Identification of highly boosted $H \rightarrow \gamma\gamma$ decays with the ATLAS detector using deep neural networks

Nathaniel Hey

Master of Science
Data Science
School of Informatics
University of Edinburgh
2023

Abstract

This thesis introduces two jet tagging algorithms to identify highly boosted $H \rightarrow \gamma\gamma$ decays using the ATLAS detector at the LHC. Based on the Deep Neural Network (DNN) architecture, the first algorithm's performance is comparable to an existing algorithm designed for highly boosted $Z \rightarrow e^+e^-$ decays. The DNN jet tagger is also multifunctional and highly effective for identifying $Z \rightarrow e^+e^-$ decays. Notably, it displayed enhanced rejection rates for background $(\tau)\tau$-jets. The second algorithm leverages an Adversarial Neural Network (ANN) architecture for mass-decorrelated classification. While it exhibited a slight performance decrease compared to the DNN-based tagger, it demonstrated a 27.8% reduction in mutual information between the mass feature and scalar discriminant metric, substantiating its capability for mass-decorrelated jet identification.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Nathaniel Hey*)

Acknowledgements

I would like to express my thanks to my supervisors, who have been supportive and flexible in my unique situation while writing this dissertation. As well as my academic advisor, who was of immense help in navigating this.

Of course, my parents and grandparents deserve my immense gratitude for all the support they have provided me throughout this work and my higher education.

Table of Contents

Chapter 1

Introduction

The twentieth century saw the successful development of the Standard Model (SM) of particle physics, one of the most experimentally verified physical theories ever developed [1, 2, 3, 4, 5]. The Standard Model predicts the existence of the Higgs boson [6, 7, 8] which for nearly five decades evaded discovery until 2012. The discovery was made as a significant part of the Large Hadron Collider (LHC) research programme and was a significant motivating reason for constructing the world's largest particle collider [9].

The existence of the Higgs boson is implied by the Brout-Englert-Higgs mechanism. This is the mechanism that spontaneously breaks the gauge symmetry that governs electroweak interactions [1, 8, 10]. The electroweak symmetry breaking allows other fundamental particles to acquire mass as they interact with the Higgs field [6, 11], the Higgs boson is an excitation of this field [6]. The first run of the LHC at the centre-of-mass energy $\sqrt{s} = 7$ TeV found the mass, parity, spin and decay modes to be as predicted by the Standard Model [12, 13, 14, 15, 16, 17, 18, 19]. LHC runs 2 and 3 are at greater energies, $\sqrt{s} = 13$ TeV and 13.6 TeV respectively. New physics within the sensitivity of the data collected at these runs may include particles predicted by the minimal supersymmetric extension of the Standard Model [20]. The increasingly large statistics generated in these runs could also produce changes and more detailed descriptions of the observables measured in the first LHC run.

Highly precise measurements and analysis systems are required to study any new physics. The first component is the physical detector that measures and records particle collision events at the LHC. The LHC has two general-purpose detectors: the CMS detector [21] and the ATLAS detector [9]. This thesis focuses on improving the analysis of $H \to \gamma\gamma$ run 2 data captured by the ATLAS detector. Identifying such decays is done by analysing the properties of jets that are clustered using the anti-k_T algorithm [22]. This field is called jet tagging within high-energy particle physics; different classification algorithms are used to tag the jets. Jet tagging $H \to \gamma\gamma$ decays are the focus for several reasons. These decays do not have a jet tagging algorithm of the types developed here. Secondly, the available data is of high quality and quantity, and finally, the interest in increasingly detailed descriptions of the Higgs boson is very high due to its relation to the mechanism by which other particles possess mass. For these reasons, developing $H \to \gamma\gamma$ jet tagging algorithms for the ATLAS detector may prove useful tools for physicists at the LHC.

1.1 The ATLAS Detector at the LHC

When the LHC started operations at CERN in 2008, it was the largest particle collider ever constructed [9] and remains so over a decade later. The LHC accelerates groups of up to 10^{11} protons and collides these proton groups together in pp collisions at a rate of 40 million collisions per second [9]. The LHC has been designed to collide protons with a maximum centre-of-mass energy of 14 TeV [9]. At such high collision energies, the protons are accelerated to 99.999999% of the speed of light[1]. This causes the decay products to become highly collimated [23]. Resolving the strongly boosted decay products requires highly specialised detection hardware and software.

