu$ in 10.71±0.16% and 10.63±0.15% of cases respectively [10]. A WW system therefore decays into $e + \mu + 2\nu$ in 2.91±0.08% of all cases.

The product of the branching ratio and the VBF Higgs production cross section for $H \to \tau\tau \to e\mu 4\nu$ and $H \to WW \to e\mu 2\nu$ are therefore 14.6 ± 1.3 fb and 24.1±2.3 fb respectively. Using the relation $N = \mathcal{L}_{\text{int}}\sigma$, where N is the number of events, $\mathcal{L}_{\text{int}} = 139\text{fb}^{-1}$ the integrated luminosity and σ the cross section, one can find the expected number of events for these processes. For $H \to \tau\tau \to e\mu 4\nu$ and $H \to WW \to e\mu 2\nu$ these are 1950 ± 170 and 3300 ± 190 events respectively.

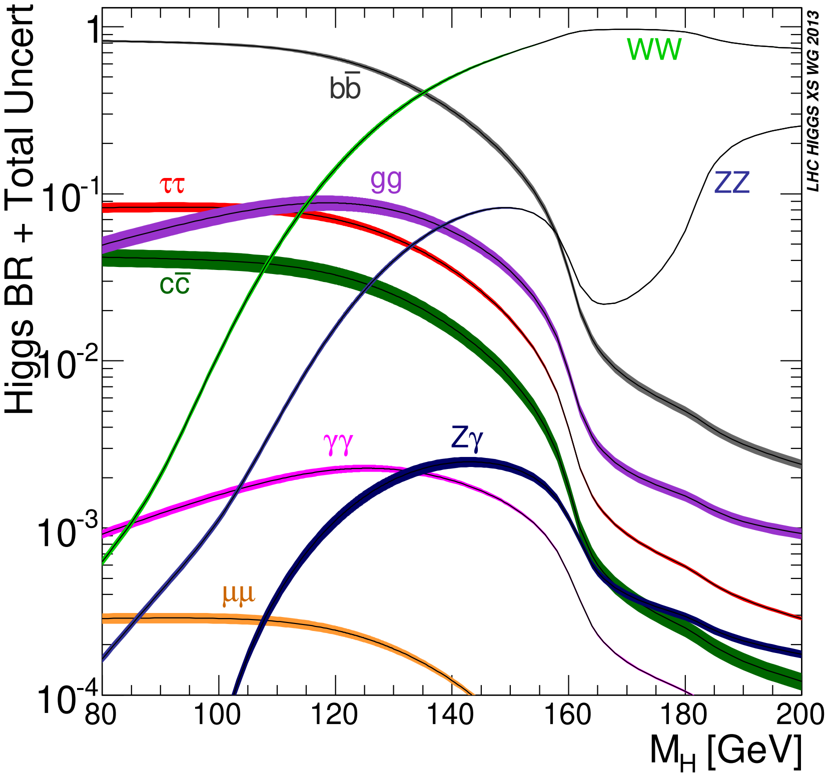


Figure 3: The Branching Ratios of the different Higgs decay modes as a function of the mass of the Higgs-boson as predicted by the Standard Model [12].

Decay Channel	Branching Ratio/%
$H \rightarrow WW$	22.0 ± 0.9
$H \rightarrow \tau\tau$	6.26 ± 0.35

Table 2: Branching Ratios for $H \rightarrow WW$ and $H \rightarrow \tau\tau$ as predicted by the Standard Model at $m_H = 125.36$ GeV [14].

3 Artificial neural networks

The increase in computing power allowed the development and implementation of computer algorithms that would improve automatically through experience. A method in machine learning is supervised learning, where an algorithm tries to find a certain function that maps an input to an output by analyzing different input-output pairs. The main goal of machine learning is either regression, where the relationship between an independent variable and a dependent variable is estimated, or classification, where a certain item is classified to a certain category. One approach in machine learning is the usage of artificial neural networks. This thesis explores the usage of artificial neural networks in classifying different recorded events from pp-collisions in the LHC to their respective processes.

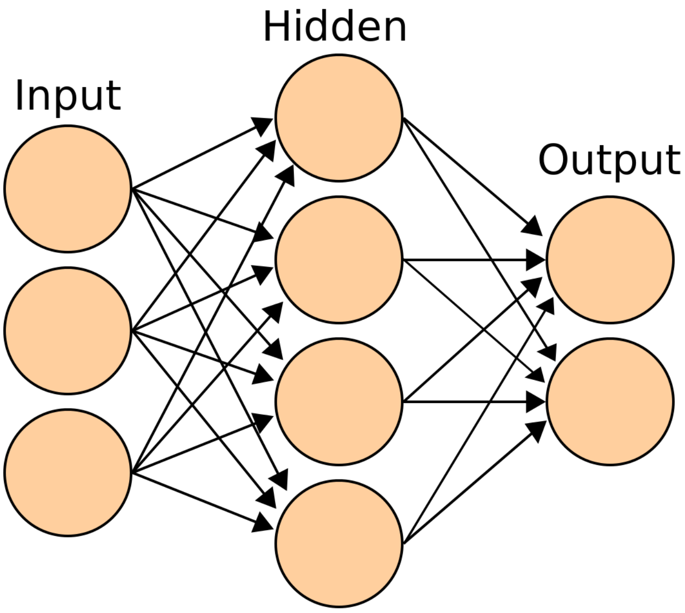


Figure 4: An example artificial neural network with one hidden layer [15]

3.1 Architecture of a neural network

An artificial neural network is composed of units called nodes. In this thesis, a feedforward neural network is used, where the connections do not form a cycle. Therefore every node is organized in layers whose nodes are connected to nodes in another layer. The input nodes form the first layer of neural network. For each input variable, there is an input node, which takes the value of the variable as its input. The last layer of the neural network is the output node, whose output is the prediction of the network. The layers of nodes between these input and output layers are referred to as hidden layers (see figure 4).

A node that receives inputs from nodes in the previous layer computes an output by non-linearly transforming the sum of the inputs. This non-linear transformation is called the activation function. A bias may be added to the argument of the function. The output of a node with an activation function f is therefore:

$$x_i^{j+1} = f\left(\sum_k w_k x_k^j + b_i^{j+1}\right) \tag{1}$$

Where x_i^{j+1} is the output of the i-th node in the $j+1$st layer, and w_k the weights provided by the connections from the neurons in the previous layer, and b the bias. In this thesis, two activation functions are considered: the Rectified Linear Units (ReLU) for hidden layers, defined by:

$$f_{\text{ReLU}}(x) = \max(0, x) \tag{2}$$

and Softmax for output layers, defined by:

$$f_{\text{Softmax}}(x_i) = \frac{e^{x_i}}{\sum_k e^{x_k}} \tag{3}$$

respectively, where x_i is the input to the i-th node. The reason for the usage of Softmax in classification is due to the fact that the sum of the outputs of the output nodes sum to 1, allowing the outputs of these nodes to be considered as probabilities. $f_{\text{Softmax}}(x_i)$ is then the prediction of the model, $\hat{y}_i$, for node i. This contrasts with the true value, y_i. For example, if there were two output nodes, the first for signal events and the second for background events, and if a signal event were to be classified as a background event with a 100% certainty, the true values would be $y_1 = 1, y_2 = 0$, while the predicted values would be $\hat{y}_1 = 0, \hat{y}_2 = 1$.

The goal of the training process is to find a combination of weights and biases that minimizes the error between the predicted values and true values.

The error as a function of the weights and biases is called the loss function and may have different forms, depending on the structure of the network. This thesis will utilize the categorical cross entropy as the loss function, commonly used in classification problems. The categorical cross entropy loss function is defined as:

$$L = -\sum_i y_i \cdot \log \hat{y}_i \tag{4}$$

Where $\hat{y}_i$ is the prediction of the model, and y_i the target value.

3.2 Minimizing the loss function

An approach to finding a combination of weights and biases that minimizes the loss function is stochastic gradient descent (SGD) [16]. Since the error is a function of the weights and biases of the node connections, for a feedforward network, it spans a hyperplane of $\sum_1^{N-1} n_i n_{i+1} + n_i$ dimensions, where n_i is the number of nodes in the i-th layer and N is the number of layers. To find the minimum of this function, a combination of weights and biases (represented as a vector $\hat{w}$) is selected and the gradient of the loss function in respect to the weights and biases is calculated. A term proportional to the gradient is then added to the vector, which effectively brings the point of the vector closer to the minimum of the loss function (see equation 5). The size of this increment is determined by the learning rate η, and the amount of data used to compute the gradient is called the batch size. Therefore, the combination of weights for a given iteration is given by:

$$\hat{w}_{i+1} = \hat{w}_i - \eta \nabla_{\hat{w}} L \tag{5}$$

, where L is the loss function, η the learning rate, $\hat{w}_i$ the vector of weights and biases for the iteration i, and $\nabla_{\hat{w}}$ the gradient operator in respect to the weights.

The hyperparameters, which are parameters which remain constant during training, here are therefore the learning rate and batch size. An epoch is when the entire dataset is passed through the training.

3.2.1 Adam

Adam, short for *Adaptive Moment Estimation* [17], is a modification of SGD that utilizes first and second moment vectors, m and v, which are the running averages of the gradient and squared gradient respectively. With Adam, each iteration is defined by:

$$
\begin{aligned}
m_w^{(t+1)} &\leftarrow \beta_1 m_w^{(t)} + (1-\beta_1)\nabla_w L^{(t)} \\
v_w^{(t+1)} &\leftarrow \beta_2 v_w^{(t)} + (1-\beta_2)(\nabla_w L^{(t)})^2 \\
\hat{m}_w &= \frac{m_w^{(t+1)}}{1-\beta_1^{t+1}} \\
\hat{v}_w &= \frac{v_w^{(t+1)}}{1-\beta_2^{t+1}} \\
w^{(t+1)} &\leftarrow w^{(t)} - \eta \frac{\hat{m}_w}{\sqrt{\hat{v}_w}+\epsilon}
\end{aligned}
$$

Where $m_w^{(t)}$ and $v_w^{(t)}$ are the first and second moment vectors at iteration t for the weight w respectively, and $\hat{m}_w$ and $\hat{v}_w$ the bias-corrected first and second moment estimates respectively. L is the loss function. The initial moments, m_w^0 and v_w^0 are set to 0, and the parameters β_1 and β_2 are set to 0.99 and 0.999 respectively, and ϵ to 10^{-8}. Adam is often used for its computational efficiency and low memory requirements and which makes it suitable for large amounts of data [17].

3.3 Training and validation

The process of minimizing the loss function with respects to weights and biases by exposing the network to a training dataset is called training.

After this process of training the neural network, the model is then used to predict the output of a second dataset, a process known as validation. This provides an evaluation of the model and helps with the optimization of the hyperparameters.

In this thesis, a process called k-fold cross validation is used, the entire dataset is split into k equally sized parts. For a given fold one of these parts are used for validation, another for testing and the rest for training. For every fold, a different combination of parts is used for training, validation, and testing so that training, validation and testing has been done at least once on every part. In this thesis, 5-fold cross validation is used.

However, overtraining may occur, which is when the predictions are modelled too closely to the training data themselves, and not the general trend which it wishes to model. This may manifest in the validation loss being higher than that of the training loss. In order to combat this, several regularization methods were developed. In this thesis the usage of learning rate decay and L2 regularization will be examined.

3.3.1 L2 regularization

L2 Regularization penalizes networks with large weights and biases by adding a quadratic term to the loss function.

$$L_{l2} = L(w_1, \dots w_n, b_1, \dots, b_n) + \lambda \sum_{i=1}^{n} w_i^2 \tag{6}$$

Where L is the original loss function, and λ a positive parameter smaller than 1. Therefore, network configurations with large weights are suppressed depending on the size of the parameter λ.

3.3.2 Learning rate decay

The learning rate can be decreased by a certain increment every iteration, which is based on the fact that the combination of weights and biases approaches the minimum of the loss function every epoch. The learning rate for iteration i depends on the decay rate by the following equation:

$$\eta_{i+1} = \eta_i \frac{1}{1 + \text{Decay} \times i} \tag{7}$$

In this thesis, the decay rate is set to a constant of 0.001.

3.4 Hyperparameters

For this thesis, the following hyperparameters remain as constants: Number of epochs, set to 50, and decay, set to 0.001. The influence of the following hyperparameters on the performance of the network are investigated: L2 Parameter λ, learning rate η, batch size and network architecture.

4 Experiment

4.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [18], located in Geneva, Switzerland, is the largest and most energetic particle collider in the world. With a circumference of 27 km, it accelerates protons in two parallel circular beams, focused using hundreds of quadrupole magnets and bent by dipole magnets. The beams intersect at four points at which the protons are collided with each other. These intersection points house different particle detectors used for different experiments, the biggest of them being the ATLAS detector.

In 2015, a center-of-mass energy of $\sqrt{s} = 13$ TeV was reached, and marked the start of "Run-2", which lasted until 2018. During Run-2, the performance of the particle accelerator was further improved by increasing its instantaneous luminosity. The luminosity is defined as the ratio between the event rate and the interaction cross section. Therefore:

$$\mathcal{L}\sigma = \dot{N}. \tag{8}$$

The integrated luminosity is then the integral of the luminosity with respect to time.

$$\mathcal{L}_{\text{int}} = \int \mathcal{L} dt. \tag{9}$$

The total recorded luminosity by the ATLAS experiment in Run-2 corresponds to 139.0 ± 2.4 fb^{-1}. This value will be used for this paper.

4.2 The ATLAS Experiment

The ATLAS experiment is one of the four main experiments conducted with the LHC. It utilizes large concentric cylinder shaped detectors to do precision measurements of Standard Model processes including Higgs physics, and to search for phenomena from physics beyond the Standard Model.

4.2.1 Coordinate system of the ATLAS Detector

Due to the geometry of the ATLAS detectors, a cylindrical coordinate system is used. The direction of the beam defines the z-axis, while the positive x-axis points into the center of the LHC, and the positive y-axis upwards. The direction of the velocity vector of a particle is uniquely determined by two angles, θ and ϕ, where θ is the angle from the beam axis, ϕ around the beam axis (see figure 5). Other useful variables are the pseudorapidity η, and ΔR, the distance between two particles in the pseudorapidity-azimuthal angle space. The pseudorapidity η, maps θ from $[-\pi, \pi]$ onto $[-\infty, \infty]$ and is defined by:

$$\eta = -\ln \left[\tan\left(\frac{\theta}{2}\right)\right] \tag{10}$$

while ΔR is defined as:

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} \tag{11}$$

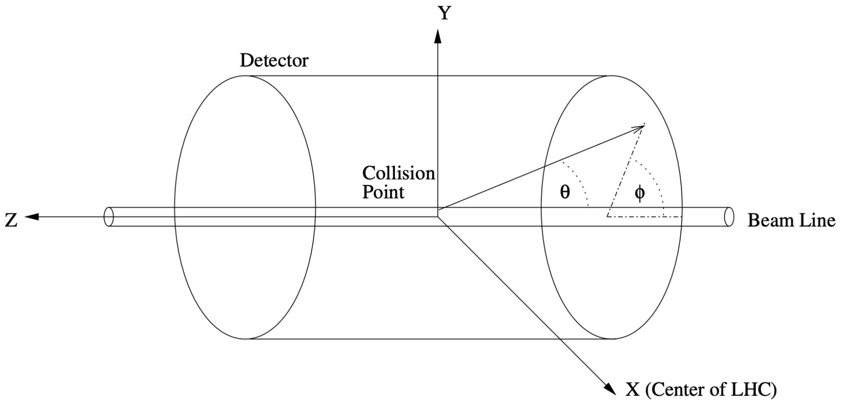


Figure 5: The Coordinate System of the ATLAS Detector. ϕ lies in the $x-y$ plane while θ lies in the plane containing the z-axis at an angle ϕ [19].

4.2.2 The ATLAS Detector

The ATLAS detector is a cylindrical general-purpose particle detector with a length of approximately 44 meters and a diameter of 25 meters. It is composed of different parts, each specialized to measure certain quantities.

The innermost part, the inner detector, is composed of three components, the pixel detector, semiconductor tracker and transition radiation tracker, whose main purpose is the measurement of momentum and charge of charged particles as well as vertex reconstruction, a process, which reconstructs the location of a certain decay event [20]. The inner detectors cover a pseudorapidity range of $|\eta| < 2.5$ [8]. The solenoid magnet surrounds the inner detector and creates a magnetic field with a strength of 2T, which bends the path of charged particles to determine their charge and momentum [20].

Beyond this are the calorimeters, which measures the amount of energy deposited by traversing electrons, photons and hadrons. This part is composed of the electromagnetic and hadronic calorimeters. The main principle of a calorimeter is that a traversing particle in a dense material may interact with the material to emit secondary particles, which in turn create further secondary particles. This phenomenon is known as a shower, whose energies are deposited into the detector. The electromagnetic calorimeter is a liquid argon calorimeter which measures the deposited energy of electrons, positrons and photons and has an energy resolution of $10\%/\sqrt{E}$ [21]. The hadronic calorimeters measure the energy deposited by hadrons. The hadronic sampling calorimeters have an energy resolution of $50\%/\sqrt{E}$ [22]. The calorimeters cover a pseudorapidity range of $|\eta| < 4.9$ [20]. The calorimeters are sampling calorimeters, where the material that produces the shower, the active material, is different to that which measures the deposited energy, the passive material. Surrounding the calorimeters are the barrel toroidal magnets, which provide a magnetic field of 8T [20].

Lastly, the muon spectrometer measures the momentum of muons which pass through the electromagnetic calorimeter undetected. The the monitored drift tubes of the spectrometer, which measure the curvature of the muon tracks using the magnetic field from the toroidal magnets, which provide a magnetic field of 8T [20]. The momentum resolution for muons up to 1 TeV was measured to be 10% at most [23]. A diagram of the ATLAS detector can be seen in figure 6.

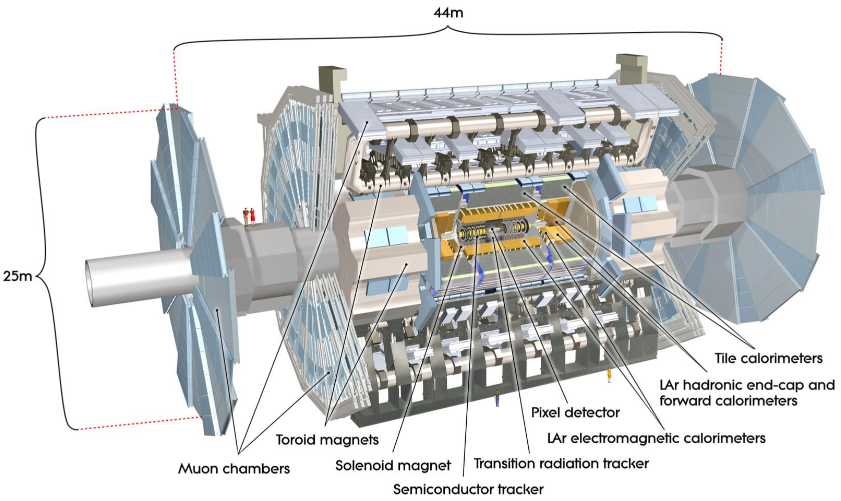

44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

Figure 6: A cut-away view of the ATLAS detector [20].

4.2.3 The Trigger System

The trigger system is designed to filter relevant data from the 40 million proton collisions per second in the ATLAS detector. The selection is done in two stages: A level-1 hardware based trigger and a high-level software based trigger. The level-1 hardware trigger uses electronics to analyze data from the calorimeters and muon solenoids. A decision to pass a certain data is made in only 2.5 microseconds [20]. Only 100,000 events per second can pass through the hardware trigger at most. The high-level trigger uses software to analyze the data coming from the level-1 trigger. Around 1000 events per second pass the high-level trigger and are recorded [20].

4.2.4 Event Reconstruction

The goal of event reconstruction is to use signals from the detectors to identify and reconstruct properties such as spacial trajectory, momentum, mass and charge of different particles. For this thesis, electrons, muons and jets are of relevance. An electron, being a charged particle, leaves a signal in the inner detector and in the electromagnetic calorimeter, while a muon leaves no signal in the electromagnetic calorimeter since it is minimally ionizing, but instead in the muon spectrometer. For the identification of electrons and muons, there are 3 different types of identification requirements: Loose, Medium and Tight, whose definitions can be seen in [24] for electrons and [25] for muons. In this thesis, the loose identification requirements are used for electrons and medium for muons. Jets are identified using the anti-k_t algorithm with a radius parameter of 0.4 [26] [27]. Jets originating from bottom quarks can be identified through vertex reconstruction. A jet originating from a bottom quark has its vertex displaced from the collision point. Machine learning algorithms are used to estimate the probability that a certain jet originates from a bottom quark. If the probability exceeds a certain threshold, the jet is called a b-tagged jet.

5 Signal and Background Processes and Event Generation

In order to evaluate the sensitivity of the event selection and to train the neural networks, simulated event samples for the signal and background processes are used. The samples are produced with Monte Carlo event generators for each process separately. Each event generated by the Monte Carlo generators carry a weight in order to make the probability distribution of simulated events consistent with actual data.

5.1 Signal Processes

The signal processes here are the VBF $H \to \tau\tau \to e\mu{+}4\nu$ and VBF $H \to WW \to e\mu{+}2\nu$ processes as mentioned in section 2.3. Such events are characterized by 2 jets with a large missing energy due to neutrinos, and a lack of jets origination from bottom quarks (b-jets). Other Higg boson production modes are then classified as background processes. The VBF signals are generated using the Powheg-Box v2 generator [28][29][30][31][32] with the PDF4LHC15 NLO parton distribution function (PDF) [33] set along with the Pythia 8 unterlying-event activity (UEPS) model [34].