The state-of-the-art hardware that forms the centre of this work is the ATLAS (A Toroidal LHS ApparatuS) detector[2] [9], the device that generates data from collisions that will be the focus of this paper. The multipurpose ATLAS detector nearly surrounds the collision point in cavern 1 of the LHC [9, 24]. It comprises several critical components designed for high-precision measurements, as seen in the cut-away model in Figure 1.1. First, an inner tracking detector is surrounded by a solenoid that provides a powerful axial magnetic field that enables the tracking of charged particles in conjunction with silicon-pixel detectors [25] and the transition radiation tracker (TRT) [9]. The TRT is vital as it provides electron identification information [24]. The next major system within the ATLAS detector is the calorimeter system, which consists of electromagnetic and hadron calorimeters. Electromagnetic calorimetry is

[1] Where c is the speed of light in natural units, therefore $c = 1$.

[2] ATLAS has a right-handed coordinate system that originates at the interaction point (IP) in the detector's centre [23]. From the IP, the z-axis follows the beam pipe; the x and y-axes point to the LHC ring centre and upwards, respectively [23]. Cylindrical coordinates (r, ϕ) measure in the transverse plane with ϕ as the azimuthal angle about the z-axis [23]. Pseudorapidity (η) is in terms of the polar angle (θ), $\eta = -\ln\tan(\theta/2)$, and angular distance is $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ [23].

achieved with high-granularity lead/liquid-argon (LAr) and copper/LAr calorimeter modules depending on location within the ATLAS detector [9]. Hadron calorimetry uses steel/scintillator-tile calorimeters and tungsten/LAr calorimeters [9, 24]. Finally, a muon spectrometer relies upon three superconducting air-core toroidal magnets that provide vast magnetic fields to deflect negatively charged muons for high-precision tracking.

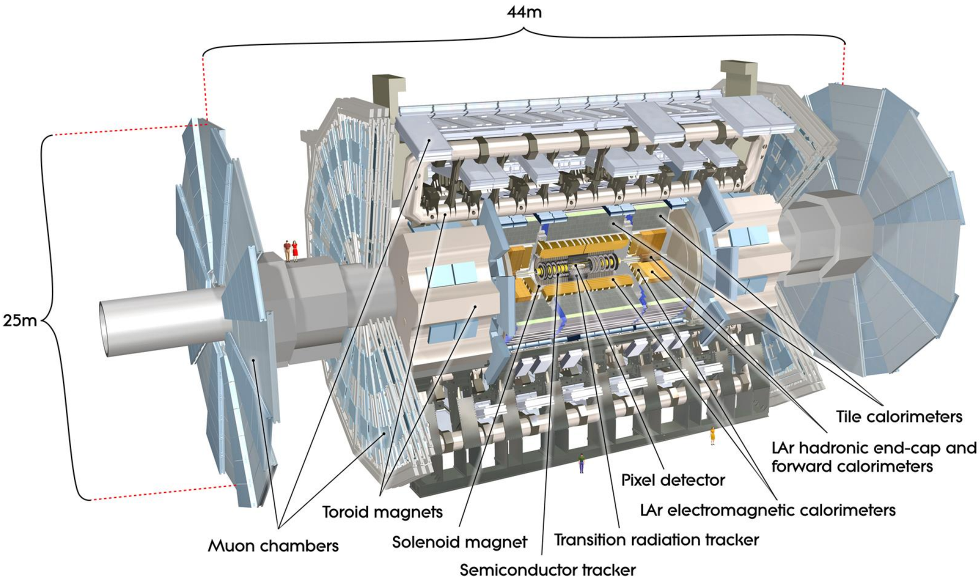

44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

Figure 1.1: Digital cut-away model of the ATLAS detector [9]. The different components required for $H \rightarrow \gamma\gamma$-jet detection, their location and scale in the detector can be seen.

When in operation, the ATLAS detector records collision events at a rate of 200 Hz [9] and records many different variables for each event. These data must then be analysed and the collision events categorised. This stage requires dedicated jet tagging algorithms for the particle decay being studied.

1.2 Photon Identification

Naturally, the ATLAS detector component that is primarily involved in photon reconstruction and identification is the electromagnetic calorimeter [9]. As photons interact with the material of the calorimeter, an electromagnetic shower is created and deposits its energy in a small group of neighbouring calorimeter cells [26]. The signatures of photons and electrons are similar, so reconstruction for both these particles occurs in parallel. Due to the similarity between photons and electrons, discriminating between the diphoton signal class and the electron background classes is important to any jet tagging algorithm focussing on a photon decay product signal.

The standard ATLAS reconstruction of photons is done by first searching for seed clusters in the calorimeter cells [26]. How these clusters are created has been a significant source of research and development, resulting in the widely used superclusters [27]. These are dynamic and variably sized seed clusters and have improved the ability to recover energy carried by photons generated by bremsstrahlung, or electron or photon conversions [27]. There are two types of detectable photons: the converted and the unconverted. Converted photons are those that interact with a charged particle, such as an atomic nucleus, and convert into an electron-positron pair [28]. Unconverted photons are those that do not undergo this process. Converted photons can be tracked, and their superclusters are matched to a conversion vertex (or vertices) [27]. Unconverted photons are superclusters that are not matched to a conversion vertex nor any track information [26, 27].

From this information, the ATLAS standard method for photon identification uses one-dimensional selection criteria that are based on electromagnetic shower shape feature variables [27]. This cut-based selection approach is designed to select photons efficiently and reject backgrounds [27]. Primary identification is done at a *Tight* operating point, with less restrictive operating points (*Medium* and *Loose*) used as triggers for the ATLAS detector [27]. *Tight* photon identification is performed separately for converted and unconverted photons as their associated showers are topologically distinct. This is a result of the conversion electron-positron pair, which are electrically charged, resulting in deflection caused by the powerful magnetic field within the ATLAS detector [27].

The ATLAS standard method for photon identification performs well over a large kinematic range; however, for highly boosted $H \rightarrow \gamma\gamma$ decays, the performance significantly degrades. This is due to the aforementioned Lorentz boosting of heavy boson decay products [29], causing the angular distance between decay products to be reduced. The angular distance of two decay products is inversely proportional to the transverse momentum[3] [23]. Therefore, greater momentum due to larger velocities leads to smaller angular distances.

To develop identification algorithms at the top end of the kinematic range is extremely important, especially as Run 3 is underway at the LHC at the higher centre-of-

[3] The angular distance ΔR of a two-body decay can be approximated by $\Delta R \approx \frac{2m}{p_T}$ where m is the parent particle mass and p_T is its transverse momentum.

mass energy of $\sqrt{s} = 13.6$ TeV [30]. Therefore, a larger proportion of highly boosted heavy bosons will be produced for detection in the ATLAS detector in the near term. A class of identification algorithms using deep neural networks has recently been effective for other highly boosted heavy bosons.

1.3 Deep Neural Networks

Deep neural networks (DNNs) are a broad class of algorithms capable of classification and regression depending on the architecture and downstream task. DNNs are characterised by possessing input and output layers with several so-called hidden layers between them [31]. The hidden layers consist of hidden nodes, performing linear and non-linear transformations on the input data and mapping it to the output [31]. An optimisation algorithm, which since the 1980s has been some form of stochastic gradient descent [32], iteratively adjusts the weights and biases of the hidden nodes [31]. This is done during the backpropagation step to minimise an objective (loss) function that is carefully chosen for the task [32, 31]. These networks automatically and adaptively learn hierarchies in the feature variables, enabling the network to capture complex representations of the input data [31].

DNN jet tagging algorithms have recently been utilised for particle identification using the ATLAS detector. Successfully developed DNN algorithms include those for identification and reconstruction of hadronic boson decays [33], decays into boosted di-τ systems [34] and for highly boosted $Z \rightarrow e^+e^-$ decays [23].

As mentioned in Section 1.2, electron and photon identification are closely related due to the similarity in their signatures in the electromagnetic calorimeter [26]. Therefore, the approach to identifying $Z \rightarrow e^+e^-$ decays is extremely relevant for this thesis. Similar techniques are used to identify both of these highly boosted decays.

1.3.1 Mass Decorrelation with Adversarial Neural Networks

The power of DNNs when presented with complex input data is that they are highly able and flexible when learning how to map the inputs to outputs. However, in the case of jet tagging, highly useful topological jet features are strongly correlated to the jet mass [35]. This can result in the DNN over-relying on the mass feature, underutilising the jet substructure information and increasing systematic uncertainty as signal efficiency degrades at less frequent signal masses [36]. It leads to sculpting of the backgrounds dependent on the signal mass distribution. As the relationship between the mass feature and other topological features is complex and extremely difficult to characterise.

One successful approach has been to add a second DNN to act as an adversary to train the jet tagger to be invariant to the mass [36]. The jet tagger learns to classify the jets in the input data; the adversary network then takes the output from the jet tagger and attempts to predict the binned mass value [36]. The adversary is optimised on its ability to make this mass prediction, whilst the jet tagger is optimised on the combination of its classification loss function and the adversary loss [36]. This training scenario is termed an adversarial neural network (ANN).