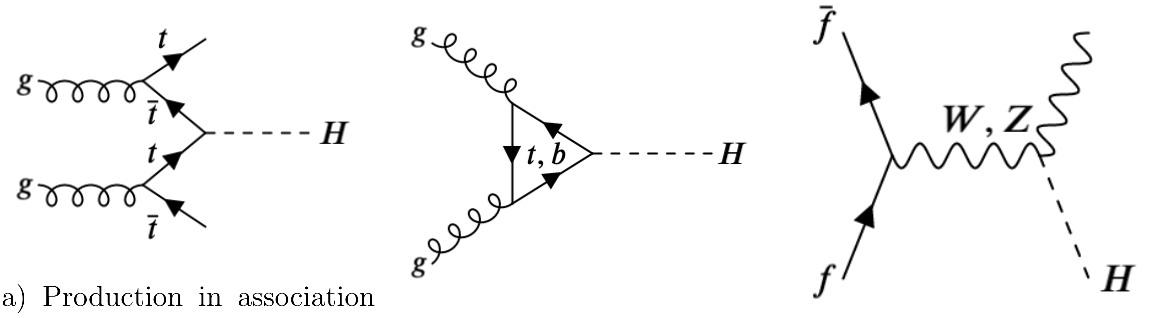
5.2 Background Processes

5.2.1 Other Higgs boson generation modes

The other Higgs boson production processes, namely ggF, ttH, and VH are already described in section 2.2. Example Feynman diagrams for these processes can be seen figure 7. For production in association with a pair of top quarks, the Higgs boson, along with the two jets originating from the top quarks leave a same signal as VBF Higgs boson production, thereby making it a background process. This background can be reduced by applying a veto against b-jets, since top quarks decay nearly exclusively into a bottom quark and W boson. For gluon fusion, if there are two other jets, the detectors also record the same signal as that of signal events. In the case of Higgs-Strahlung, the Z or W boson decaying into a pair of quarks, which are detected as jets, leave a signal identical to that of the signal process. The generators used to simulate these events can be seen in table 3.

Process	Generators	PDF Set	UEPS Model	Reference
ggF	Powheg-Box v2	PDF4LHC15 NNLO	Pythia 8	[28]
VH	Powheg-Box v2	PDF4LHC15 NNLO	Pythia 8	[35]
ttH	Powheg-Box v2	NNPDF2.3 NLO	Pythia 8	[36]

Table 3: The Monte Carlo generators used to simulate events of other Higgs boson production events.



(a) Production in association with a pair of top quarks

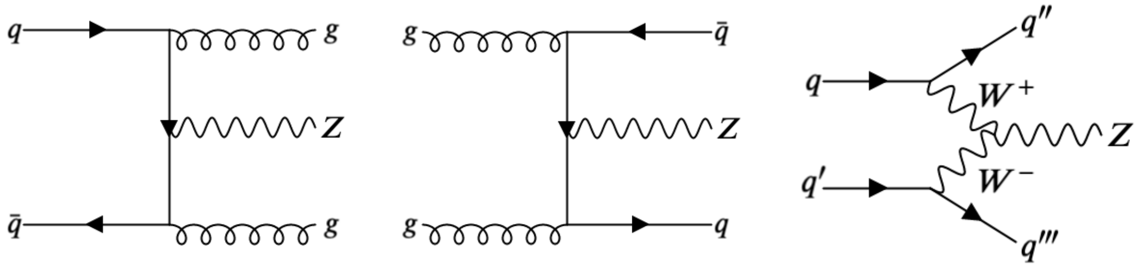
(b) Gluon-Fusion

(c) Higgs-Strahlung

Figure 7: Example Feynman diagrams of the different production modes of the Higgs Boson considered as background events:(a) Production in association with a pair of top quarks (b) Gluon-Fusion (c) Higgs-Strahlung.

5.2.2 Z+Jets

Z bosons are also produced in pp-collisions in the LHC through processes seen in figure 8. The total Z-boson production cross section multiplied by the leptonic branching ratio is measured to be 1.981 ± 0.007 (stat) ± 0.038 (sys) ± 0.042 (lumi) nb [37].



(a) Strong production

(b) Strong production

(c) Electroweak production

Figure 8: Example Feynman diagrams for dominant Z-boson production processes, where (a) and (b) are strong production processes and (c) an electroweak production process.

It was also measured that in 10% of cases, the Z-Boson decays into a lepton and an antilepton. The branching ratios of the leptonic Z-boson decays take a similar value of BR($Z \to l\bar{l}$) = 3.36% for a specific lepton l [10].

For a $Z \to \tau\tau$ decay where the Z-boson is created through strong production, the final end products are identical to that of VBF signals (2 jets, 2 τ's) but m_{jj} and $\Delta\eta_{jj}$ are smaller than that of VBF processes. For borg electroweakly and strongly produced $Z \to \tau\tau$ events, $m_{\tau\tau}$ differs from that of VBF signals due to the difference in mass between the Z-boson and Higgs boson. It is also possible that for a $Z \to ee$ or $Z \to \mu\mu$ that one of the electrons or muons are falsely identified as another particle or jet. Although the probability is small, this can also produce a $e\mu$ system. This is also the reason why in this thesis, only the $e\mu$ end state is considered: in order to decrease the contribution of $Z \to ee$ and $Z \to \mu\mu$ processes. The difference from VBF signals is that there is no missing transversal energy since no neutrinos are produced in this process.

5.2.3 W+Jets

W-bosons are also produced at the LHC. The cross section for W-boson production multiplied with the leptonic branching ratio for one generation was measured to be 11.83 ± 0.02 (stat) ± 0.32 (sys) ± 0.25 (lumi) pb for W^+ and 8.79 ± 0.02 (stat) ± 0.24 (sys) ± 0.18 (lumi) pb for W^- [37]. A W-Boson may decay into a lepton and a antineutrino or into a quark and antiquark of different types. When a W boson is generated along with 3 jets, where one of the jets is misidentified as an electron, and if the W boson decays into a muon and muon neutrino pair, this gives a same signal as that of a VBF event. However, this process has a different m_{jj} and $m_{\tau\tau}$ than that of VBF processes due to the difference in mass between the W and Higgs boson.

The W- and Z + Jets processes are both simulated using the generator Sherpa 2.2.1 [38] with the PDF set NNPDF3.0 NNLO and UEPS model Sherpa 2.2.1.

5.2.4 Diboson-Production (VV)

The following diboson-production processes as considered:WW, ZW and ZZ production. Examples of Feynman diagrams for these three processes can be seen in figure 9. The cross section of WW production was measured to be 115±5.8(stat)±5.7(exp)±6.4(theo)±3.6(lumi) pb, that of WZ production 40.9±3.4(stat)$^{+3.1}_{-3.3}$(sys)±0.4(theo)±1.3(lumi) pb, and that of ZZ production 14.6$^{+1.9}_{-1.8}$(stat)$^{+0.5}_{-0.3}$(sys)±0.2(theo)±0.4(lumi) pb [39]. For ZZ production, one Z boson may decay into a quark antiquark pair, which are detected as jets, and the other into a $\tau^-\tau^+$ pair, thereby creating the same end product as that of VBF processes. Similarly, for WZ production, the W boson may decay into a pair of quarks of different generations, which are detected as jets, while the Z boson may decay into $\tau^-\tau^+$ pair. For WW production, if one of the W bosons decay leptonically into a muon and muon neutrino pair and the other hadronically into two quarks, which are detected as two jets and if there is another jet, that is misidentified as an electron, it may leave a signal identical to that of VBF H events.

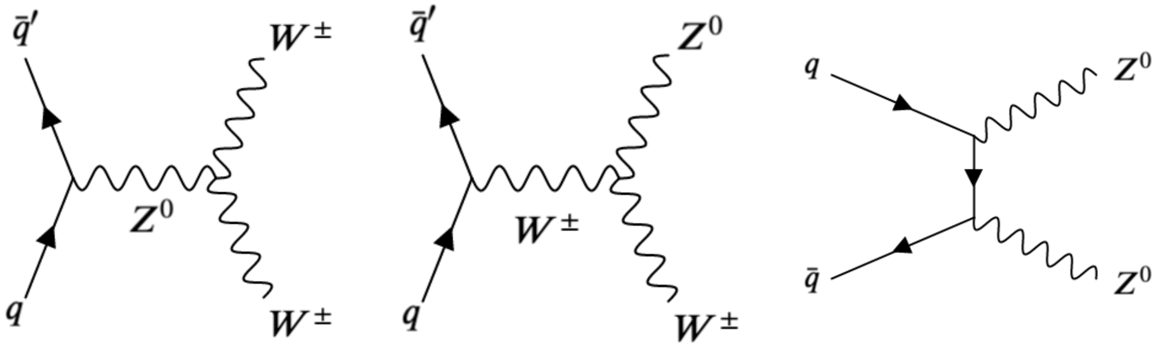


Figure 9: Example Feynman diagrams for different diboson production modes.

This process is simulated using the generator Sherpa 2.2.1, with the PDF set NNPDF3.0 NNLO and UEPS model Sherpa 2.2.1 [40].

5.2.5 Top Quark Pair Production

Pairs of top quarks are produced through processes shown in figure 10. This process was measured to have a total cross section of 781 ±7 (stat) ± 62(sys) ± 20(lumi) pb [41]. These processes are a background processes since top quarks decay almost exclusively into a bottom quark and W boson. The two b quarks turn into jets, which are detected by the detectors, and the two W bosons may decay into $e+\nu_e+\mu+\nu_\mu$. This leaves a signal that is identical to that of VBF events. This background can be reduced by applying veto against b-tagged jets, since top quarks decay nearly exclusively into a bottom quark and W boson. Top quark pair production is simulated using the Powheg-Box v2 generator with a PDF set NNPDF2.3 NLO and UEPS model Pythia 8.

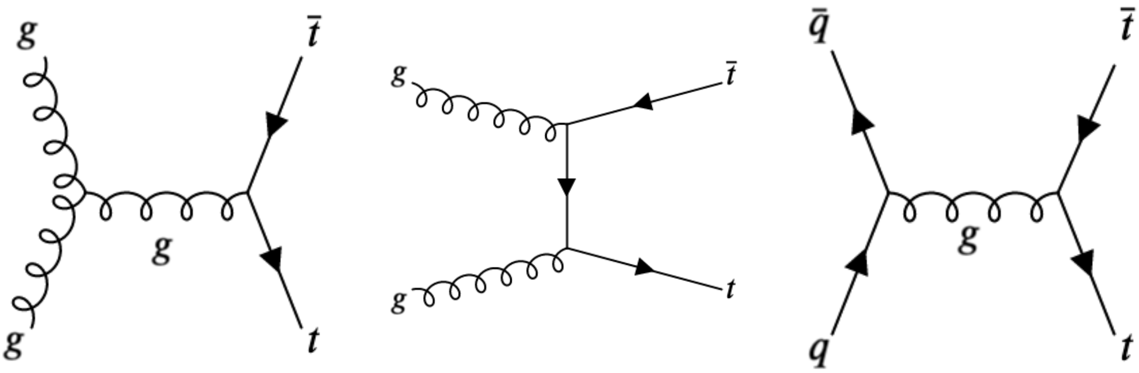


Figure 10: Example Feynman diagrams for different production modes of a pair of top quarks.

5.2.6 Single Top Quark Production

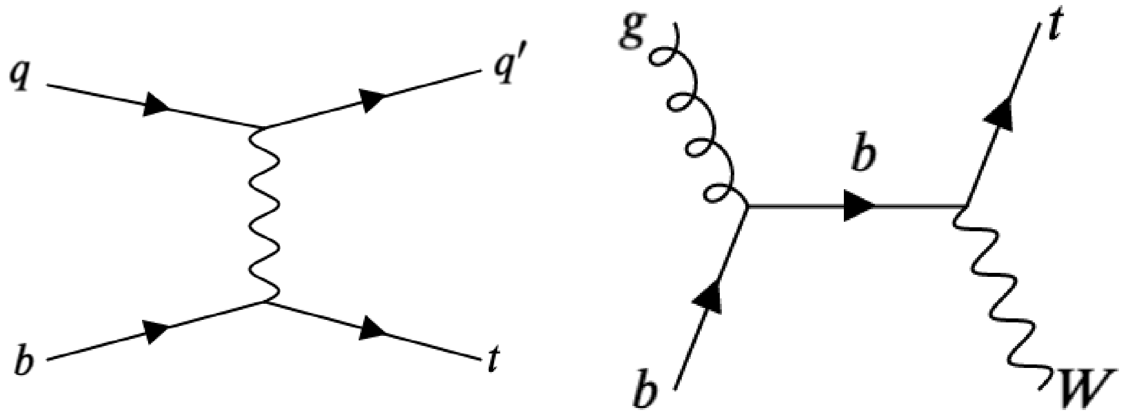
Single top quark production may occur as well, although it is not as common as top quark pair production. The most common modes of single top quark production can be seen in figure 11.

The total cross section of the t-channel top quark production was measured to be 219.0 ± 1.5(stat)± 13.0(sys) pb [42], and for that of the tW production 63.1 ± 1.8(stat) ± 6.4(sys) ± 2.1(lumi) pb [43]. This process is simulated with the Powheg-Box v1 generator [44] with the PDF set NNPDF2.3NLO and UEPS model Pythia 8.

Single top quark and top quark pair production are collectively referred to as the top background.

5.3 Fakes

Jets that are misidentified and incorrectly reconstructed as tau lepton decay products are called fakes. This process is poorly modeled by Monte Carlo simulations, furthermore,



(a) t-channel

(b) tW-production

Figure 11: Example Feynman diagrams for the dominant Single Top Quark Production modes in the LHC.

the rarity of the event makes it difficult to generate enough events. Therefore data driven methods are used to estimate this background process [45].

An overview of the Monte Carlo generators and the cross section of these signal and processes can be seen in table 4.

Process	Generator	PDF Set	UEPS Model	Tot. $\sigma(\times$ BR) [pb]
VBF	Powheg-Box v2	PDF4LHC15 NLO	Pythia 8	$3.766\ ^{+0.45\%}_{-0.33\%}$(scale) $\pm2.1\%$(PDF$+\alpha_s$) $\times10^6$
ggH	Powheg-Box v2	PDF4LHC15 NNLO	Pythia 8	$48.61\ ^{+4.27\%}_{-6.48\%}$(theory) ±1.85 (PDF) $^{+2.59\%}_{-2.62\%}(\alpha_s)$
$t\bar{t}H$	Powheg-Box v2	PDF4LHC15 NNLO	Pythia 8	$0.507\ ^{+5.8\%}_{-9.2\%}$(scale) $\pm3.6\%$(PDF$+\alpha_s$)
VH	Powheg-Box v2	NNPDF2.3 NLO	Pythia 8	$2.238\ ^{+1.28\%}_{-2.84\%}$(scale) $\pm3.16\%$(PDF$+\ \alpha_s$)
$Z+$Jets	Sherpa 2.2.1	NNPDF3.0 NNLO	Sherpa 2.2.1	1981 ± 7 (stat) ± 38 (sys) ± 42 (lumi)
$W+$Jets	Sherpa 2.2.1	NNPDF3.0 NNLO	Sherpa 2.2.1	11.83± 0.03 (stat) ± 0.40 (sys) ± 0.30 (lumi)
VV	Sherpa 2.2.1	NNPDF3.0 NNLO	Sherpa 2.2.1	115±5.8(stat)±5.7(exp)±6.4(theo)±3.6(lumi) pb (WW) 40.9 ± 3.4(stat)$^{+3.1}_{-3.3}$(sys)±0.4(theo)±1.3(lumi) (WZ) $14.6^{+1.9}_{-1.8}$(stat)$^{+0.5}_{-0.3}$(sys)±0.2(theo)±0.4(lumi) (ZZ)
Single top	Powheg-Box v1	NNPDF2.3 NLO	Pythia 8	282.1 ± 2.3(stat) ± 14.5(sys) ± 2.1(lumi)
Top pair	Powheg-Box v2	NNPDF2.3 NLO	Pythia 8	781 ±7 (stat) ± 62(sys) ± 20(lumi)

Table 4: An overview of the Monte Carlo generators and cross sections or product of cross section and leptonic branching ratio of the signal and background processes. Note that the cross sections and their uncertainties of Higgs production modes are theoretical, while others are experimental.

6 Event Selection

A series of event selection requirements, also known as *cuts*, were applied to the input dataset to increase the signal-to-background ratio. Only events, which pass all requirements are used for the training of the neural network. A summary of the event selection requirements can be seen in table 5. In the table, the preselection cuts refer to requirements that have already been applied onto the initial dataset and the additional cuts refer to those that were applied additionally in order to amplify the signal-to-background ratio. The additional cuts are primarily inspired by those used in the test of CP-invariance in VBF production of the Higgs-boson in the $H \to \tau\tau$ channel [46], where Boosted Decision Trees were trained instead of neural networks.

Category	Event Selection Requirements
Preselection	1 e with $p_T > 15$ GeV & $\lvert\eta\rvert < 2.47 \wedge \lvert\eta\rvert \notin [1.37, 1.52]$ 1 μ with $p_T > 13$ GeV & $\lvert\eta\rvert < 2.5$ 2 leptons with $\Delta R_{ll} < 2.5$ & $\Delta\eta_{ll} < 1.5$ At least 1 jet with $p_T^{\text{Jets}} > 20$, $\lvert\eta_{\text{Jets}}\rvert < 4.5$ GeV $p_T^{j_1} > 40$ GeV $E_T^{\text{miss}} > 20$ GeV $N_\tau = 0$ $0.1 < x_1 < 1$ $0.1 < x_2 < 1$
Additional Cuts	$m_{\tau\tau}^{\text{coll}} > m_Z - 25 \text{ GeV} \approx 66.18$ GeV $p_T^{l_1} > 18$ GeV $p_T^{l_2} > 14$ GeV $30 < m_{ll} < 100$ GeV $N_{\text{b-Jets}} = 0$ $N_{\text{Jets}} \geq 2$ $p_T^{j_2} > 30$ GeV $m_{jj} > 300$ GeV

Table 5: A summary of event selection requirements applied to the initial dataset. When multiple candidates of the same type exist, such as leptons or jets, they are ordered by transverse momentum, i.e. l_1 has a larger p_T than l_2. l_1 is then called the leading lepton and l_2 the subleading lepton.

6.1 Preselection requirements

This thesis, events with one electron with $p_T > 15$ GeV and $\lvert\eta\rvert < 2.47 \wedge \lvert\eta\rvert \notin [1.37, 1.52]$ along with one muon with $p_T > 13$ GeV and $\lvert\eta\rvert < 2.5$ are used. This is due to the fact that the tracker of the inner detector only covers a pseudorapidity range of $\lvert\eta\rvert < 2.5$ and that a pseudorapidity range of $1.37 < \eta < 1.52$ is not covered by the electromagnetic calorimeters. The transverse momentum of the jets must exceed 20 GeV and the absolute value of the pseudorapidity of jets must be smaller than 4.5. The requirement for the missing transversal energy was set to 20 GeV to reject events without neutrinos. The transversal momentum of the leading jet must exceed 40 GeV in order to select VBF events. The number of hadronically decaying τ leptons must be 0 since only leptonic decays of τ leptons are considered. The number of leptons are required to be two since this paper considers $e\mu$ end states. The number of jets must exceed 1. The difference in R and η between the leptons must be less than 2.5 and 1.5 respectively to suppress signals originating from Z-bosons, which are produced with low transverse momenta and therefore their decay products have larger angle separation. The visible momentum fractions x_1 and x_2 of their respective τ-lepton and di-τ mass calculated in the collinear approximation [47] was set between 0.1 and 1 to suppress events where the direction of the missing energy does not agree with expected di-τ decay kinematics.

6.2 Additional Selection Requirements

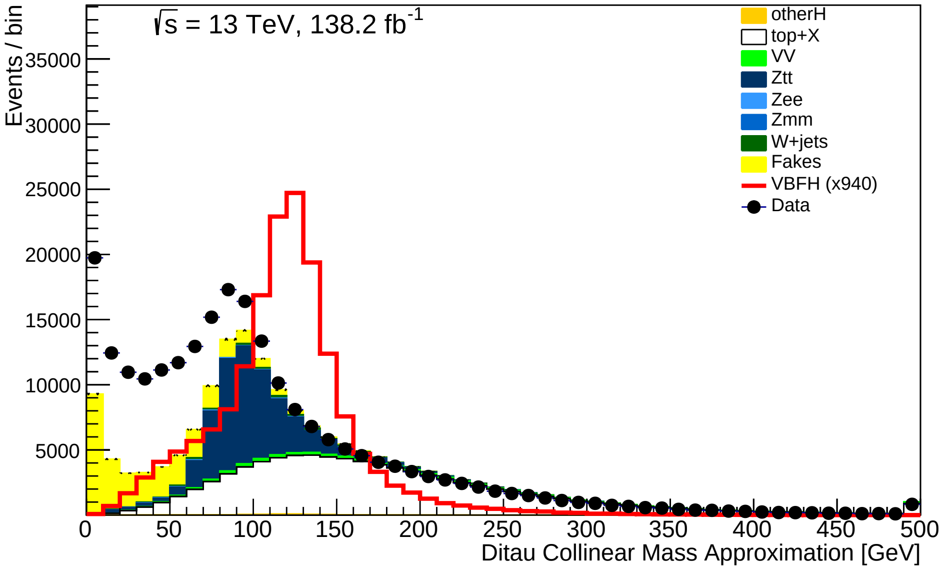
This section considers the distribution of each observable before their respective additional selection requirements in order to illustrate how these requirements amplify VBF signals.