Incorporating mass decorrelation into the training strategy of a jet tagger is an effective method for reducing the sensitivity to the systematic uncertainties derived from the dominance of the mass feature [36]. Whilst achieving the primary goal of high jet tagging performance [36].

1.4 Research Objectives

Given the complexity of the task and its importance to high-energy particle physics research, looking beyond the Standard Model, this thesis undertakes two primary objectives to advance the jet tagging of $H \rightarrow \gamma\gamma$ decays.

1. Improve highly boosted $H \rightarrow \gamma\gamma$ decay identifications using deep neural networks.
2. Develop a deep neural network that makes its $H \rightarrow \gamma\gamma$ identifications decorrelated from the mass feature of the input data. While maintaining a high degree of discriminatory power.

Achieving these objectives could substantially contribute to both theoretical and experimental particle physics by enhancing jet tagging capabilities and minimising biases.

The paper is organised as follows: Chapter 2 elaborates on the methodology, providing information on the simulated dataset, exploratory data analysis, feature analysis, network architectures, and evaluation methods. In Chapter 3, detailed results of the models' performance are reported. Chapter 4 discusses these results and possible routes for future research. Finally, Chapter 5 offers conclusive remarks on this thesis' results, implications and potential contributions to the field.

Chapter 2

Methodology

This chapter presents the methodology for developing and evaluating the jet tagging algorithms. Initially, the data utilised for the study is discussed, beginning with its acquisition. Justification for the before-experimentation selection of features within the dataset is also provided. After this, an examination of exploratory data analysis is conducted, scrutinising the attributes of the data with respect to the target classes. Feature analysis follows, elaborating on the significance of various features in influencing the performance of the final models. The mass feature is examined in detail due to one of the stated aims of developing a mass-decorrelated jet tagger. Finally, the architectures of the neural networks are presented along with their underlying rationale, and the evaluation methods applied to assess the jet tagging algorithms are detailed.

2.1 Data

Acquiring high-quantity and quality data is fundamental in developing classification algorithms employing machine learning techniques like deep neural networks. When trained on robust datasets, such neural networks exhibit high flexibility and expressiveness, enabling them to capture complex and nonlinear relationships between input features and output targets [37]. For the purposes of this study, data generated through Monte Carlo event simulation is utilised. This is a methodology commonly employed in the domain of high-energy particle physics [38, 39, 40, 41, 42, 43].

2.1.1 Monte Carlo Event Generation

Monte Carlo event generation is a broad computational technique employed to simulate complex systems and processes by generating random samples from known distributions [44]. Instead of attempting deterministic calculations following classical equations, probability distributions are defined, and samples from these distributions are randomly sampled. This method can be applied to deterministic but highly complex scenarios where analytic solutions are intractable. Naturally, it can also be applied to the simulation of quantum particles and their interactions, an inherently stochastic process [44].

In this thesis, simulated data from three MC event generators, `SHERPA` [40], `MadGraph` [41], and `PYTHIA8.3` [45] are employed for the training of jet tagging algorithms. The feature variables in the simulated data are those measurable and capable of being recorded by the ATLAS detector at the LHC [9]. Information for the generation of specific particle collision events, namely the signal $H \to \gamma\gamma$ decays and background noise $Z \to e^+e^-$ decays, q/g-jets, e/γ-jets, and $(\tau)\tau$-jets, are included. These datasets are thus utilised to develop jet tagging algorithms compatible with actual data from the ATLAS detector.

It should be noted that the simulated dataset is divided into three segments. Comprising 60% of the total number of events is the training dataset, while validation and testing segments each constitute 20% of the data. In the training dataset there are around 280,000 $H \to \gamma\gamma$ decays, 2.3 million $Z \to e^+e^-$ decays, 3.7 million q/g-jets, 1.6 million e/γ-jets and 700,000 $(\tau)\tau$-jets. All analytical procedures and model training are conducted on the training dataset. Performance checks during the training phase are carried out using the validation dataset. Lastly, the testing dataset is reserved exclusively for the final evaluation of the models.

2.1.2 Features

Previous work on other decay processes uses the tracking information provided by the ATLAS detector's inner tracking detector [9, 23]. As photons have zero electric charge, they undergo no deflection due to the solenoid's magnetic field [9], calorimeter information is more relevant. Due to the reduced utility in the tracking features for the signal process $H \to \gamma\gamma$, the training data contains significantly fewer features than previous work on different decay processes [23]. The feature variables that data has been MC generated for are provided in Table 2.1.