For the analysis, events with at least two jets are considered since, VBF events produce two jets. The transverse momentum requirements of the leptonically decaying τ's were slightly increased from their preselection requirements since the signal to background ratio was greater for $p_T^{\tau_1} > 18$ GeV and $p_T^{\tau_2} > 14$ GeV (see plots 12b and 13c).

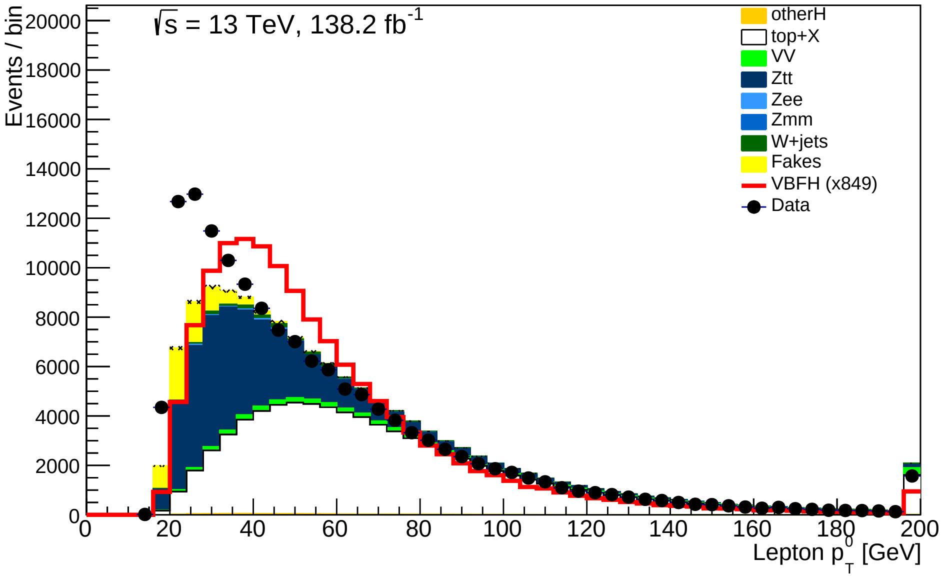
The location of the peak of the distribution of the di-τ mass calculated in the collinear approximation [47] for VBF signals, as seen in figure 12a, is higher than that of the background processes. Therefore in order to increase the signal to background ratio, only events with $m_{\tau\tau}^{\text{coll}} > 66.18$ GeV were accepted. In order to suppress signals from top pair production and diboson production, only events with $30 \leq m_{ll} \leq 100$ GeV were accepted. As seen in figure 13d, most signals originating from VBF processes have a m_{ll} in this range. Only $p_T^{j_2}$ of greater than 30 were selected, since as seen in figure 15g, the signal to background ratio is much higher in the $p_T^{j_2} > 30$ GeV region. Only events with $m_{jj} > 300$ GeV are taken to suppress signals from Z bosons. It can be seen in figure 15h that background signals have a small contribution in this region. The number of b-jets was required to be 0, also known as a b-veto, to suppress top background processes. As seen in figure 14e, the signal to background ratio is significantly greater for $N_{b-\text{Jets}}$=0. A table of the number of events before and after the additional selection requirements for each process can be seen in table 6.

Process	# Events (No Add. Cuts)	# Events (With Cuts)
VBFH	262.39 ± 0.44	74.28 ± 0.23
Other Higgs	1006.7 ± 2.0	53.53 ± 0.48
Top Background	138 630 ± 140	1075 ± 12
Diboson	6829 ± 19	399.4 ± 3.3
$Z \to \tau\tau$	71 440 ± 140	3500 ± 20
$Z \to ee$	3300 ± 150	26 ± 12
$Z \to \mu\mu$	362 ± 48	10.7 ± 3.2
W+Jets	12 720 ± 470	159 ± 35
Fakes	420 500 ± 640	402 ± 16
All Backgrounds	654 750 ± 830	5630 ± 47
$s/b\,[10^{-3}]$	4.007 ± 0.008	13.19 ± 0.12
$s/\sqrt{s+b}$	0.3242 ± 0.0006	0.984 ± 0.005

Table 6: The number of event before and after all additional selection requirements for different simulated signal and background processes as well as the signal-to-background ratio s/b and significance $s/\sqrt{s+b}$.

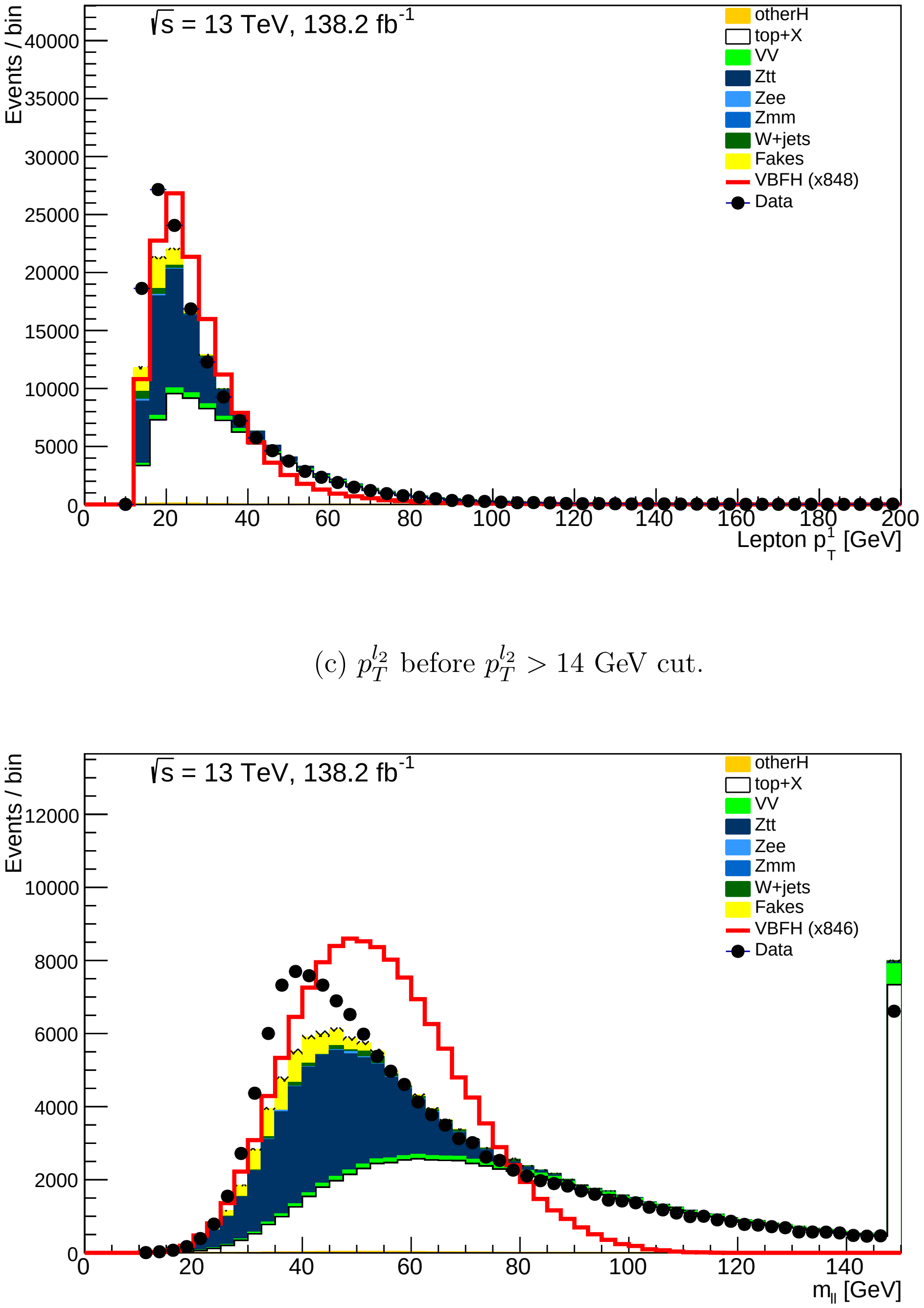


(a) Di-τ mass calculated in the collinear approximation ($m_{\tau\tau}^{\text{coll}}$) before $m_{\tau\tau}^{\text{coll}}$>66.18 GeV cut



(b) $p_T^{l_1}$ before $p_T^{l_1} > 18$ GeV cut.[a]

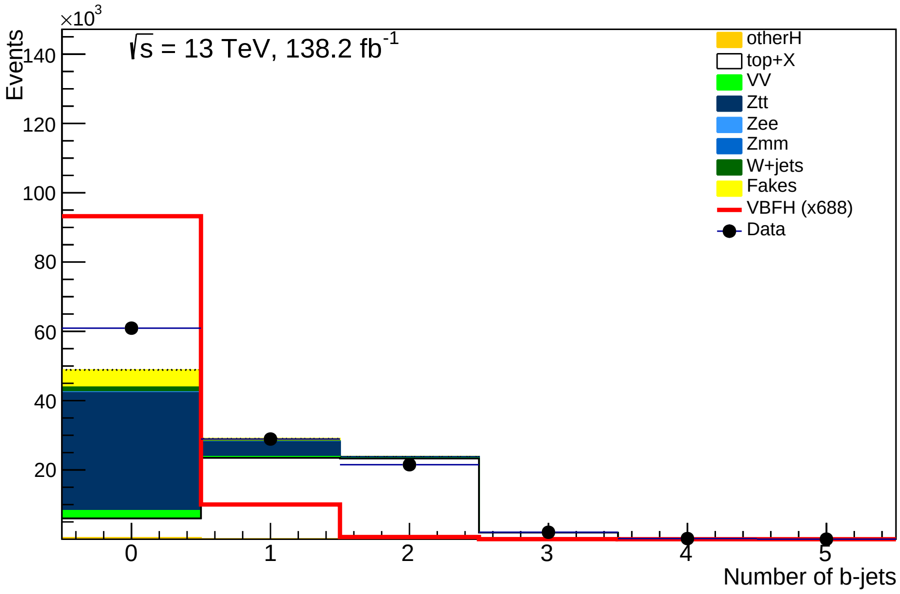
[a] The numbering of the leptons and jets in the figure follow the zero-based numbering convention used for programming, while in text it follows one-based numbering.



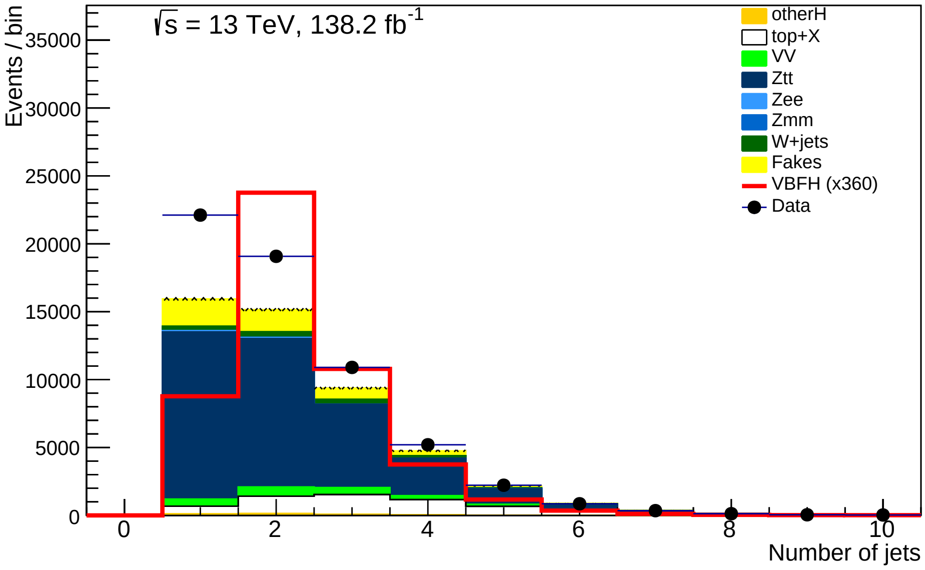
(c) $p_T^{l_2}$ before $p_T^{l_2} > 14$ GeV cut.

(d) Di-τ Mass (m_{ll}) before the $30 < m_{ll} < 100$ GeV cut.

Figure 13: The distribution of different variables before their respective additional selection requirements.

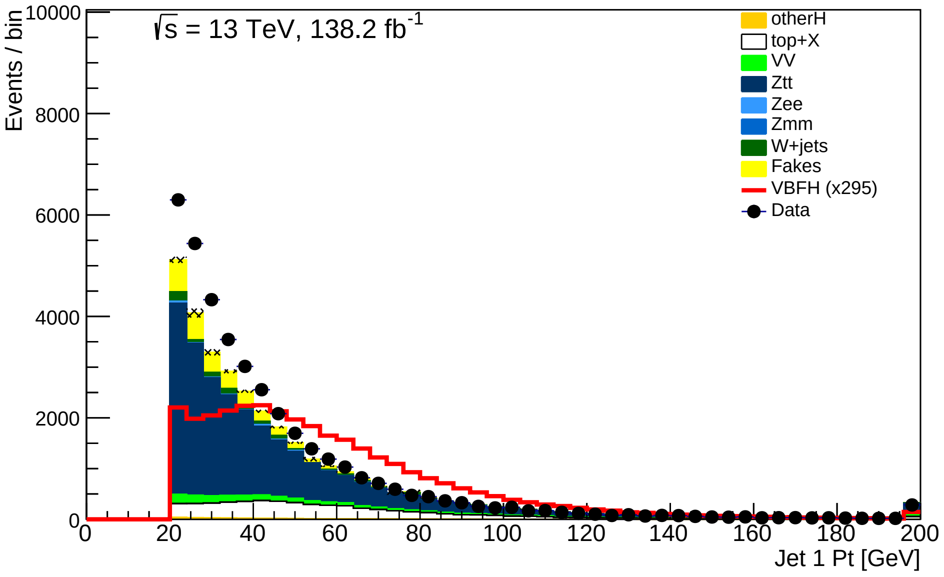


(e) Number of b-Jets before the $N_{\text{b-Jets}} = 0$ cut.

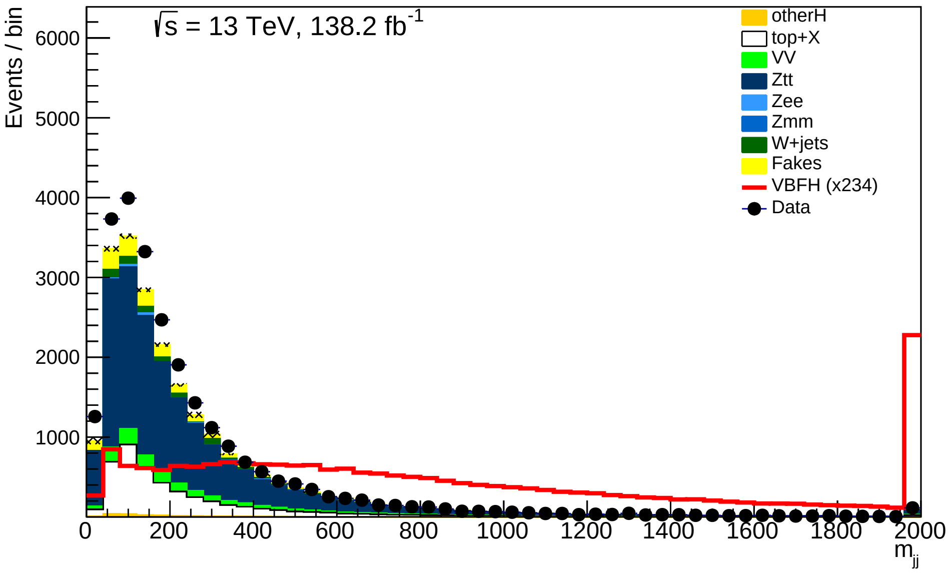


(f) Number of Jets before the $N_{\text{Jets}} \geq 2$ cut

Figure 14: The distribution of different variables before their respective additional selection requirements.



(g) $p_T^{j_2}$ before $p_T^{j_2} > 30$ GeV cut.



(h) m_{jj} before the $M_{jj} > 300$ GeV cut.

Figure 15: The distribution of different variables before their respective additional selection requirements.

7 Neural network analysis

In this section, an artificial neural network is developed and optimized to separate VBF $H \to \tau\tau \to e\mu 4\nu$ signals from background events. The influence of different hyperparameters, architecture and input variable sets on the performance of the neural network is examined. Events are classified as signal events if the neural network output is larger than a certain threshold. The performance of a neural network was determined by its maximum significance, where the significance is defined by:

$$\text{Sign.} = \frac{s}{\sqrt{s+b}} \tag{12}$$

Where s is the number of signal events, and b the number of background events that are classified as a signal event for a given threshold. The total uncertainty of the significance score was calculated through Gaussian propagation of error. For each network, the optimal threshold was determined by the threshold that gave the maximum significance.

Before the data was exposed to the neural network for training, a new weighting was defined to be used specifically for the training. In order to increase the neural network's ability to learn signals, the weights of the background events were normalized to yield the same expected amount of events as the signal events.

7.1 Optimizing the hyperparameters for a fixed architecture.

In order to consistently select an optimized combination of hyperparameters, the same process of selection was used for optimizing all neural networks. The initial network, before the optimzation, was selected to have a 50-50-50 architecture, 2 output nodes for signal and background events (which in reality corresponds to only one output node since only the output of the signal node is used), an L2 parameter, λ, of 10^{-5}, a learning rate of 0.01, batch size of 200, 50 epochs and a decay of 0.001. As a first step, the L2 parameters was varied using $\lambda \in \{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ to ensure there is no overtraining in the later steps. Then, using the optimal L2 parameter, the learning rate was varied for $\eta \in \{0.0001,\ 0.0005,\ 0.001,\ 0.005,\ 0.01,\ 0.05,\ 0.1\}$. Once the learning rate with the largest significance was selected, the batch sized was varied for batch sizes $\in \{50, 100, 200, 500, 1000, 2000\}$. The batch size which had the best results, as well as the selected L2 parameter and learning rate, is then termed the optimized combination of hyperparameters.

7.2 Input variables

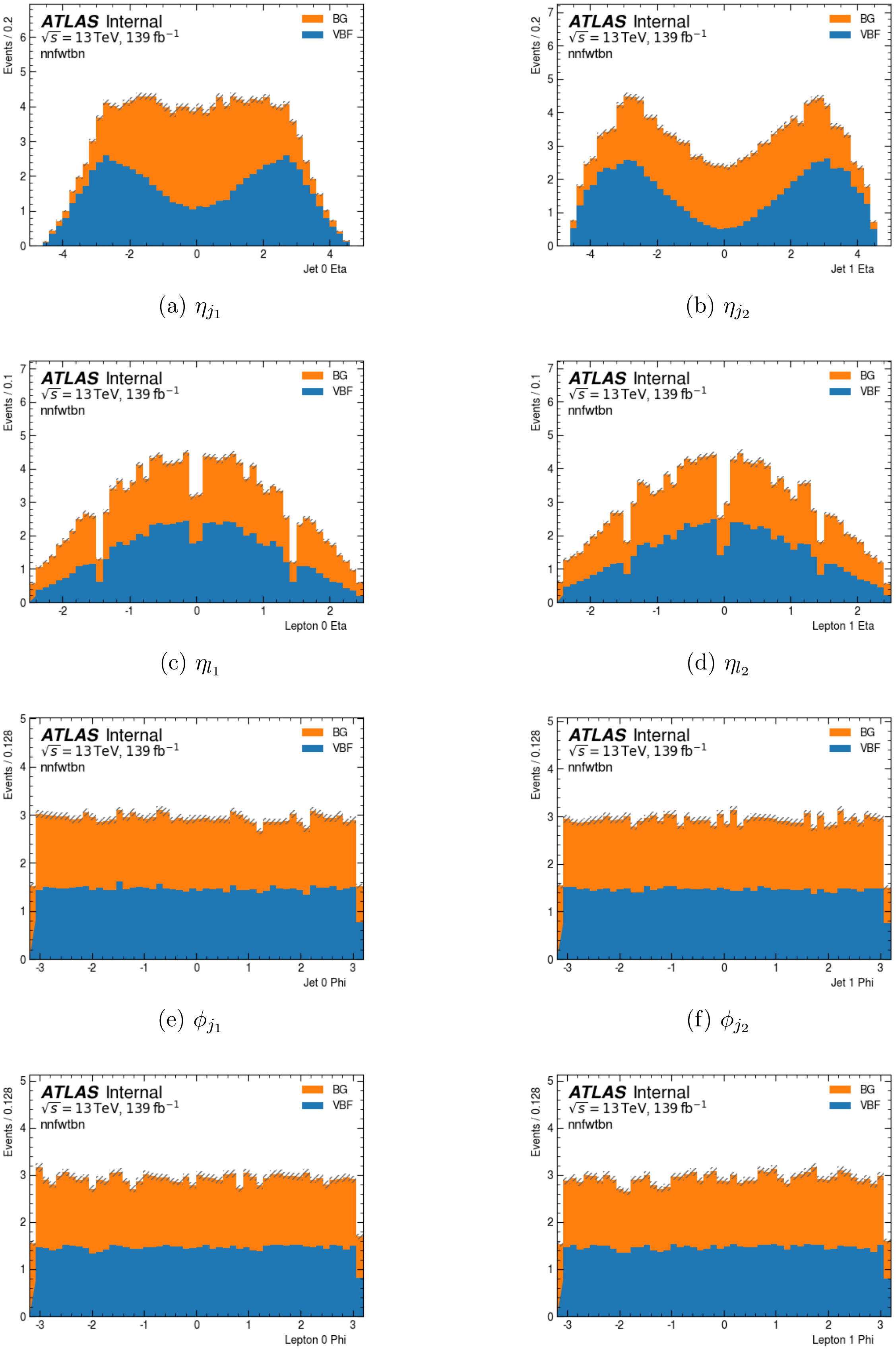
Several observables are used as input variables for the training of the neural network. Table 7 shows a list of high-level variables that will be used as input variables. This section aims to define these variables. η, ϕ, p_T are the pseudorapidity, ϕ angle and transverse momentum of leptons and jets. E_T^{miss} is the missing energy, the amount of energy needed to maintain conservation of momentum of the resulting particles from the pp collision.