Table 2.1: Feature variable information that has been simulated using MC event generation with `SHERPA`, `MadGraph` and `PYTHIA8.3`. These features are used as input variables for the jet tagging algorithms.

Feature	**Description**
p_T Bin	Binned transverse momentum of the jet. This momentum component is perpendicular to the direction of the particle beam in the collider.
$\|\eta\|$ Bin	Binned absolute pseudorapidity of the jet. Pseudorapidity is a spatial coordinate that describes the angle of a particle relative to the beam. $\eta = -\ln\tan(\theta/2)$ where θ is the polar angle with respect to the anticlockwise beam direction [46].
m	Mass of the jet.
$\Delta R(c_1, j)$	Angular distance of the leading cluster to the jet axis.
$\Delta R(c_2, j)$	Angular distance of the sub-leading cluster to the jet axis.
$\Delta R(c_1, c_2)$	Angular separation between the two leading clusters.
$r_{N=1}^{(\beta=1)}$	Ratio of the energy correlation functions calculated from the two leading tracks [23, 47]. This is sensitive to the jet's N-prong substructure [47].
$\max\left(\frac{E_{layer}}{E_{jet}}\right)$	Maximum fraction of energy deposited by a jet in a single layer of the EM calorimeter [48].
f_{EM}	Fraction of energy deposited in the EM calorimeter [48].
E/p	Ratio of the cluster energy to the track momentum [23].
Planar Flow	Spread of the jet's energy over the area of the jet calculated from the two leading tracks. Values around 1 indicate an even spread, and values around 0 indicate a linear spread [23].
Width	Width of the jet, defined as the p_T weighted average of the ΔR distances between the neutral particle flow objects and jet axis direction [23].
Balance	The ratio of the difference between cluster energies and the total energy of the clusters $\frac{E_{cluster}^{1st} - E_{cluster}^{2nd}}{E_{cluster}^{1st} + E_{cluster}^{2nd}}$.
$N^{Const.}$	Multiplicity of particle-flow objects per jet.
N^{Trk}	Multiplicity of tracks per jet.

2.2 Exploratory Data Analysis

In this section, the features are initially visualised in a couple of ways to investigate the inherent distribution of the data in the features with respect to the target classes.

As there are continuous and categorical feature variables, the analyses are separated to present the most relevant plots and information to the feature type. Figure 2.1 contains four bar plots, one for each categorical variable. Each bar plot within the figure illustrates the frequency distribution of each feature across the different types of particle collision events. In absolute terms, the signal process is less frequent than the background processes.

First, consider the top row plots in Figure 2.1; these show the frequency distributions across decay types of $|\eta|$ bin and p_T bin features. In the $|\eta|$ bin plot, all classes are most numerous in the first two bins, but overall, the distribution is quite flat. The p_T bin bar plot shows that the distribution of all the background jets is relatively flat, but the distribution of the signal jets increases as the bin index increases. Pseudorapidity and transverse momentum are highly dependent on the energy of the particle beam in the LHC. The classifiers should be able to function within a range of beam energies. This will require the decisions of the classifier models to be decorrelated with respect to these features. How this will be implemented will be described in section 2.4 when the network architectures are presented.

Two similar distributions can be observed in the bar plots for the N^{Trks} and $N^{Const.}$. The signal jets are only significant in relatively low bin indices compared to the total number of bins. The background jets all have significantly greater spreads across the bins. This suggests that these will be useful features to the jet taggers for distinguishing the signal from the background jets.

The ten continuous feature variables are visualised against the target classes using violin plots in Figure 2.2. These plots allow the distribution and the probability density of the data for each feature to be displayed simultaneously. It should be noted that the most significant outliers in all plots comprise the default values for the feature variable; these are recorded when a measurement for this feature cannot be made. The default values for individual feature variables that possess them are contained in Appendix A.

The violin plots for the signal jets are similar to the $Z \rightarrow e^+e^-$ background decays. Across all features, the signal process and $Z \rightarrow e^+e^-$ decays have the most qualitatively similar distributions and probability densities. It should be noted, however, that the signal process has some slightly more complex probability densities for the width and the $\Delta R(c_1, c_2)$ features. The implication is that the classifiers will find discriminating the signal jet from the Z-boson background jet the most difficult.