Variable Set	DNN Input Variables
High Level	$m_{\tau\tau}^{\text{MMC}}$, m_{jj}, ΔR_{ll},$C_{jj}(\tau_1)$,$C_{jj}(\tau_2)$,p_T^{tot},$m_{\tau\tau}^{\text{vis}}$ $m_T^{l_1,E_T^{\text{miss}}}$, $E_T^{\text{miss}}/p_T^{l_1}$,$E_T^{\text{miss}}/p_T^{l_2}$,$p_T^{j_3}$

Table 7: The high level variable set used as the input for training the neural network.

$m_{\tau\tau}^{\text{MMC}}$ is the invariant mass of the di-τ system calculated using the missing-mass calculator [48]. m_{jj} is the invariant mass of the jets. $C_{jj}(\tau) = \exp\left[\frac{-4}{(\eta_{j_1}-\eta_{j_2})^2}\left(\eta_\tau - \frac{\eta_{j_1}+\eta_{j_2}}{2}\right)^2\right]$ is called the centrality, which has a value of 1 when the object is halfway in η between the two jets and $1/e$ when the object is aligned with one of the jets and $< 1/e$ when the object is not between the jets in η. p_T^{tot} is the magnitude of the vectorial sum of the transverse momenta of leptons, jets and missing energy. $m_{\tau\tau}^{\text{vis}}$ is the visible mass of the di-τ system. $m_T^{l,E_T^{\text{miss}}}$ is the transverse mass, defined as $\sqrt{2p_T^l E_T^{\text{miss}} \cdot (1 - \cos\Delta\phi_{l,E_T^{\text{miss}}})}$. The transversal momentum of the 3rd jet, $p_T^{j_3}$ was estimated by subtracting the magnitudes transversal momenta of the leading and subleading leptons and jets from $(\sum p_T)_{\text{scalar}}$.

The distribution of these variables for signal and background processes can be seen in figures 16 to 17.



(a) η_{j_1}

(b) η_{j_2}

(c) η_{l_1}

(d) η_{l_2}

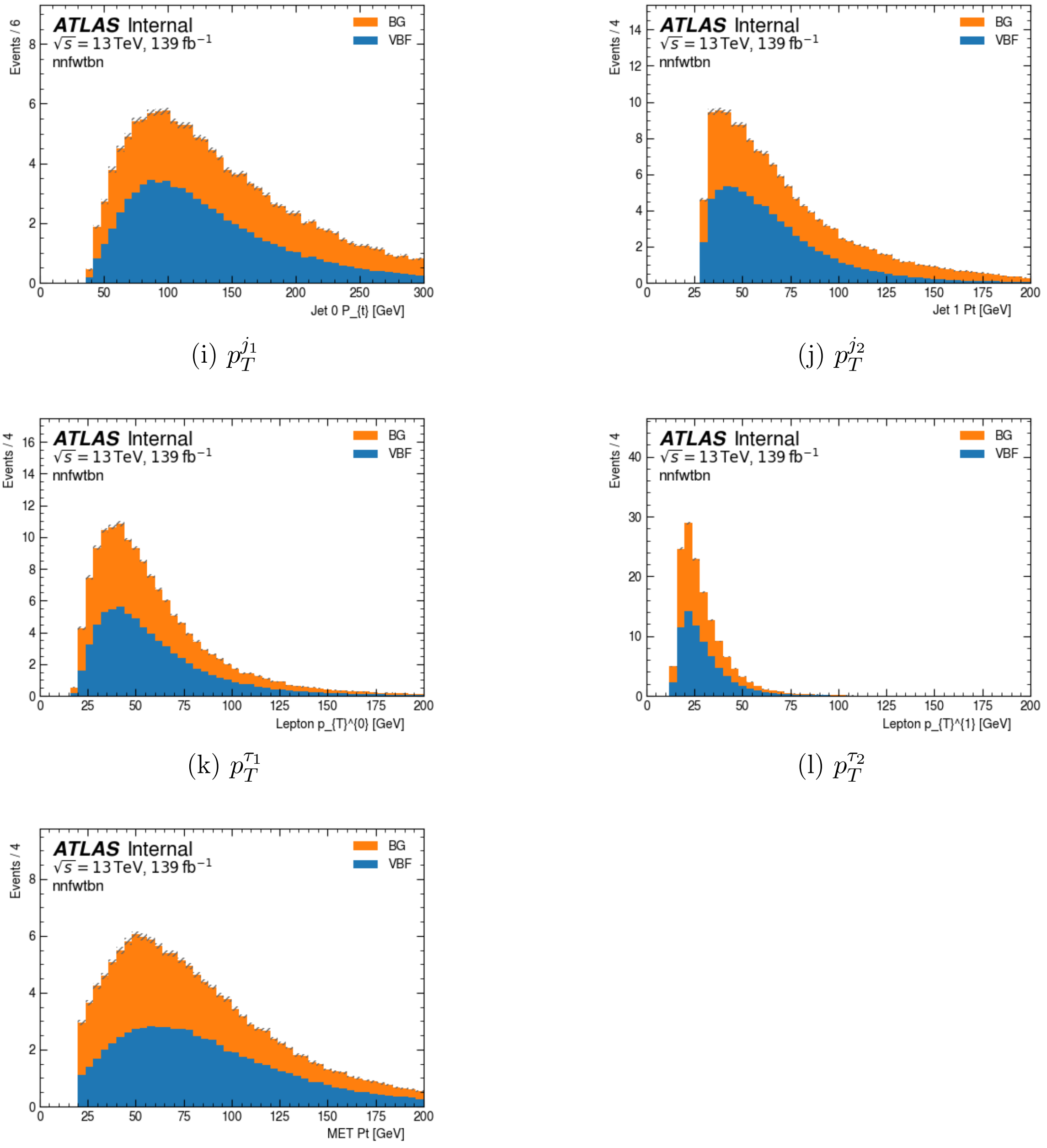
(e) ϕ_{j_1}

(f) ϕ_{j_2}

(g) ϕ_{l_1}

(h) ϕ_{l_2}

Figure 16: Distribution of variables in the Low Level (1) variable category (stacked). Signal events are shown in blue and background events as orange.



(i) $p_T^{j_1}$

(j) $p_T^{j_2}$

(k) $p_T^{\tau_1}$

(l) $p_T^{\tau_2}$

(m) E_T^{miss}

Figure 17: Distribution of variables in the Low Level (1) variable category (stacked). Signal events are shown in blue and background events as orange.

7.3 Influence of Network Architecture

Using the optimization process above, firstly the influence of the network architecture on the performance of the network was investigated. For this part, only high-level variables were used as input (see table 7). The following network architectures were investigated:

50-50
30-30-30
50-50-50
70-70-70
100-100-100
100-50-25
50-50-50-50

Table 8: The different network architectures that were investigated

Where each number represents the amount of nodes in a hidden layer. Out of these architectures, the full optimization process was done for the architectures 50-50, 50-50-50, 50-50-50-50 and 100-50-25. For the 30-30-30, 70-70-70 and 100-100-100 architectures, the optimal hyperparameter combination of the network with the 50-50-50 architecture was taken. The optimization procedure for the four network architectures can be seen in figures 18 to 20. The selected hyperparameter for each optimization step for each network architecture can be seen in tables 9 to 11. A table of the expected number of events for each process that pass the classification threshold can be seen in table 12.

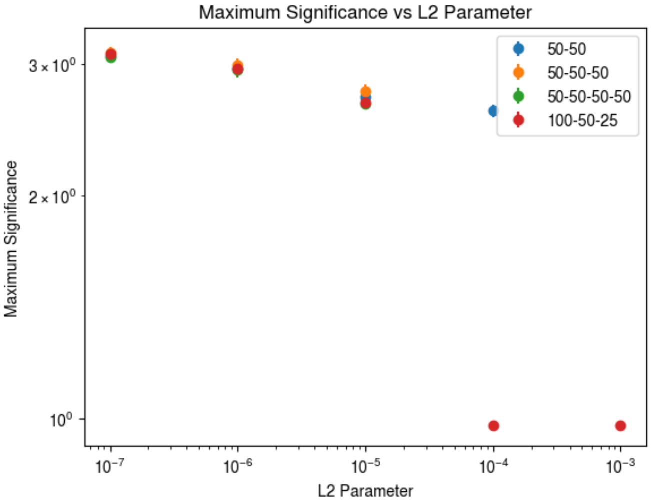


Figure 18: Maximum Significance as a function of the L2 λ Parameter for different network architectures. The other hyperparameters are: Batch Size = 200, Learning Rate=0.01.

Architecture	Selected L2 λ
50-50	10^{-4}
50-50-50	10^{-5}
50-50-50-50	10^{-5}
100-50-25	10^{-5}

Table 9: The selected L2 λ parameters for each network architecture as determined by their maximum significance and presence of overtraining.

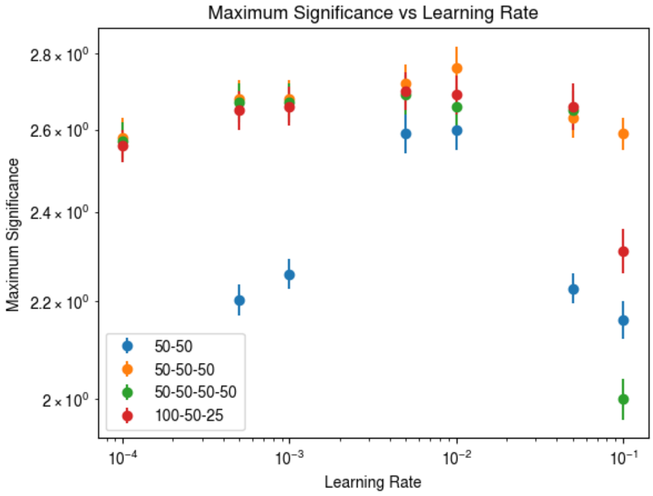


Figure 19: Maximum significance as a function of the learning rate for different network architectures. The other hyperparameters are: Batch Size = 200, and L2 λ parameter as determined in the previous step.

Architecture	Selected η
50-50	0.01
50-50-50	0.01
50-50-50-50	0.005
100-50-25	0.005

Table 10: The selected learning rate for each network architecture as determined by their maximum significance.

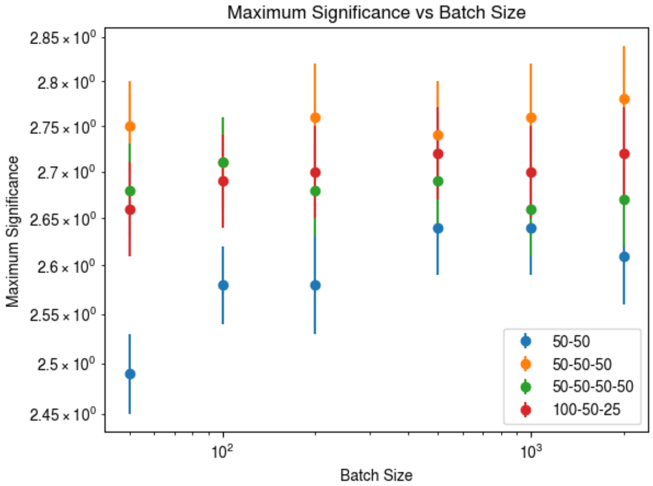


Figure 20: Maximum significance as a function of the batch size for different network architectures. The other hyperparameters are as determined in the previous steps.

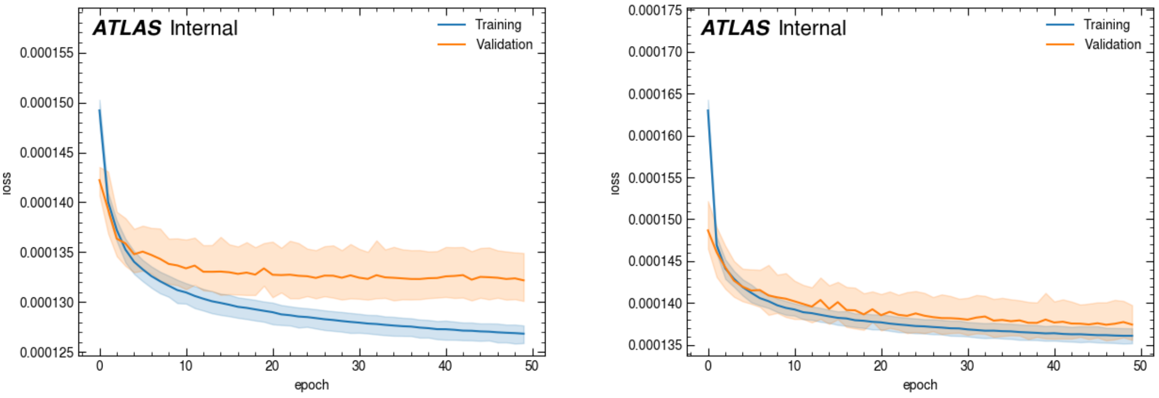
Architecture	Selected batch size
50-50	1000
50-50-50	2000
50-50-50-50	100
100-50-25	500

Table 11: The selected batch sizes for each network architecture as determined by their maximum significance.

Process	50-50	50-50-50	50-50-50-50
VBFH	61.36 ± 0.29	56.37 ± 0.27	47.92 ± 0.24
Other H	7.51 ± 0.24	6.25 ± 0.22	8.25 ± 0.25
VV	16.17 ± 1.28	10.63 ± 0.95	16.15 ± 1.30
Top	48.34 ± 4.10	34.77 ± 3.43	49.11 ± 4.12
$W+$ Jets	10.40 ± 4.91	7.26 ± 4.34	9.66 ± 4.52
$Z \to ee$	-1.10 ± 1.16	-1.15 ± 1.16	-1.15 ± 1.16
$Z \to \tau\tau$	114.29 ± 5.52	82.25 ± 5.00	119.18 ± 5.86
$Z \to \mu\mu$	0.18 ± 0.13	0.18 ± 0.13	0.27 ± 0.15
Fakes	21.42 ± 5.47	14.52 ± 4.66	18.82 ± 5.30
Signal/Bkg	0.29	0.36	0.29
Significance	2.65±0.05	2.80±0.06	2.72±0.05

Table 12: The expected number of events for each process that pass the signal classification threshold for the optimized networks with 50-50, 50-50-50 and 50-50-50-50 architectures.

During the optimization of the L2 parameter, the reason why a L2 parameter of $\lambda = 10^{-5}$ was often chosen, even though other parameters gave better significance was due to the presence of overtraining for $\lambda = 10^{-6}$ and 10^{-7}. This is evident from comparing the loss evaluated with the validation and training data sets. An example for a network architecture of 50-50-50 with a batch size of 200 can be seen in figure 21.

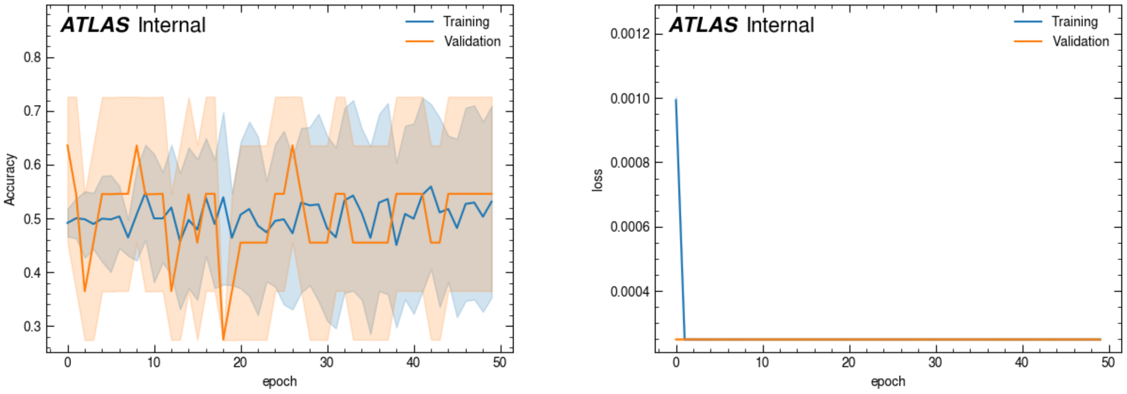


(a) L2 $\lambda = 10^{-7}$

(b) L2 $\lambda = 10^{-6}$

Figure 21: The loss as a function of the epoch for the training and validation samples for a network with a 50-50-50 architecture, learning rate of 0.01 and a batch size of 200.

Furthermore, it was observed that for large L2 parameters such as $\lambda = 10^{-4}, 10^{-3}$, the parameter would prevent the network from learning from the data. This caused the network to give an output of around 0.5 for both signal and background events. Hence when scanning over the signal classification threshold, if the threshold is below 0.5, all events are classified as signal events, resulting in the points in figure 18 that lie on 1. This is evident when one takes a look at the training and validation loss and accuracy as a function of the epoch. An example for a network with a 50-50-50 architecture and a L2 parameter of $\lambda = 10^{-3}$, batch size of 200 and learning rate of 0.01 can be seen in plot 22. As seen in figure 22a, the accuracy of the neural network does not significantly increase as a function of the epoch, but instead lies around 0.5. This would imply a random classification.



(a) Accuracy

(b) Loss

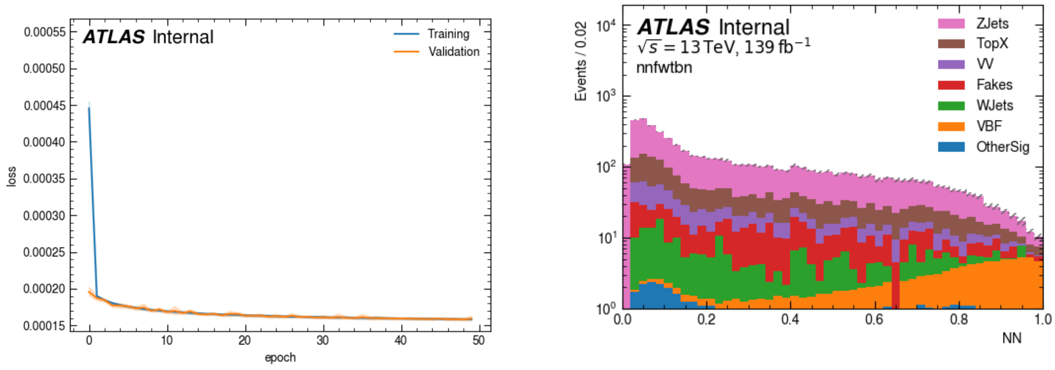
Figure 22: Training and validation accuracy and loss as a function of the epoch for a neural network with a 50-50-50 architecture, L2 parameter of $\lambda = 10^{-3}$, learning rate of 0.01 and batch size of 200.

The phenomenon where a small L2 parameter causes overtraining and where a large L2 parameter causes the net to not learn is a recurring phenomenon, which will be seen and be referred to in later optimization processes as well.

The optimal hyperparameter combination for these four network architectures as well as the best significance and optimal neural network output threshold are summarized in table 13. The a histogram of the output of the neural network as well as the training and validation loss as a function of the epoch can be seen in figures 23 to 26.

Architecture	L2 λ	η	Batch Size	Max. Sign.	NN Threshold
50-50	10^{-4}	0.01	1000	2.65 ± 0.05	0.86
50-50-50	10^{-5}	0.01	2000	2.80 ± 0.06	0.90
50-50-50-50	10^{-5}	0.005	100	2.72 ± 0.05	0.86
100-50-25	10^{-5}	0.005	500	2.72 ± 0.05	0.86

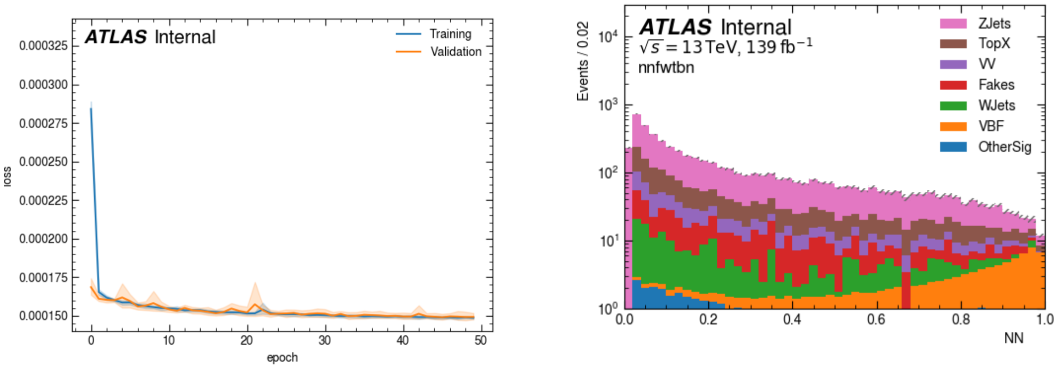
Table 13: The optimized combination of hyperparameters (L2 parameter λ, learning rate, batch size) for different network architectures along with their maximum significance and optimal neural network output threshold.



(a) Training and validation loss as a function of the epoch.

(b) The neural network output for different processes plotted on a logarithmic scale (stacked).

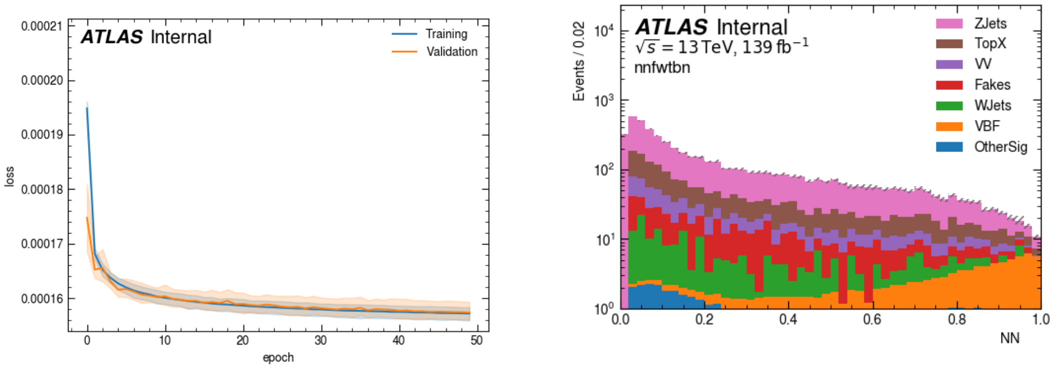
Figure 23: The training and validation loss and neural network output for the optimized network with a 50-50 architecture (L2 $\lambda = 10^{-4}$, η=0.01, batch size = 1000). The total of the background events were normalized to the signal events.



(a) Training and validation loss as a function of the epoch.

(b) The neural network output for different processes plotted on a logarithmic scale (stacked).

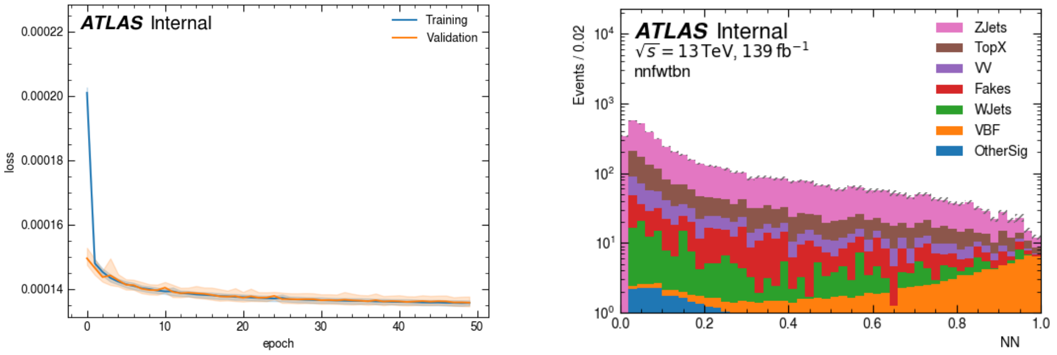
Figure 24: The training and validation loss and neural network output for the optimized network with a 50-50-50 architecture (L2 $\lambda = 10^{-5}$, η=0.01, batch size = 2000). The total of the background events were normalized to the signal events.



(a) Training and validation loss as a function of the epoch.

(b) The neural network output for different processes plotted on a logarithmic scale. (stacked).

Figure 25: The training and validation loss and neural network output for the optimized network with a 50-50-50-50 architecture (L2 $\lambda = 10^{-5}$, η=0.005, batch size = 100). The total of the background events were normalized to the signal events.



(a) Training and validation loss as a function of the epoch.

(b) The neural network output for different processes plotted on a logarithmic scale.

Figure 26: The training and validation loss and neural network output for the optimized network with a 100-50-25 architecture (L2 $\lambda = 10^{-5}$, η=0.005, batch size = 100). The total of the background events were normalized to the signal events.

7.3.1 Effect of Number of Layers

In order to see the effect of the number of layers on the performance of the neural network, the maximum significance was determined depending on the number of layers (with 50 nodes each). This can be seen in figure 27. The training and validation loss as a function of the epoch as well as the output of the neural network of these networks can be seen above in figures 23 to 25. A summary of the expected number of events for each process that pass the neural network signal classification threshold can be seen above in table 34.

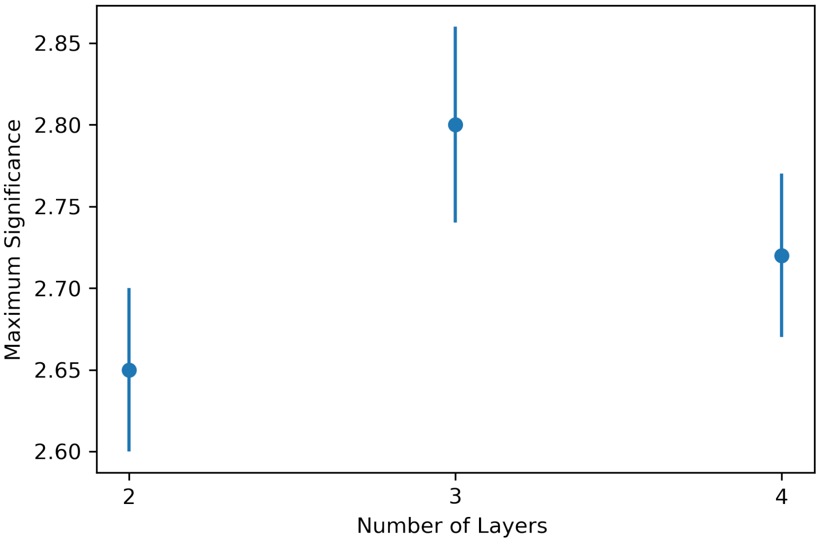
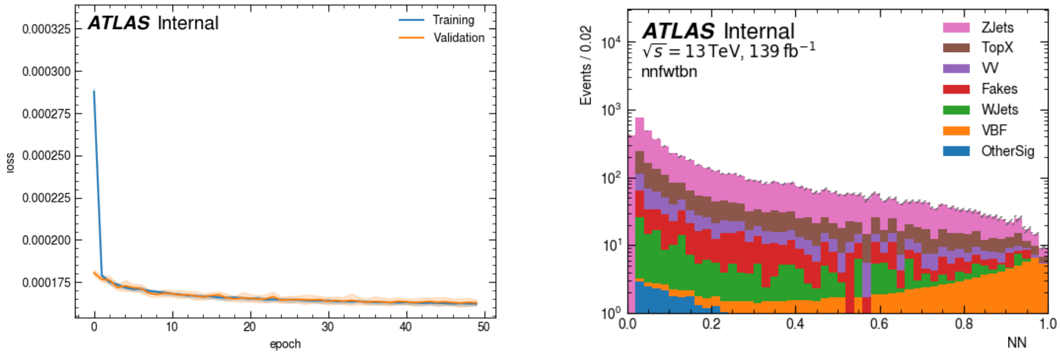


Figure 27: Maximum significance as a function of the number of layers with 50 nodes each. The other hyperparameters are as determined in the optimization process.

As it can be seen from the figure, the network with three layers had the best nominal value for the significance, however not significantly more than that of the network with 4 layers. It can be seen that the network with 3 layers performed significantly better than the one with two layers.

7.3.2 Influence of number of nodes per layer

As an additional investigation, for the optimized 50-50-50 architecture, the number of nodes was further varied to see its impact on the significance. This was done with 30, 50, 70 and 100 nodes per layer. The other hyperparameters were fixed to the optimal values of the 50-50-50 neural network. The maximum significance as a function of the number of nodes per layer can be seen in figure 31. The training and validation loss as a function of the epoch as well as the output of the neural network of these networks can be seen in figures 28 to 30. A summary of the expected number of events for each process that pass the neural network signal classification threshold can be seen in table 34.

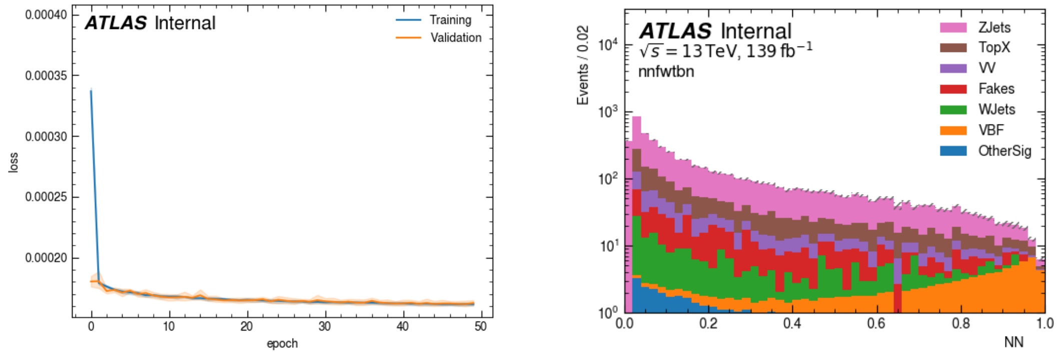


(a) Training and validation loss as a function of the epoch.

(b) The neural network output for different processes plotted on a logarithmic scale.

Figure 28: The training and validation loss and neural network output for the optimized network with a 30-30-30 architecture (L2 $\lambda = 10^{-5}$, η=0.01, batch size = 2000). The total of the background events were normalized to the signal events.

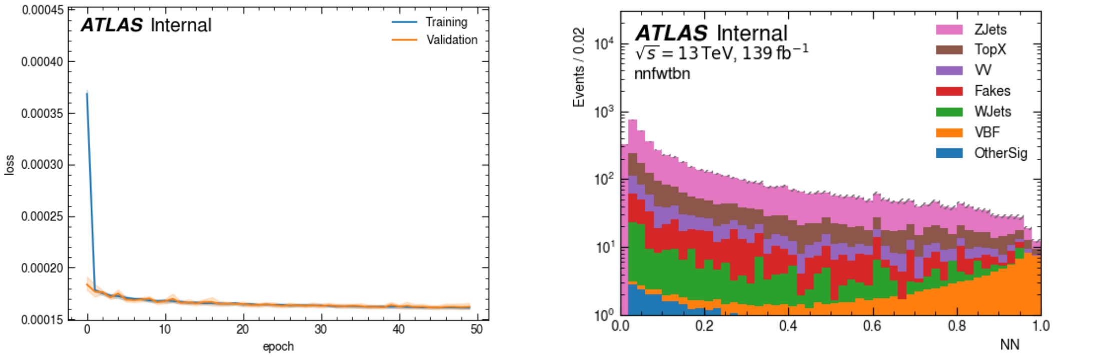
Process	30-30-30	70-70-70	100-100-100
VBFH	56.34 ± 0.28	55.23 ± 0.27	58.71 ± 0.28
Other H	6.48 ± 0.22	6.05 ± 0.21	6.99 ± 0.23
VV	12.86 ± 1.23	11.13 ± 0.99	12.74 ± 1.08
Top+X	39.56 ± 3.71	33.69 ± 3.41	42.81 ± 3.86
W Jets	8.67 ± 4.47	6.26 ± 4.27	7.55 ± 4.34
Zee	-1.15 ± 1.16	-1.15 ± 1.16	-1.15 ± 1.16
Ztt	89.06 ± 4.89	76.85 ± 4.82	91.69 ± 5.16
Zmm	0.27 ± 0.15	0.18 ± 0.13	0.18 ± 0.13
Fakes	12.90 ± 4.37	11.28 ± 4.06	11.28 ± 4.06
Signal/Bkg	0.33	0.38	0.34
Significance	2.71±0.06	2.82±0.07	2.79±0.06



(a) Training and validation loss as a function of the epoch.

(b) The neural network output for different processes plotted on a logarithmic scale.

Figure 29: The training and validation loss and neural network output for the optimized network with a 70-70-70 architecture (L2 $\lambda = 10^{-5}$, η=0.01, batch size = 2000). The total of the background events were normalized to the signal events.



(a) Training and validation loss as a function of the epoch.

(b) The neural network output for different processes plotted on a logarithmic scale.

Figure 30: The training and validation loss and neural network output for the optimized network with a 100-100-100 architecture (L2 $\lambda = 10^{-5}$, η=0.01, batch size = 2000). The total of the background events were normalized to the signal events.

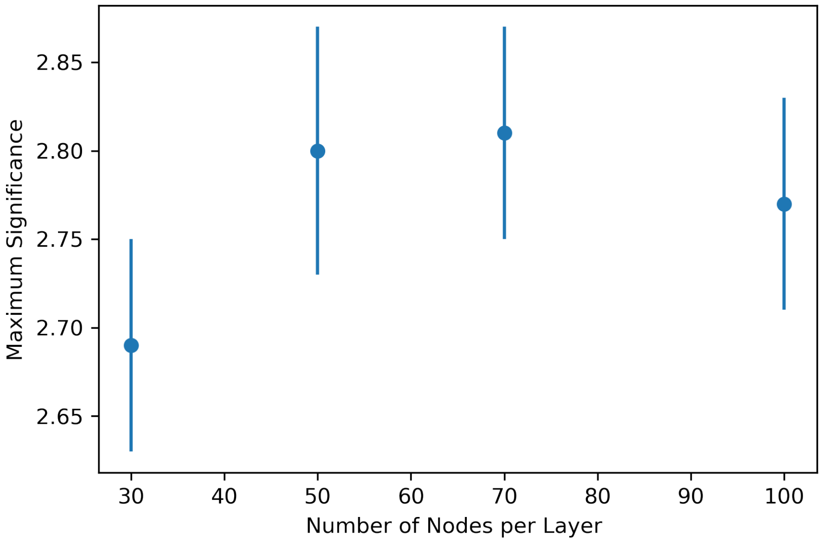


Figure 31: Maximum significance as a function of the number of nodes per layer for a neural network with three layers, learning rate of 0.01 and batch size of 2000 (optimal configuration for 50-50-50 net))

As seen in the figure 31, apart from the network with 30 nodes per layer, increasing the number of nodes did not lead to a significant increase in the performance of the network.

7.3.3 Influence of a non-flat architecture

A neural network with a 100-50-25 architecture was also investigated, its optimization process can be seen above in figures 18 to 44. The training and validation loss as a function of the epoch as well as the output of the network can be seen in figure 26. A table of expected number of events that pass the neural network classification threshold can be seen in table 14. As seen in table 13, the optimal configuration returned a maximum significance of 2.72 ± 0.05, which was not significantly better than that of the 50-50-50 architecture, which was 2.80 ± 0.06. A further investigation which also utilized a 100-50-25 architecture, but where the L2 parameter was scaled to the number of nodes per layer can be seen in section 7.4.

Process	100-50-25
VBFH	49.32 ± 0.25
Other H	8.56 ± 0.25
VV	15.65 ± 1.29
Top	48.41 ± 4.08
W+Jets	9.12 ± 4.47
$Z \to ee$	-1.54 ± 1.24
$Z \to \tau\tau$	131.43 ± 6.30
$Z \to \mu\mu$	1.06 ± 0.81
Fakes	20.93 ± 5.51
Signal/Bkg	0.28
Significance	2.72±0.05

Table 14: The expected number of events for each process that pass the signal classification threshold for the optimized networks with and without scaling the L2 parameter to the number of nodes (100-50-25 architecture, HL variables).

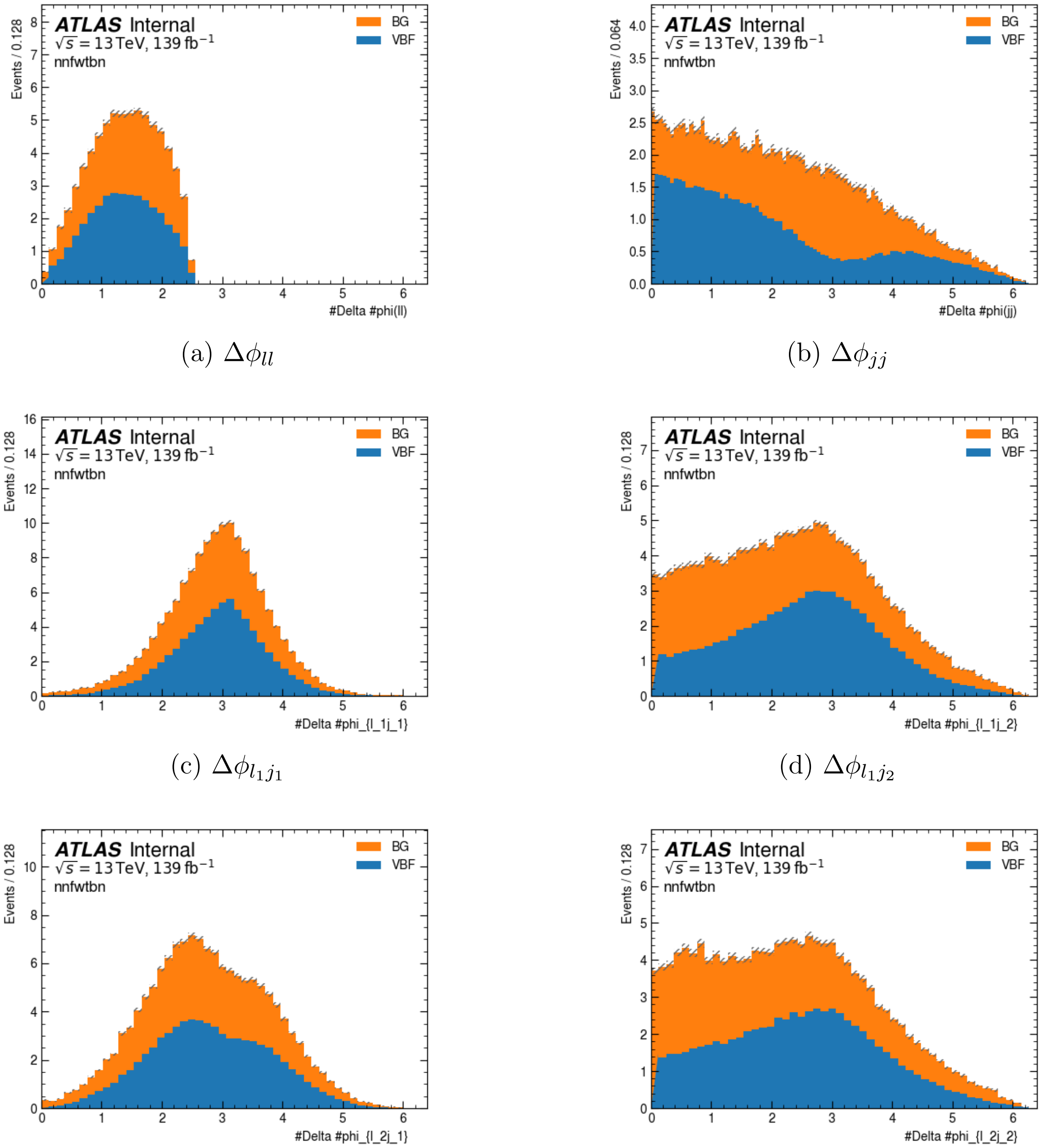
7.4 Influence of Input Variables

The high level variables seen in table 7 are calculated from low level variables, derived directly from the 4-momentum of the leptons and jets. Since all information in the high-level variables derive from low-level variables, a neural network should in principle be able to discern signal from background events using only low level variables.

In order to investigate the influence of different sets of input variables on the performance of the neural network, several groups of input variables were defined. These can be seen in table 15. $(\sum p_T)_{\text{scalar}}$ is the sum of the magnitudes of the transversal momenta of jets and leptons. The distribution of these variables can be seen in figures 32 to 35

Variable Set	DNN Input Variables
Low Level (1)	η_{j_1},η_{j_2},η_{l_2},η_{l_1},ϕ_{j_1},ϕ_{j_2},ϕ_{l_1},ϕ_{l_2}, p_T^{j1},p_T^{j2},p_T^{l1},p_T^{l2}, E_T^{miss}
Low Level (2)	$\Delta\phi_{ll}$, $\Delta\phi_{jj}$,$\Delta\phi_{l_1j_1}$,$\Delta\phi_{l_2j_2}$,$\Delta\phi_{l_1j_2}$,$\Delta\phi_{l_2j_1}$ + Low Level (1) without ϕ's
Low Level (3)	$(\sum p_T)_{\text{scalar}}$ + Low Level (2)
Low Level (4)	$\Delta\phi_{E_T^{\text{miss}},\tau_1}$,$\Delta\phi_{E_T^{\text{miss}},\tau_2}$,$\Delta\phi_{E_T^{\text{miss}},j_1}$,$\Delta\phi_{E_T^{\text{miss}},j_2}$ + Low Level (3)

Table 15: The Variable set used as the input for training the neural network.