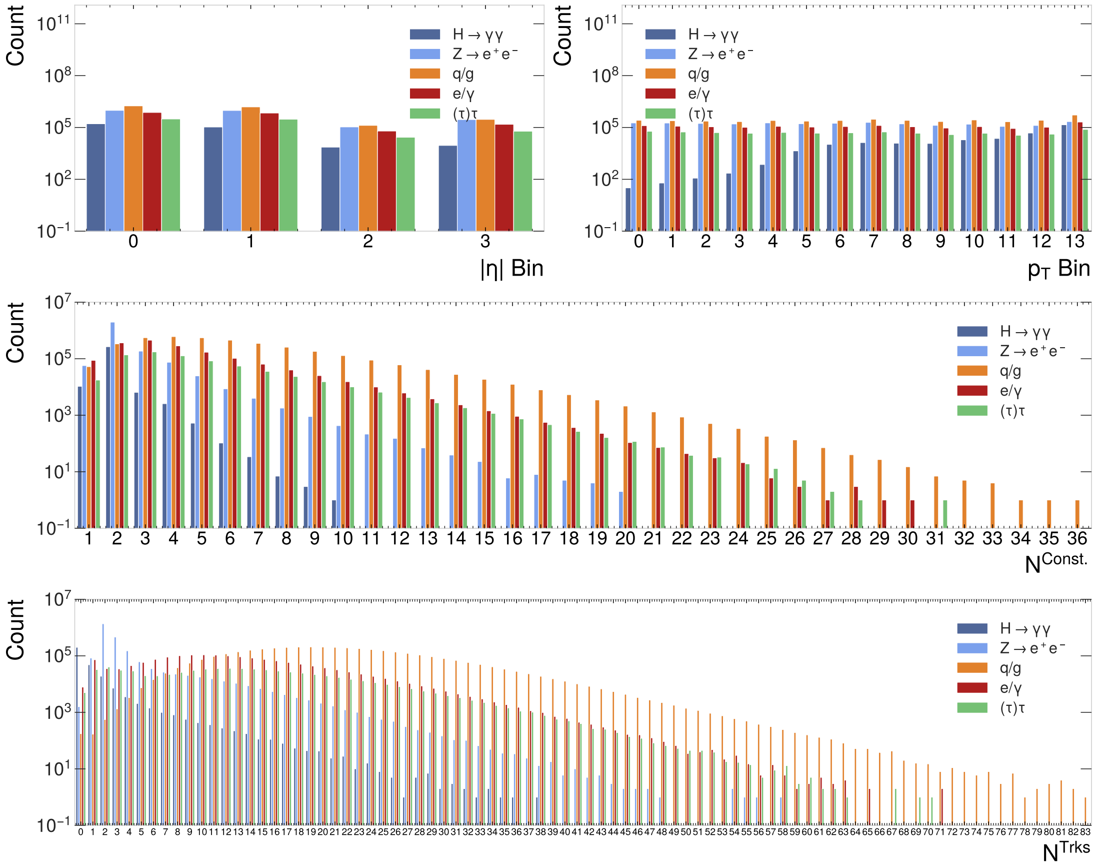


Figure 2.1: Bar plots for the categorical feature variables. Distributions of all jets are relatively flat across all pseudorapidity bins. Distributions for background jets are similar across transverse momentum bins, although the signal count increases with bin index. Distributions are similar for N^{Trks} and $N^{Const.}$, with signal counts at lower bin indices and backgrounds spread across more bins. Enlarged bar plots for N^{Trks} and $N^{Const.}$ can be found in Appendix B.

Additionally, the other background jets possess even more similar violin plots to one another across all features. The jet taggers, therefore, may find discriminating background jets from one another an even harder task. This is not a significant issue, as correctly identifying the signal jet and reducing background jets is the main aim of this thesis.

From Figure 2.2, the continuous feature variables that will likely be of greatest utility to the deep neural networks as they are learning to discriminate between the different decay processes are $r_{N=1}^{(\beta=1)}$, E/p, f_{EM}, $\max\left(E_{layer}/E_{jet}\right)$ and mass. This is due to the median, interquartile range and probability density of the signal class in these features being the most different from the background classes. The second jet tagger aims to identify the signal jet decorrelated to the mass feature. As the mass feature clearly contains highly useful information to the network, developing a mass-decorrelated network requires specific architecture decisions; these will be further discussed in section 2.4. However, seeing the information in the mass violin plot indicates this task's difficulty.

Finally, for the continuous features, it is noted that the features shown with the least utility displayed within the violin plots are the features where the distributions and probability densities are most similar across all jets. This is seen most evidently in the violin plot for the balance; to a lesser degree, it is also observed for the planar flow. These features likely contain limited useful information for the classifiers.

2.3 Feature Analysis

This section analyses the features via correlation coefficients and the mutual information metric. Furthermore, the relationship between various feature variables and the mass feature is examined in more detail, given the objective of decorrelating the second model's classification from this particular feature.

2.3.1 Correlation Coefficients

A succinct method for examining correlations between features is to use a correlation coefficient matrix as seen in Figure 2.3. This figure displays a matrix of the features and their corresponding Spearman rank correlation coefficient. This nonparametric measure enables the calculation of correlation for general monotonic relationships and is deemed suitable for scenarios where nonlinear correlations are anticipated [49]. The range of this correlation coefficient is [-1,1]; positive values are interpreted as positively correlated, and negative values as negatively correlated, 0 indicates no correlation [49].

Notably, the binned absolute pseudorapidity ($|\eta|$ bin) and the binned transverse momentum (p_T bin) of the jet have relatively small correlation coefficients with all other features, especially $|\eta|$ bin. This suggests that these features contain significant information not present in the other features. As expected, the mass is the largest positive correlation for the p_T bin feature.