(a) $\Delta\phi_{ll}$

(b) $\Delta\phi_{jj}$

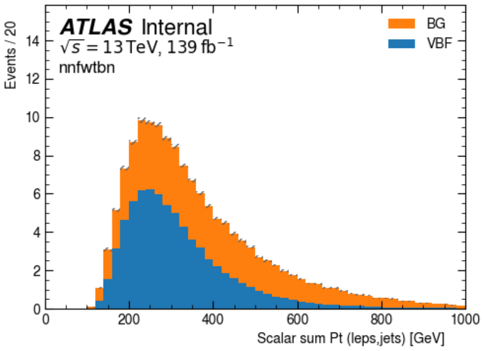
(c) $\Delta\phi_{l_1j_1}$

(d) $\Delta\phi_{l_1j_2}$

(e) $\Delta\phi_{l_2j_1}$

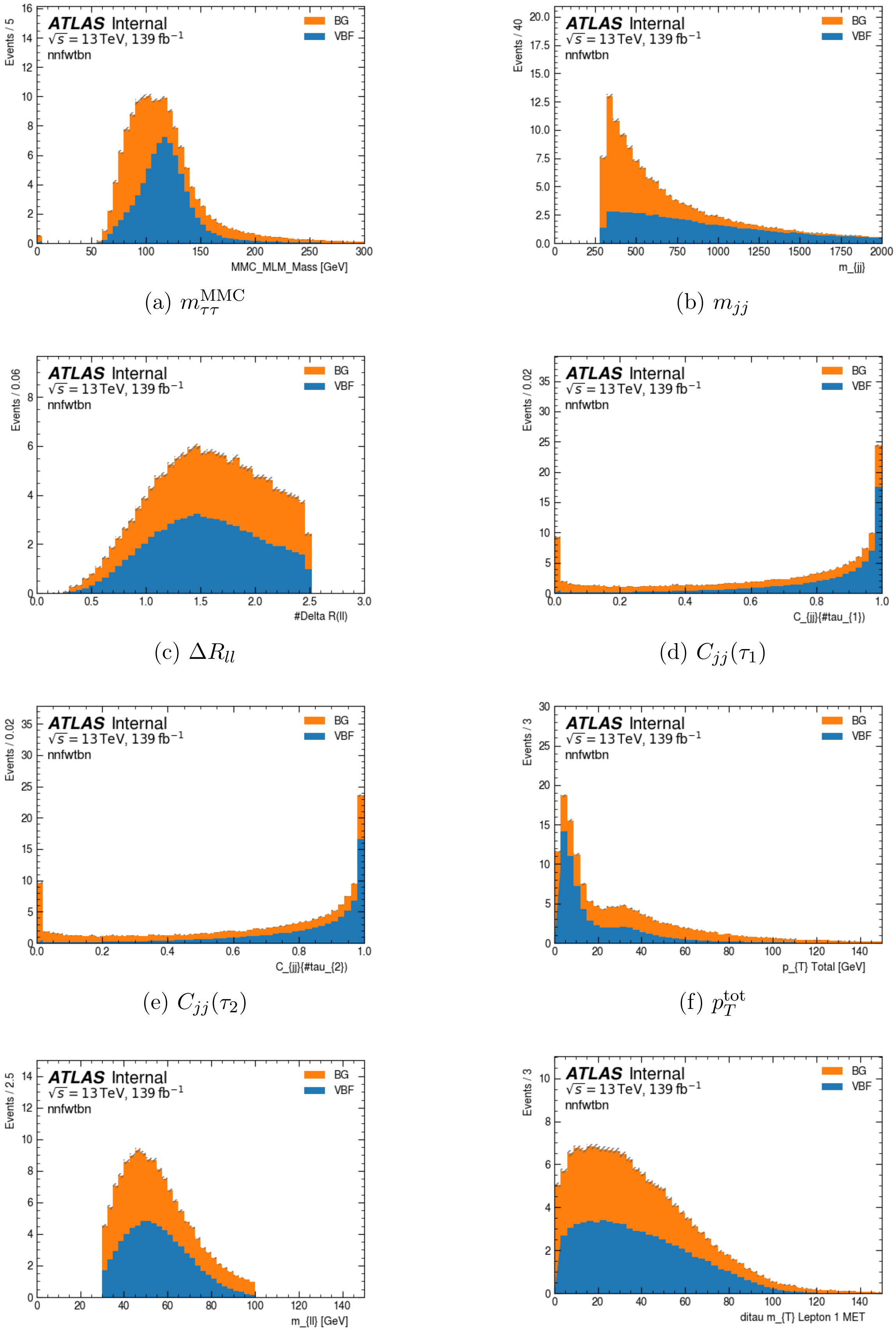
(f) $\Delta\phi_{l_2j_2}$

Figure 32: Distributions of variables in the Low Level (2) variable category (stacked).



(a) $(\sum p_T)_{\text{scalar}}$

Figure 33: Distribution of variables in the Low Level (3) variable category (stacked).



(a) $m_{\tau\tau}^{\mathrm{MMC}}$

(b) m_{jj}

(c) ΔR_{ll}

(d) $C_{jj}(\tau_1)$

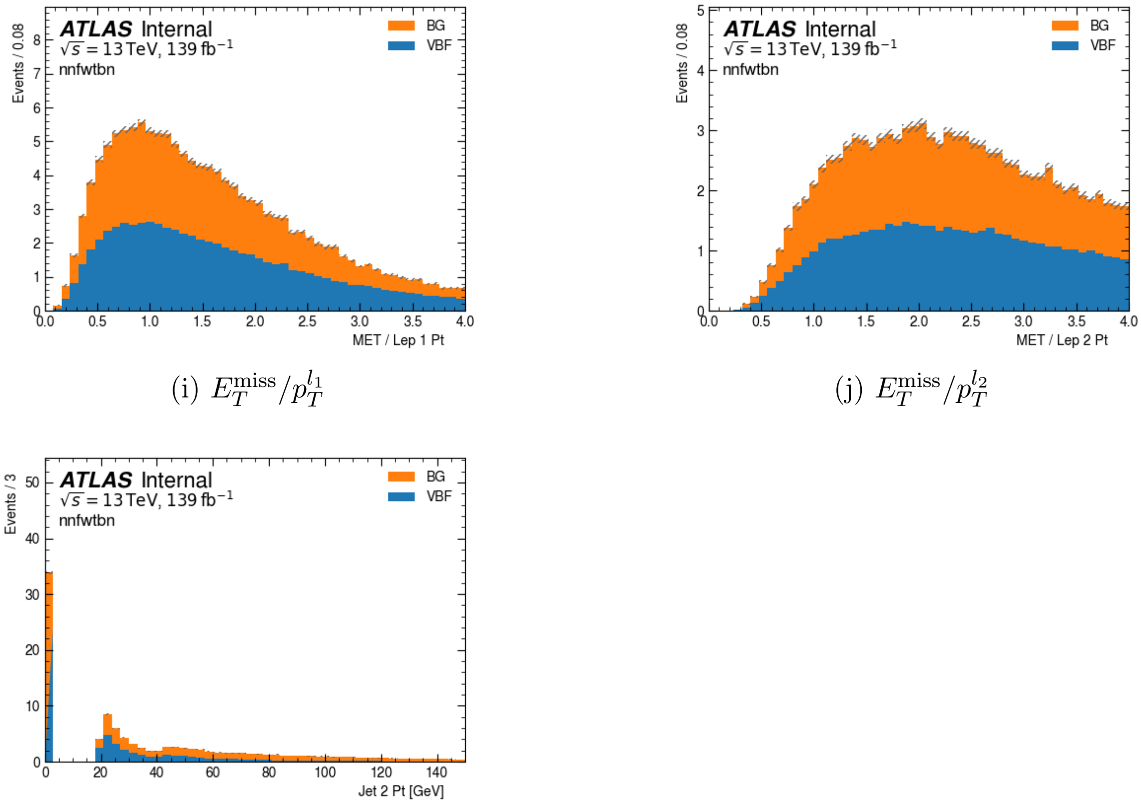
(e) $C_{jj}(\tau_2)$

(f) p_T^{tot}

(g) $m_{\tau\tau}^{\mathrm{vis}}$

(h) $m_T^{\tau_1,E_T^{\mathrm{miss}}}$

Figure 34: Distribution of variables in the High Level variable category (stacked).



(i) $E_T^{\text{miss}}/p_T^{l_1}$

(j) $E_T^{\text{miss}}/p_T^{l_2}$

(k) $p_T^{j_3}$

Figure 35: Distribution of variables in the High Level variable category (stacked).

The full optimization processes were done for the variable category Low Level (1), High Level and lastly with all variables. The influence of the additional low level variables such as low level (2),(3) and (4) were investigated using the optimized low level net. The results of the optimization procedure can be seen from figures 36 to 38. The output of the neural network and training and validation loss as a function of the epoch can be seen in figure 39 to 41, and a summary of the optimized hyperparameters as well as the best significance and optimal NN threshold can be seen in table 19. A table of the expected number of events that pass the signal classification threshold for each process can be seen in table 20.

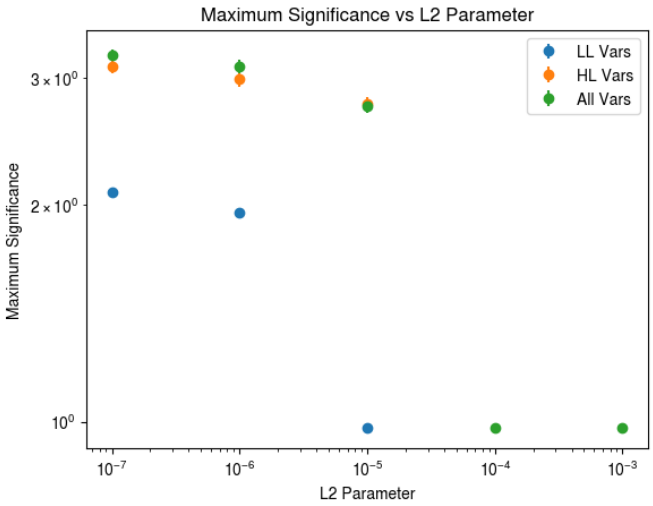


Figure 36: Maximum Significance as a function of the L2 λ Parameter for neural networks utilizing different input variables. The other hyperparameters are: Batch Size = 200, Learning Rate=0.01

Variable Set	Selected L2 λ
LL (1)	10^{-6}
HL	10^{-5}
All	10^{-5}

Table 16: The selected L2 λ parameters for each variable category as determined by their maximum significance.

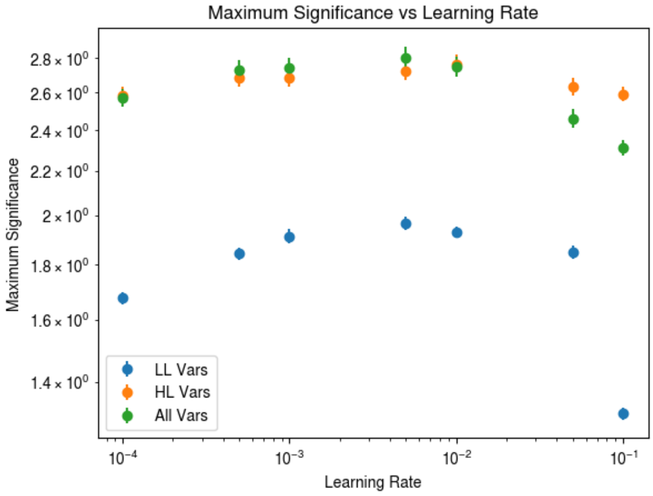


Figure 37: Maximum significance as a function of the learning rate for neural networks utilizing different input variables. The other hyperparameters are as determined in the previous steps.

Variable Set	Selected η
LL (1)	0.005
HL	0.01
All	0.005

Table 17: The selected learning rate for each network architecture as determined by their maximum significance.

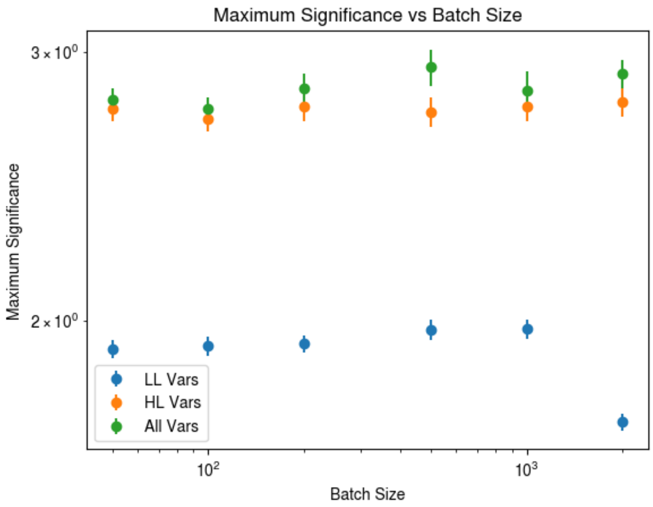
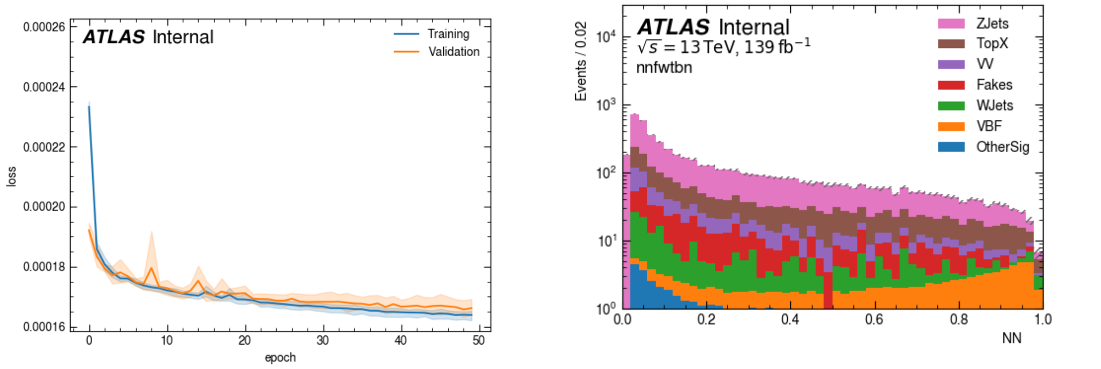


Figure 38: Maximum significance as a function of the batch size for neural networks utilizing different input variables. The other hyperparameters are: Batch Size = 200, and L2 λ parameter as determined in the previous step.

Variable Set	Selected batch size
LL (1)	2000
HL	2000
All	500

Table 18: The selected batch sizes for each network architecture as determined by their maximum significance.



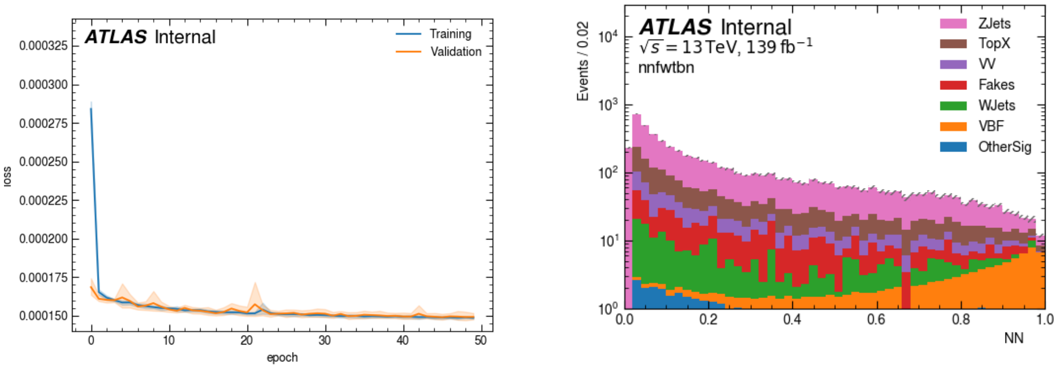
(a) Training and validation loss as a function of the epoch.

(b) The neural network output for different processes plotted on a logarithmic scale (stacked).

Figure 39: The training and validation loss and neural network output for the optimized network utilizing Low Level (1) variables (50-50-50 architecture, L2 $\lambda = 10^{-6}$, η=0.01, batch size = 2000). The total of the background events were normalized to the signal events.

Variable Set	L2 λ	η	Batch Size	Max. Sign.	NN Threshold
LL(1)	10^{-6}	0.01	2000	2.01 ± 0.03	0.80
HL	10^{-5}	0.01	2000	2.80 ± 0.06	0.90
All	10^{-5}	0.005	50	2.93 ± 0.08	0.92

Table 19: The optimized combination of hyperparameters (L2λ, learning rate, batch size) for different network architectures along with their maximum significance and optimal neural network answer threshold



(a) Training and validation loss as a function of the epoch.

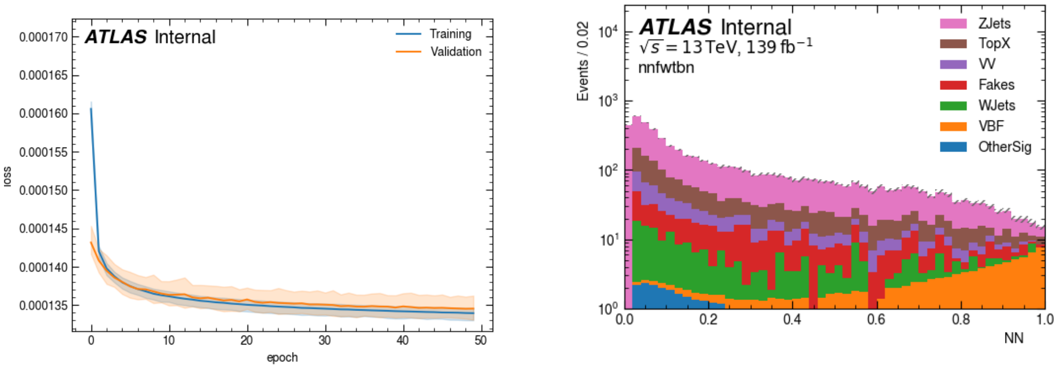
(b) The neural network output for different processes plotted on a logarithmic scale (stacked).

Figure 40: The training and validation loss and neural network output for the optimized network utilizing high level variables (50-50-50 architecture, L2 $\lambda = 10^{-5}$, η=0.01, batch size = 2000). The total of the background events were normalized to the signal events.

As it can be seen in table 19, the neural network utilizing all variables has the highest nominal value for the significance, although it is not significantly higher than that of the low level variables. Furthermore, the networks using only high level or all variables perform significantly better than the network using only low level variables. Performance of the neural network using high level variables was not significantly better than the one using all variables, however for computing efficiency and storage optimization, one may consider using only using high level variables.

7.4.1 Varying the low-level variables

The influence of additional low-level variables were investigated using the optimized low-level neural network above (50-50-50 architecture, L2 $\lambda = 10^{-6}$, $\eta = 0.01$, Batch Size = 200) . Upon the original low-level (1) variable set, additional low-level varible sets were added as input variables for the neural network. A list of low level input variables, organized into variable sets, can be seen in table 21. For each addition, its influence on the maximum significance was investigated. The Low Level (1) variable set was defined as the set of ϕ, η and p_T of jets and leptons as well as the magnitude of te missing transversal energy. The ϕ angle of the leptons and jets were redefined so that the ϕ of the missing energy was at 0, i.e. for every event, for every lepton and jet, ϕ was replaced with $([\phi - \phi_{E_T^{\text{miss}}} + \pi] \mod 2\pi) - \pi$, where the modulo operator ensures that the result is between $-\pi$ and π. The Low Level (2) variable set contains the differences of ϕ between all leptons and jets and variables from Low Level (1). In this case, these variables replace the ϕ's in the Low Level (1) variable set. Low Level (3) contains the variable $(\sum p_T)_{\text{scalar}}$,defined as the scalar sum of all transversal momenta (i.e. of jets, leptons and missing energy) along with those in Low Level (2), and Low Level (4) contains the differences in ϕ between the missing energy and the leptons and jets, along with those in Low Level (3). For the calculation of the $\Delta\phi$'s, the original unrotated ϕ were used. A Summary of the results can be seen in table 21.



(a) Training and validation loss as a function of the epoch.

(b) The neural network output for different processes plotted on a logarithmic scale (stacked).

Figure 41: The training and validation loss and neural network output for the optimized network utilizing all variables (50-50-50 architecture, L2 $\lambda = 10^{-5}$, η=0.005, batch size = 50). The total of the background events were normalized to the signal events.

As it can be seen in table 21, replacing ϕ's with $\Delta\phi$'s between the leptons and jets slightly increased the performance of the net. This is most likely due to the differences in the distribution of $\Delta\phi_{jj}$ for background and VBF signals (see figure 42).

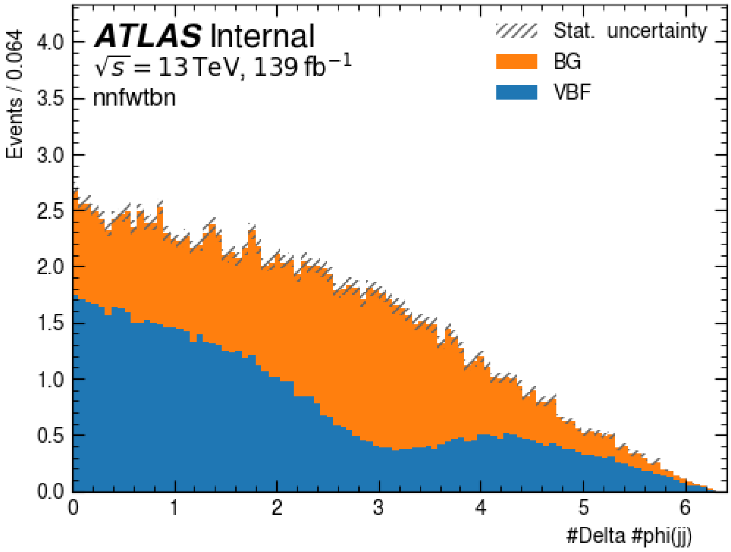


Figure 42: The distribution of $\Delta\phi_{jj}$ for background and VBF signal events (stacked).

However, the biggest improvement in the performance of the neural network was made with the addition of $(\sum p_T)_{\text{scalar}}$. Before the addition of this variable, the maximum significance lied on 2.068 ± 0.034, which was increased to 2.52 ± 0.05 through the addition of this variable. The addition of $\Delta\phi$'s with the missing energy however did not significantly increase the significance. Which is as expected since it can be seen in figure 43, that the distribution of these $\Delta\phi$'s with the missing energy have very similar distributions for signal and background events.

Process	50-50-50 (HL)	50-50-50 (LL)	50-50-50 (All)
VBFH	56.37 ± 0.27	58.77 ± 0.28	36.19 ± 0.21
Other H	6.25 ± 0.22	11.61 ± 0.31	4.12 ± 0.17
VV	10.63 ± 0.95	32.15 ± 1.57	4.42 ± 0.70
Top	34.77 ± 3.43	165.41 ± 7.09	21.76 ± 2.79
W+Jets	7.26 ± 4.34	21.57 ± 5.97	6.43 ± 4.31
$Z \rightarrow ee$	-1.15 ± 1.16	0.41 ± 0.62	0.00 ± 0.00
$Z \rightarrow \tau\tau$	82.25 ± 5.00	281.37 ± 7.53	49.15 ± 3.88
$Z \rightarrow \mu\mu$	0.18 ± 0.13	-0.68 ± 0.97	0.09 ± 0.09
Fakes	14.52 ± 4.66	27.19 ± 5.93	4.30 ± 2.52
Signal/Bkg	0.36	0.11	0.53
Significance	2.80±0.06	2.01±0.03	2.93 ± 0.08

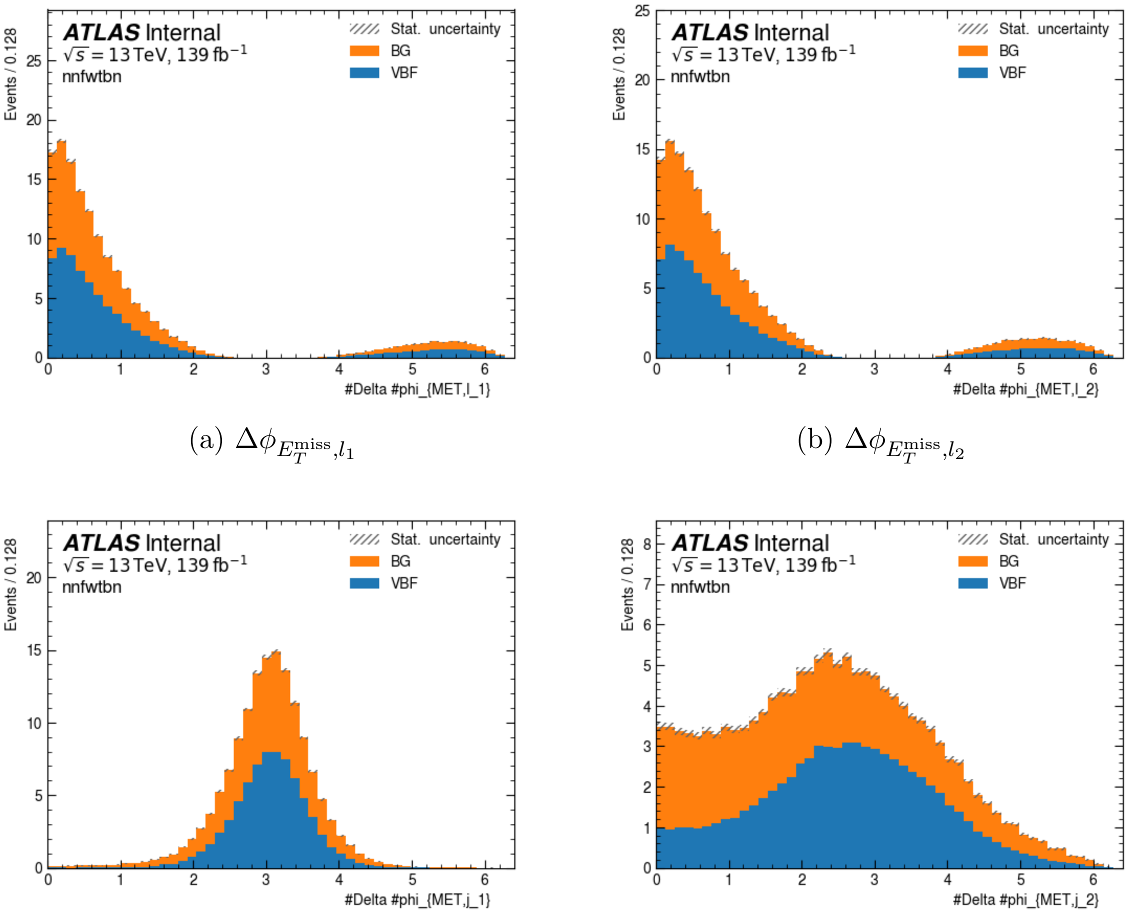
Table 20: The expected number of events for each process that pass the signal classification threshold for the optimized networks utilizing high level, low level and both variable sets (50-50-50 architecture).

Input Variable Sets	Max. Signifiance	Optimal Threshold
Low Level (1)	2.010 ±0.028	0.80
Low Level (2)	2.068 ±0.034	0.86
Low Level (3)	2.52 ± 0.05	0.90
Low Level (4)	2.58±0.05	0.88

Table 21: The maximum significance of the optimized low level neural network for different combinations of input variable sets, defined in table 7

7.5 Scaling the L2 parameter to the number of nodes

A further investigation was done which scaled the L2 parameter proportionally to the number of nodes per layer as suggested by [49], i.e. if the first layer with a hundred nodes had an L2 parameter of $\lambda = 10^{-5}$, a layer with 50 nodes would have an L2 parameter of $\lambda = 5 \cdot 10^{-6}$. This was applied to the network with the 100-50-25 architecture and the optimization process was done. For these optimization processes, only the high level input variables were used. An overview of the optimization process can be seen in figures 44 to 46. A table of the number of expected events that pass the signal classification threshold can be seen in table 27. The optimization process for the network without the L2 parameter scaling is identical to that of section 7.2 but displayed again here for comparison.



(a) $\Delta\phi_{E_T^{\text{miss}},l_1}$

(b) $\Delta\phi_{E_T^{\text{miss}},l_2}$

(c) $\Delta\phi_{E_T^{\text{miss}},j_1}$

(d) $\Delta\phi_{E_T^{\text{miss}},j_2}$

Figure 43: The distributions of $\Delta\phi$'s between the missing energy and leptons and jets for background and signal events (stacked). Background events are normalized to signal events.

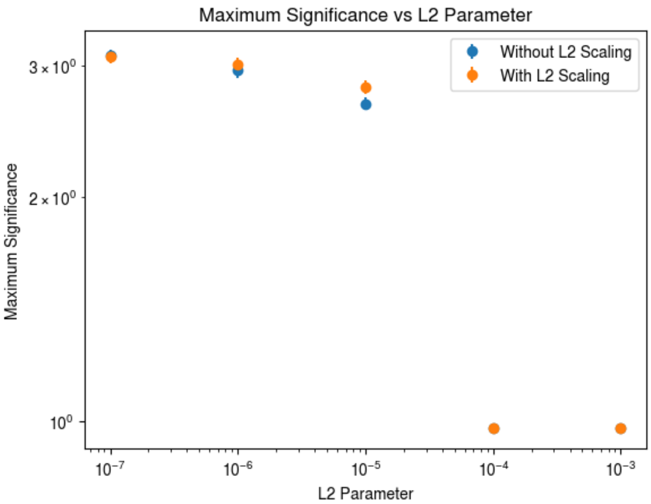


Figure 44: Maximum Significance as a function of the L2 λ Parameter in the first hidden layer for different network architectures. The other hyperparameters are: Batch Size = 200, Learning Rate=0.01.

L2 Scaling	Selected L2 λ
No	10^{-5}
Yes	10^{-5}

Table 22: The selected L2 λ parameters for each variable category as determined by their maximum significance.

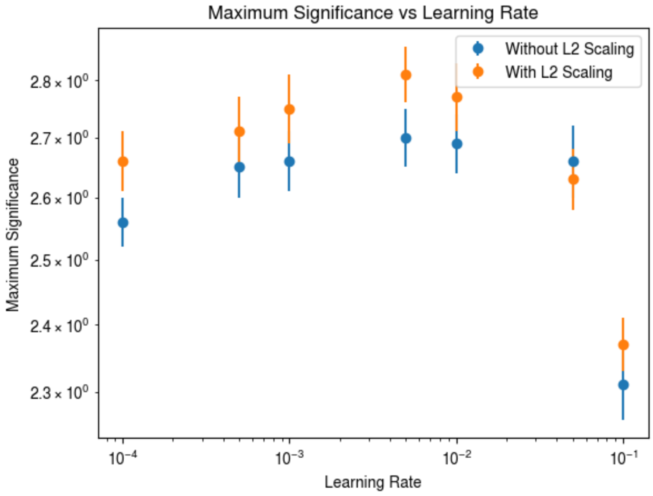


Figure 45: Maximum significance as a function of the learning rate for different network architectures. The other hyperparameters are as determined in the previous steps.

L2 Scaling	Selected η
No	0.005
Yes	0.005

Table 23: The selected learning rate for each network architecture as determined by their maximum significance.

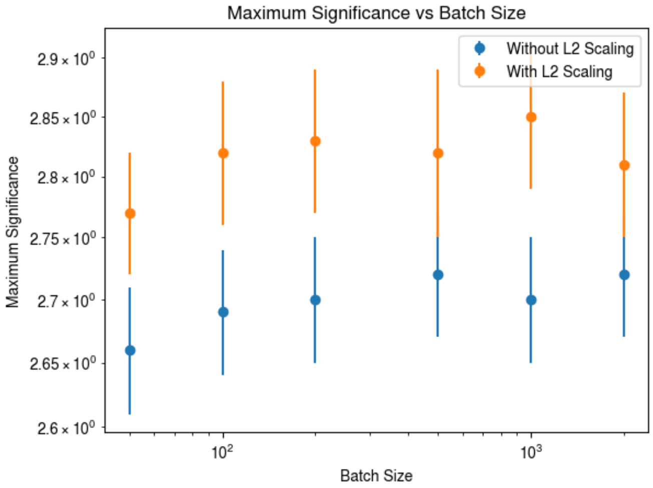


Figure 46: Maximum significance as a function of the batch size for different network architectures. The other hyperparameters are: Batch Size = 200, and L2 λ parameter as determined in the previous step.

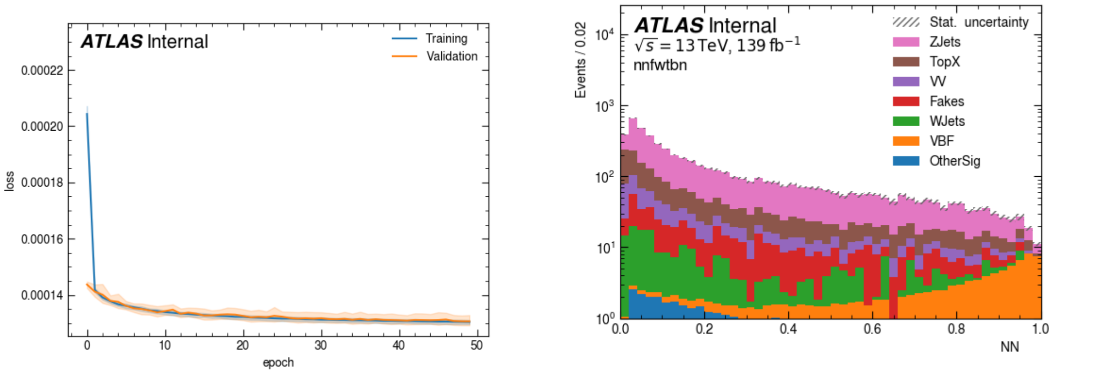
L2 Scaling	Selected Batch Size
No	500
Yes	1000

Table 24: The selected batch sizes for each network architecture as determined by their maximum significance.

It can be seen in table 26 that scaling the L2 parameter to the number of nodes per layer slightly increases the performance of the network from 2.71 ± 0.05 to 2.85 ± 0.06.

7.6 Utilization of multiple output nodes

In this section, the most dominant background processes were determined, and multiple output nodes were set up to differentiate not just signal from background events, but



(a) Training and validation loss as a function of the epoch.

(b) The neural network output for different processes plotted on a logarithmic scale (stacked).

Figure 47: The training and validation loss and neural network output for the optimized network with L2 parameter scaling (50-50-50 architecture, L2 $\lambda = 10^{-5}$, η=0.005, batch size = 50, using high level variables). The total of the background events were normalized to the signal events.

L2 Scaling	L2 λ	η	Batch Size	Max. Significance	Optimal Threshold
No	10^{-5}	0.005	500	2.72 $\pm$0.05	0.90
Yes	10^{-5}	0.005	1000	2.85 $\pm$0.06	0.86

Table 25: The maximum significance for different nets utilizing high level input variables where one scales the L2 parameter to the number of nodes per layer and the other doesn't. Other hyperparameters are as determined in the optimization processes.

Table 26: The expected number of events for each process that pass the signal classification threshold for the optimized networks that scale and don't scale the L2 parameter to the number of nodes per layer (100-50-25 architecture).

also to differentiate different types of background processes as well. In order to identify the dominant background processes, the expected number of events for each process that passes the NN output threshold was calculated. It was consistently observed, that the top background and $Z \to \tau\tau$ processes were the two dominant background processes. An example table of expected number of events for each process that passes the optimal NN threshold with the optimized 50-50-50 neural network using high level input variables can be seen in table 28.

Therefore, the neural network had a total of four output nodes, one for VBF signals, $Z \to \tau\tau$, top background and other background events each. An event was classified as one of these processes above when the output of its corresponding output node was greater than that of the other output nodes. The top background, $Z \to \tau\tau$ and other background events signals were normalized separately to the signal events. The optimization process for a neural network with a 50-50-50 architecture and multiple output nodes using high level variables was done. The results of the optimization can be seen in figures 48 to 50. A table of the expected number of events that pass the signal classification threshold can be seen in table 33. The optimization process of the neural network with the 50-50-50

Process	No L2 Scaling	With L2 Scaling
VBFH	49.32 ± 0.25	58.15 ± 0.28
Other H	8.56 ± 0.25	6.43 ± 0.22
VV	15.65 ± 1.29	10.08 ± 1.00
Top	48.41 ± 4.08	34.26 ± 3.51
W+Jets	9.12 ± 4.47	7.42 ± 4.37
$Z \rightarrow ee$	-1.54 ± 1.24	-1.59 ± 1.24
$Z \rightarrow \tau\tau$	131.43 ± 6.30	88.05 ± 5.04
$Z \rightarrow \mu\mu$	1.06 ± 0.81	0.18 ± 0.13
Fakes	20.93 ± 5.51	11.84 ± 4.24
Signal/Bkg	0.28	0.37
Significance	2.72±0.05	2.85±0.06

Table 27: The expected number of events for each process that pass the signal classification threshold for the optimized networks with and without scaling the L2 parameter to the number of nodes (100-50-25 architecture, HL variables).

VBF $\rightarrow H \rightarrow \tau\tau$	42.66 ± 0.23
VBF$\rightarrow H \rightarrow WW$	13.71 ± 0.15
Other H Production Modes	6.25 ± 0.22
Diboson Production	10.63 ± 0.95
Top Background	34.77 ± 3.43
W+Jets	7.26 ± 4.34
$Z \rightarrow ee$	-1.15 ± 1.16
$Z \rightarrow \tau\tau$	82.25 ± 5.00
$Z \rightarrow \mu\mu$	0.18 ± 0.13
Fakes	14.52 ± 4.66

Table 28: The expected number of events for each process above a cut of 0.90 for the optimized neural network with a 50-50-50 architecture, L2 parameter of $\lambda = 10^{-5}$, Learning rate of 0.01 and batch size of 2000, corresponding to a luminosity of 139 fb^{-1}.

architecture with only two output nodes are also shown for comparison.

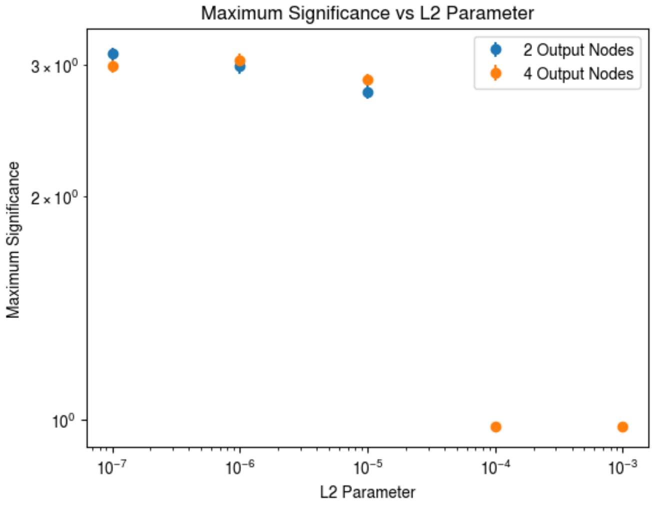


Figure 48: Maximum Significance as a function of the L2 λ Parameter for each neural network with different numbers of output nodes. The other hyperparameters are: Batch Size = 200, Learning Rate=0.01

No. Nodes	Selected L2 λ
2	10^{-5}
4	10^{-5}

Table 29: The selected L2 λ parameters for each neural network with different numbers of output nodes as determined by their maximum significance.

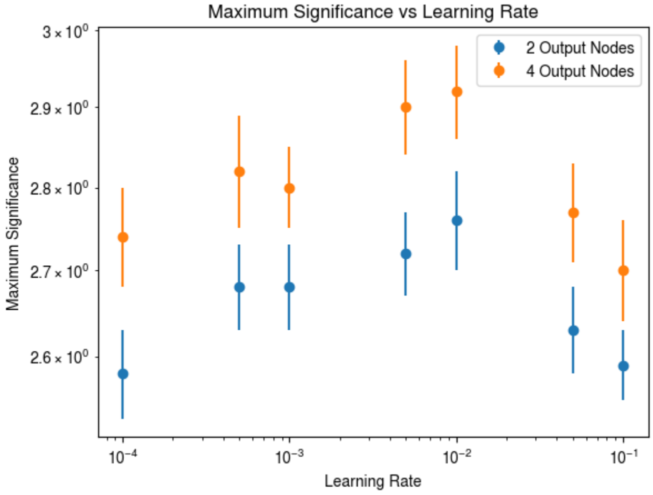


Figure 49: Maximum significance as a function of the learning rate for each neural network with different numbers of output nodes. The other hyperparameters are as determined in the previous steps.

No. Nodes	Selected η
2	0.01
4	0.01

Table 30: The selected learning rate for each neural network with different numbers of output nodes as determined by their maximum significance.

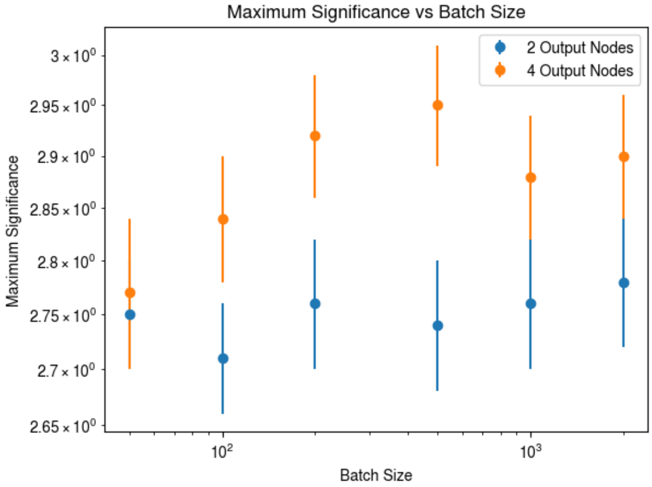


Figure 50: Maximum significance as a function of the batch size for networks with different numbers of output nodes. The other hyperparameters are: Batch Size = 200, and L2 λ parameter as determined in the previous step.

No. Nodes	Selected Batch Size
2	2000
4	500

Table 31: The selected batch sizes for each neural network with different numbers of output nodes as determined by their maximum significance.

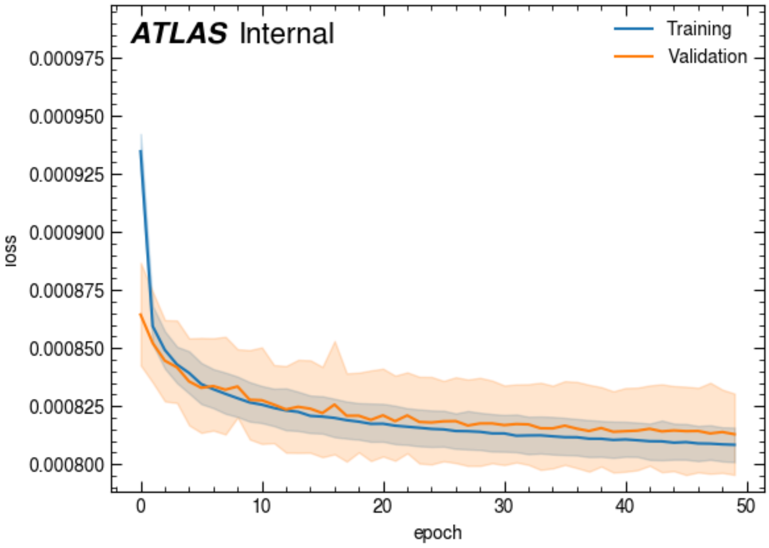


Figure 51: The training and validation loss and neural network output for the optimized network with four output nodes (50-50-50 architecture, L2 $\lambda = 10^{-5}$, η=0.01, batch size = 50, high level variables).

No. Outputs	L2 λ	η	Batch Size	Max. Signifiance	Optimal Threshold
2	10^{-5}	0.01	2000	2.80 ±0.06	0.90
4	10^{-5}	0.01	500	2.95 ±0.06	0.82

Table 32: The maximum significance for different nets utilizing high level input variables with different number of output nodes for different processes.

Process	2 Output Nodes	4 Output Nodes
VBFH	56.37 ± 0.27	45.54 ± 0.25
Other H	6.25 ± 0.22	3.90 ± 0.17
VV	10.63 ± 0.95	4.80 ± 0.62
Top	34.77 ± 3.43	15.14 ± 2.37
W+Jets	7.26 ± 4.34	1.47 ± 0.88
$Z \rightarrow ee$	-1.15 ± 1.16	0.01 ± 0.01
$Z \rightarrow \tau\tau$	82.25 ± 5.00	50.34 ± 3.35
$Z \rightarrow \mu\mu$	0.18 ± 0.13	0.18 ± 0.13
Fakes	14.52 ± 4.66	1.62 ± 1.62
Signal/Bkg	0.36	0.59
Significance	2.80±0.06	2.95±0.06

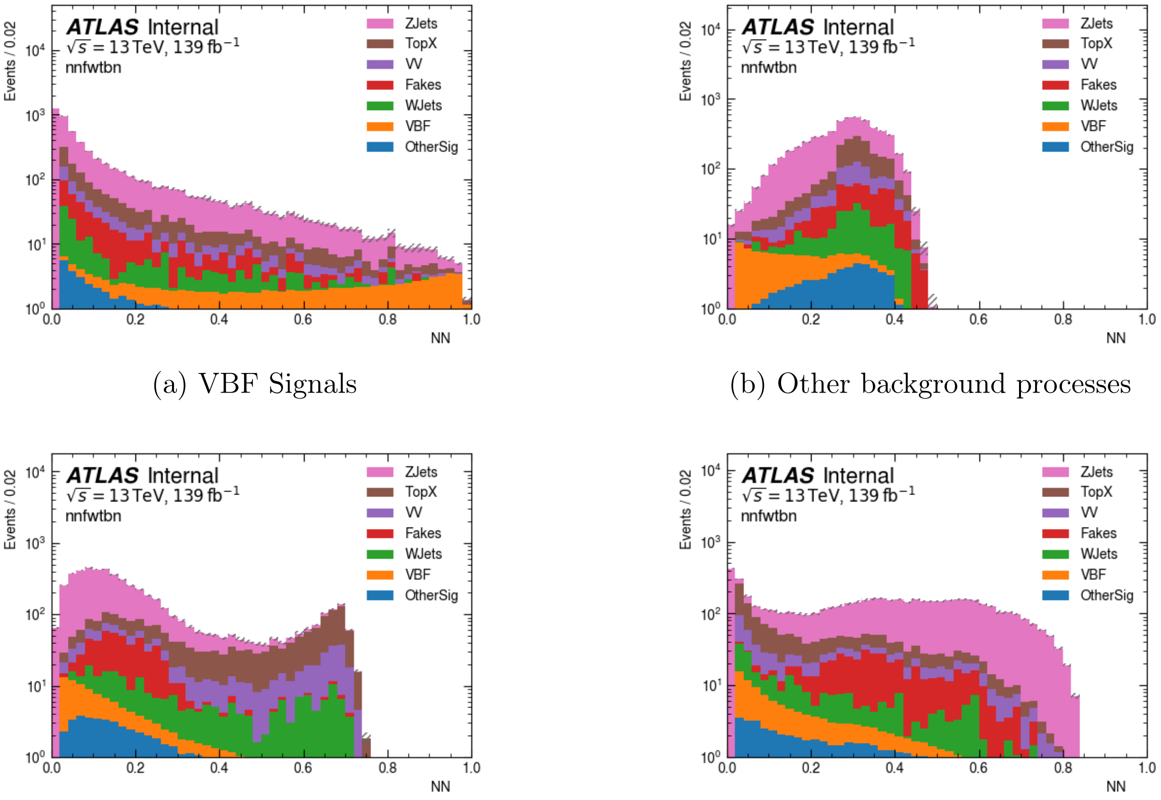
Table 33: The expected number of events for each process that pass the signal classification threshold for the optimized networks with 2 and 4 output nodes utilizing high level variables.

As it can be seen in table 32, increasing the number of nodes to consider different types of background processes has significantly increased the performance of the neural network. With two output nodes, the significance had a maximum value of 2.80 ± 0.06, while with four output nodes 2.95 ± 0.06. However, another important aspect of this neural network is its ability to accurately identify background processes as well. Histograms of the neural network output for each output node, separated into the different processes, can be seen in figure 52.

Even though the differentiation from signal to background has improved, the performance of the classification of the background processes is not ideal. The output of the node for other background processes is limited to 0.5, implying that for a given event, the neural network assigned a maximum probability of 0.5 that the event is from other background processes. This may be due to the low amount of events in the input data of this category, since the dominant background processes have been assigned to other output nodes. Similar phenomena can also be seen from other nodes as well. For the top background node, the output was limited to 0.9, and for $Z \rightarrow \tau\tau$ limited to 0.84. This implies, increasing the amount of output nodes for different processes may be used to improve the discrimination of signal events from background events, but not necessarily improve the classification of events into their respective processes.

7.7 Summary of optimized networks

An overview of the predicted expected number of events that pass the signal threshold for each process can be seen in table 34. It can be seen that the optimized 50-50-50 network utilizing all variables had the highest signal to background ratio and significance. Apart from the network utilizing low level variables, the maximum significances of all networks were between 2.6 and 2.95. It is also evident, that certain network configurations are able to significantly suppress $Z \rightarrow \tau\tau$ signals, which is one of the dominant background processes. For example, the 100-50-25 architecture using high level variables predicted that 131 $Z \rightarrow \tau\tau$ events pass the event threshold, while for the 50-50-50 architecture with



(a) VBF Signals

(b) Other background processes

(c) Top background

(d) $Z \rightarrow \tau\tau$

Figure 52: Distribution of the output of each output node (stacked)

all variables, only 49 events pass the threshold.

Process	50-50 (HL)	50-50-50 (HL)	50-50-50-50 (HL)	50-50-50 (LL)	100-50-25 (HL)	50-50-50 (All)	100-50-25* (HL)	50-50-50** (HL)
VBFH	61.36 ± 0.29	56.37 ± 0.27	47.92 ± 0.24	58.77 ± 0.28	49.32 ± 0.25	36.19 ± 0.21	58.15 ± 0.28	45.54 ± 0.25
Other H	7.51 ± 0.24	6.25 ± 0.22	8.25 ± 0.25	11.61 ± 0.31	8.56 ± 0.25	4.12 ± 0.17	6.43 ± 0.22	3.90 ± 0.17
VV	16.17 ± 1.28	10.63 ± 0.95	16.15 ± 1.30	32.15 ± 1.57	15.65 ± 1.29	4.42 ± 0.70	10.08 ± 1.00	4.80 ± 0.62
Top	48.34 ± 4.10	34.77 ± 3.43	49.11 ± 4.12	165.41 ± 7.09	48.41 ± 4.08	21.76 ± 2.79	34.26 ± 3.51	15.14 ± 2.37
W+Jets	10.40 ± 4.91	7.26 ± 4.34	9.66 ± 4.52	21.57 ± 5.97	9.12 ± 4.47	6.43 ± 4.31	7.42 ± 4.37	1.47 ± 0.88
$Z \rightarrow ee$	-1.10 ± 1.16	-1.15 ± 1.16	-1.15 ± 1.16	0.41 ± 0.62	-1.54 ± 1.24	0.00 ± 0.00	-1.59 ± 1.24	0.01 ± 0.01
$Z \rightarrow \tau\tau$	114.29 ± 5.52	82.25 ± 5.00	119.18 ± 5.86	281.37 ± 7.53	131.43 ± 6.30	49.15 ± 3.88	88.05 ± 5.04	50.34 ± 3.35
$Z \rightarrow \mu\mu$	0.18 ± 0.13	0.18 ± 0.13	0.27 ± 0.15	-0.68 ± 0.97	1.06 ± 0.81	0.09 ± 0.09	0.18 ± 0.13	0.18 ± 0.13
Fakes	21.42 ± 5.47	14.52 ± 4.66	18.82 ± 5.30	27.19 ± 5.93	20.93 ± 5.51	4.30 ± 2.52	11.84 ± 4.24	1.62 ± 1.62
Sum Bkg.	217 ± 10	155 ± 9	220 ± 10	539 ± 13	234 ± 11	90 ± 7	157 ± 9	77 ± 5
Signal/Bkg	0.29	0.36	0.21	0.11	0.28	0.53	0.37	0.59
Significance	2.65±0.05	2.80±0.06	2.72±0.05	2.01±0.03	2.72±0.05	2.93 ± 0.08	2.85±0.06	2.95±0.06

Table 34: A summary of the number of expected events that pass the signal classification threshold for different processes for different optimized neural networks as well as their signal to background ratio and significance.(* With L2 parameter scaling, **With four output nodes).

8 Conclusion

The goal of this thesis was to optimize a neural network to discriminate VBFH $H \rightarrow \tau\tau \rightarrow e\mu 4\nu$ events from background processes for data collected in pp-collisions during Run-2 of the LHC with the ATLAS detector. The effect of the variation in the L2 parameter, learning rate, batch size, input variables, and network architecture on the performance of the network was investigated, where the performance of a neural network is determined by its highest significance, defined as $s/\sqrt{s+b}$. A standardized optimization process of the L2 parameter, learning rate and batch size was implemented to find the combination which results in the highest significance for a given neural network architecture.

Trying to find the optimal combination of hyperparameters for different network architectures yielded a result that the 50-50-50, 70-70-70, 100-100-100 architecture perform similarly albeit better than the 30-30-30 architecture, with maximum significances of 2.80±0.07, 2.81±0.06, 2.77±0.06, and 2.69±0.06 respectively. This implies that simply increasing the number of nodes per layer does not improve the performance of the neural network. It was also shown that having more layers does not increase the significance. The optimized neural network with 2, 3 and 4 layers yielded significances of 2.65±0.05, 2.70±0.06 and 2.72±0.05 respectively. Furthermore, an investigation into a non-flat architecture of 100-50-25 showed that it did not discriminate significantly better than the best flat architectures.

Another investigation into the influence of the type of input variables on the performance of the neural network suggests that using both high and low variables is slightly better than using only high-level variables, while both being far better than utilizing low-level variables only. However, the performance of the neural network using low-level variables can be significantly improved by adding the variable $(\sum p_T)_{\text{scalar}}$, while adding other variables such as $\Delta\phi$'s between jets, leptons and missing energy did not significantly improve the performance of the network.

Increasing the amount of output nodes to consider different types of background processes, namely the top background and $Z \rightarrow \tau\tau$, significantly improved the discrimination of signal events from background events with a significance of 2.95±0.05 as compared to 2.80±0.06 for the best neural network with only 2 output nodes and using high level variables. However the accuracy of the classification of the background events is not ideal.

Lastly, it was consistently observed during the optimization processes, that an L2 parameter of $\lambda = 10^{-5}$ was shown to most often minimize the amount of overtraining and at the same time make sure that the network is able to actually learn. In all cases, learning rates of 0.005 or 0.01 were shown to yield best performances. The influence of the batch size on the performance of the network was often insignificant, where no single batch size would perform better than any other.

This implies, as far as the investigation of this thesis is concerned, the ideal neural network for discerning VBF signal events from background events would be a neural network with a 50-50-50 or 70-70-70 architecture with multiple output nodes utilizing all or high-level variables. The maximum significance for an optimized neural network with a 50-50-50 architecture with multiple output nodes utilizing high level variables is 2.95 ± 0.05 with a signal to background ratio of 0.59.

References

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

[2] S. L. Glashow, *Partial-symmetries of weak interactions*, Nuclear Physics **22** (1961) 579.

[3] A. Salam, “Weak and Electromagnetic Interactions”, *Elementary particle theory, Relativistic groups and analyticity*, Proceedings of the Eighth Nobel Symposium (Aspenäsgarden, Lerum, May 19–25, 1968), ed. by N. Svartholm, Almquist & Wiksell, 1968 367.

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

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

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

[7] *CERN experiments observe particle consistent with long-sought Higgs boson*, 2012, URL: `https://home.cern/news/press-release/cern/cern-experiments-observe-particle-consistent-long-sought-higgs-boson`.

[8] F. Abe et al., *Observation of Top Quark Production in $p\bar{p}$ Collisions with the Collider Detector at Fermilab*, Physical Review Letters **74** (1995) 2626, ISSN: 1079-7114, URL: `http://dx.doi.org/10.1103/PhysRevLett.74.2626`.

[9] K. Kodama et al., *Final tau-neutrino results from the DONuT experiment*, Physical Review D **78** (2008), ISSN: 1550-2368, URL: `http://dx.doi.org/10.1103/PhysRevD.78.052002`.

[10] Particle Data Group, *2018 Particle Physics Booklet*, 2018.

[11] *File:HiggsBR.svg*, URL: `https://commons.wikimedia.org/wiki/File:HiggsBR.svg`.

[12] D. de Florian et al., *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, 2016, arXiv: `1610.07922 [hep-ph]`.

[13] M. Rauch, *Vector-Boson Fusion and Vector-Boson Scattering*, 2016, arXiv: `1610.08420 [hep-ph]`.

[14] The ATLAS Collaboration, *Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment*, European Physical Journal (2016).

[15] *File:Artificial neural network.svg*, 2006, URL: `https://commons.wikimedia.org/wiki/File:Artificial_neural_network.svg`.

[16] J. Kiefer and J. Wolfowitz, *Stochastic Estimation of the Maximum of a Regression Function*, The Annals of Mathematical Statistics **23** (1952) 462, ISSN: 00034851, URL: `http://www.jstor.org/stable/2236690`.

[17] D. P. Kingma and J. L. Ba,
"ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION", 2017.

[18] *The Large Hadron Collider*,
URL: `https://home.cern/science/accelerators/large-hadron-collider`.

[19] M. Schott and M. Dunford,
Review of single vector boson production in pp collisions at $\sqrt{s} = 7$ TeV,
The European Physical Journal C **74** (2014), ISSN: 1434-6052,
URL: `http://dx.doi.org/10.1140/epjc/s10052-014-2916-1`.

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

[21] M. Aharrouche et al., *Energy linearity and resolution of the ATLAS electromagnetic barrel calorimeter in an electron test-beam*,
Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **568** (2006) 601,
ISSN: 0168-9002, URL: `http://dx.doi.org/10.1016/j.nima.2006.07.053`.

[22] G. Aad et al., *Readiness of the ATLAS Tile Calorimeter for LHC collisions*,
The European Physical Journal C **70** (2010) 1193, ISSN: 1434-6052,
URL: `http://dx.doi.org/10.1140/epjc/s10052-010-1508-y`.

[23] J. Snuverink, "The ATLAS Muon Spectrometer: Commissioning and Tracking",
PhD thesis: Twente U., Enschede, 2009.

[24] M. Aaboud et al.,
Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton–proton collision data at $\sqrt{s} = 13$ TeV,
The European Physical Journal C **79** (2019), ISSN: 1434-6052,
URL: `http://dx.doi.org/10.1140/epjc/s10052-019-7140-6`.

[25] G. Aad et al., *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s}$ =13 TeV*,
The European Physical Journal C **76** (2016), ISSN: 1434-6052,
URL: `http://dx.doi.org/10.1140/epjc/s10052-016-4120-y`.

[26] M. Cacciari, G. P. Salam, and G. Soyez, *The anti-ktjet clustering algorithm*,
Journal of High Energy Physics **2008** (2008) 063, ISSN: 1029-8479,
URL: `http://dx.doi.org/10.1088/1126-6708/2008/04/063`.

[27] G. Aad et al., *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1*, The European Physical Journal C **77** (2017),
ISSN: 1434-6052, URL: `http://dx.doi.org/10.1140/epjc/s10052-017-5004-5`.

[28] P. Nason,
A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms,
Journal of High Energy Physics **2004** (2004) 040, ISSN: 1029-8479,
URL: `http://dx.doi.org/10.1088/1126-6708/2004/11/040`.

[29] P. Nason and C. Oleari, *NLO Higgs boson production via vector-boson fusion matched with shower in POWHEG*, Journal of High Energy Physics **2010** (2010),
ISSN: 1029-8479, URL: `http://dx.doi.org/10.1007/JHEP02(2010)037`.

[30] S. Frixione, P. Nason, and C. Oleari, *Matching NLO QCD computations with parton shower simulations: the POWHEG method*, Journal of High Energy Physics **2007** (2007) 070, ISSN: 1029-8479, URL: http://dx.doi.org/10.1088/1126-6708/2007/11/070.

[31] S. Alioli, P. Nason, C. Oleari, and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, Journal of High Energy Physics **2010** (2010), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP06(2010)043.

[32] K. Hamilton, P. Nason, and G. Zanderighi, *MINLO: multi-scale improved NLO*, Journal of High Energy Physics **2012** (2012), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP10(2012)155.

[33] J. Butterworth et al., *PDF4LHC recommendations for LHC Run II*, Journal of Physics G: Nuclear and Particle Physics **43** (2016) 023001, ISSN: 1361-6471, URL: http://dx.doi.org/10.1088/0954-3899/43/2/023001.

[34] T. Sjöstrand et al., *An introduction to PYTHIA 8.2*, Computer Physics Communications **191** (2015) 159, ISSN: 0010-4655, URL: http://dx.doi.org/10.1016/j.cpc.2015.01.024.

[35] G. Luisoni, P. Nason, C. Oleari, and F. Tramontano, *HW ±/HZ + 0 and 1 jet at NLO with the POWHEG BOX interfaced to GoSam and their merging within MiNLO*, Journal of High Energy Physics **2013** (2013), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP10(2013)083.

[36] R. D. Ball et al., *Unbiased global determination of parton distributions and their uncertainties at NNLO and at LO*, Nuclear Physics B **855** (2012) 153, ISSN: 0550-3213, URL: http://www.sciencedirect.com/science/article/pii/S0550321311005463.

[37] The ATLAS Collaboration, *Measurement of $W^{\pm}$ and Z-boson production cross sections in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Physics Letters B (2016).

[38] S. Höche, F. Krauss, M. Schönherr, and F. Siegert, *QCD matrix elements + parton showers. The NLO case*, Journal of High Energy Physics **2013** (2013), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP04(2013)027.

[39] L. L. Iglesias, *Diboson production at the LHC*, Proceedings of Science (2016).

[40] R. D. Ball et al., *Parton distributions for the LHC run II*, Journal of High Energy Physics **2015** (2015), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP04(2015)040.

[41] The CMS Collaboration, *Measurement of the top quark pair production cross section in dilepton final states containing one τ lepton in pp collisions at $\sqrt{s} = 13$ TeV*, Journal of High Energy Physics (2020).

[42] D. Mueller, *Measurement of the single top quark and antiquark production cross sections in the t channel and their ratio at 13 TeV*, 2019, arXiv: 1901.05247 [hep-ex].

[43] A. M. Sirunyan et al., *Measurement of the production cross section for single top quarks in association with W bosons in proton-proton collisions at* $\sqrt{s} = 13$ *TeV*, Journal of High Energy Physics **2018** (2018), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP10(2018)117.

[44] E. Re, *Single-top Wt-channel production matched with parton showers using the POWHEG method*, The European Physical Journal C **71** (2011), ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-011-1547-z.

[45] M. Bahmani,
Data-driven estimation of fake τ *background in Higgs searches in ATLAS*, tech. rep. ATL-PHYS-PROC-2019-006, CERN, 2019, URL: https://cds.cern.ch/record/2653533.

[46] G. Aad et al., *Test of CP invariance in vector-boson fusion production of the Higgs boson in the* $H \to \tau\tau$ *channel in proton–proton collisions at s=13TeV with the ATLAS detector*, Physics Letters B **805** (2020) 135426, ISSN: 0370-2693, URL: http://dx.doi.org/10.1016/j.physletb.2020.135426.

[47] R. Ellis, I. Hinchliffe, M. Soldate, and J. Van Der Bij,
Higgs decay to $\tau^+\tau^-$*A possible signature of intermediate mass Higgs bosons at high energy hadron colliders*, Nuclear Physics B **297** (1988) 221, ISSN: 0550-3213, URL:
http://www.sciencedirect.com/science/article/pii/0550321388900193.

[48] A. Elagin, P. Murat, A. Pranko, and A. Safonov,
A new mass reconstruction technique for resonances decaying to,
Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **654** (2011) 481, ISSN: 0168-9002, URL: http://dx.doi.org/10.1016/j.nima.2011.07.009.

[49] J. Erdmann, T. Kallage, K. Kröninger, and O. Nackenhorst,
From the bottom to the top—reconstruction of t t- events with deep learning, Journal of Instrumentation **14** (2019) P11015, ISSN: 1748-0221, URL: http://dx.doi.org/10.1088/1748-0221/14/11/P11015.

Zusammenfassung

In dieser Bachelorarbeit wird die Benutzung von neuronalen Netzwerken für die Selektion von $H \rightarrow \tau\tau \rightarrow e\mu 4\nu$ Zerfallsereignissen des durch Vektorbosonfusion produzierten Higgs-Bosons im Run-2 vom ATLAS Experiment mit $\mathcal{L} = 139\text{fb}^{-1}$ untersucht. Zuerst wurde eine Vorselektion auf die Eingangsdaten angewandt, um das Signal-zu-Untergrund Verhältnis zu vergrößern. Die Monte-Carlo simulierten Daten werden dann benutzt, um neuronale Netzwerke zu trainieren. Der Einfluss von verschiedenen Hyperparametern wie die Lernrate, Batch Size, L2 Parameter, Netzwerkarchitektur und Eingangsvariablen auf die Trennfähigkeit des Netzwerkes wird untersucht, um eine optimale Kombination von Hyperparametern zu finden. Die beste Signifikanz wird von einem Netzwerk mit einer 50-50-50 Architektur mit vier Ausgangsknoten, das beide low-level und high-level Variablen verwendet, geliefert.

Acknowledgments

I would like to thank Prof. Dr. Markus Schumacher for giving me the opportunity to write my bachelor's thesis with his research group as well as supporting me during the entire course of the writing. I would also like to thank Dr. David Hohn for helping me out all the time and making time for me nearly every morning even though he must have been very busy with his own work. Lastly, I would like to thank Dr. Valerie Lang for her guidance in the beginning of this journey.

During my time in the Schumacher AG, I learned many things, not just about physics or how to write a thesis, but on how to work on such projects alone and with other people. I feel like I was able to learn more about my own strengths and weaknesses as a person, and I hope to improve upon them. I will be looking forward to seeing them again when I start my master's program.