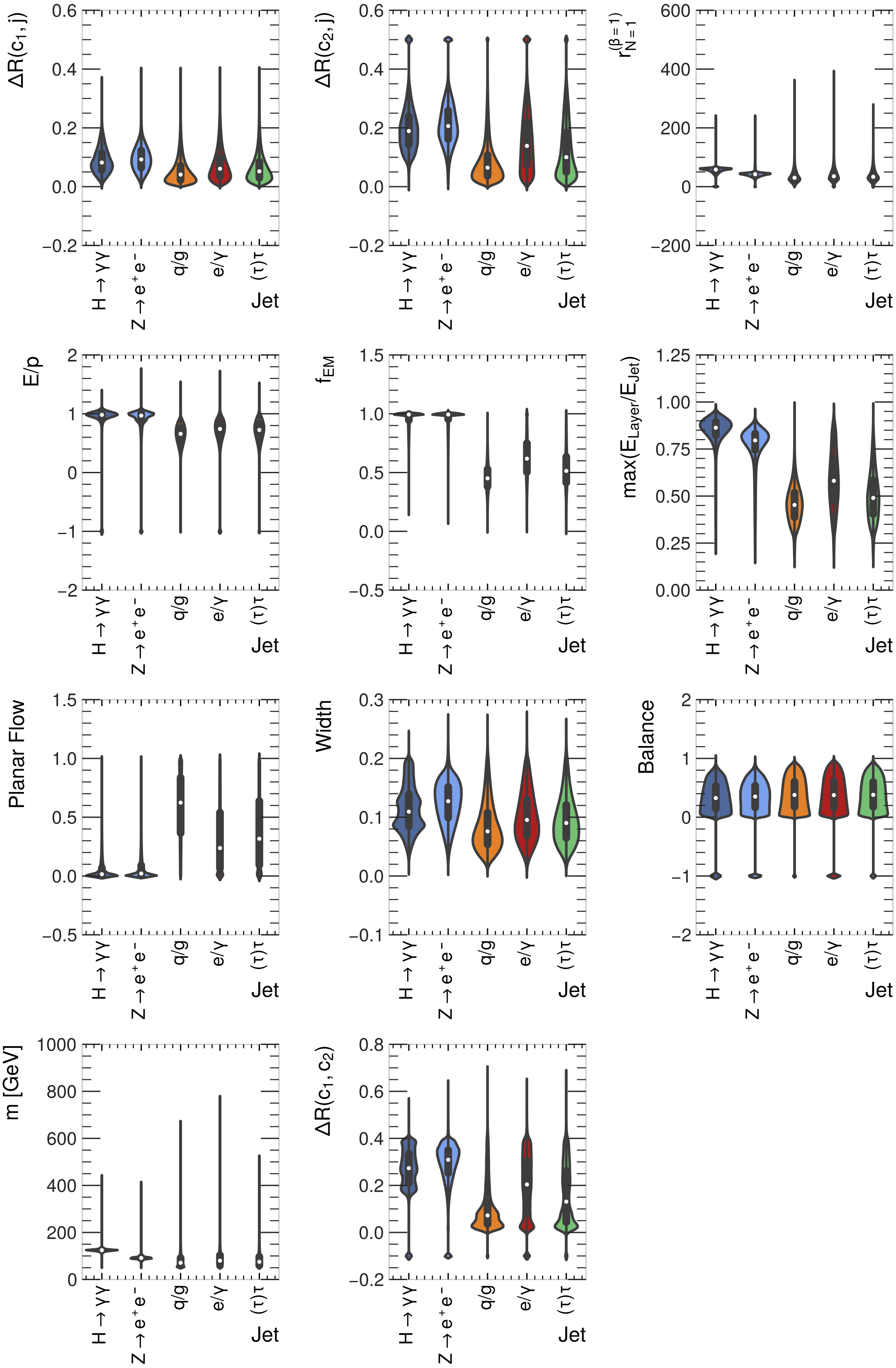


Figure 2.2: Violin plots for the continuous feature variables. These plots combine box plots and histograms by presenting the median, interquartile range and distribution shape.

More generally, many features correlate positively or negatively to many other features in the data. The relationship between the features and the target classes is complex and motivates the use of a type of classifier that is highly expressive.

2.3.1.1 Mass Feature Correlations

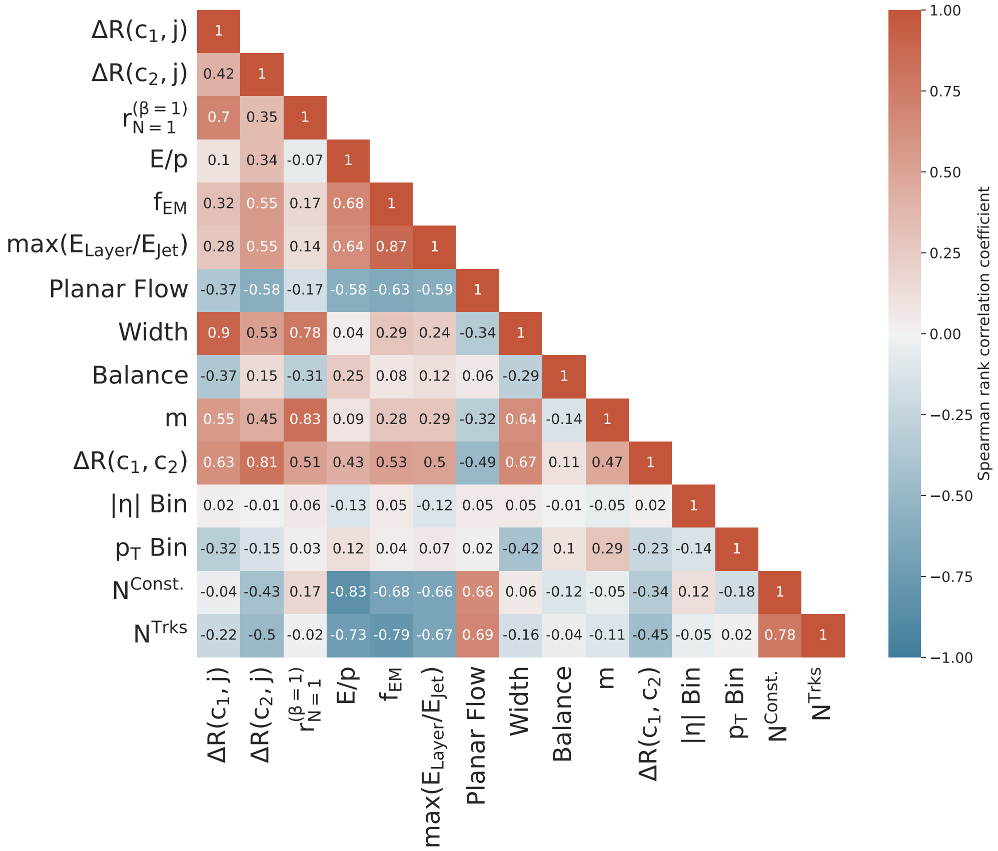
Decorrelating the identification made by a classifying model from the mass feature variable can be attempted in several ways. However, the decision of which method to use is related to how many relationships this feature has with other features. The more correlated the mass feature is to other features, the more complex the relationship, requiring more complicated network architecture and training decisions.

As seen in Figure 2.3, the mass feature (m) has a Spearman rank correlation coefficient with an absolute value greater than 0.25 with 9 of the other features. It has a value greater than 0.5 for 3 of the features. As expected, the features that mass is most strongly correlated to are topological features of the jet $\Delta R(c_1, j)$, $\Delta R(c_2, j)$, $\Delta R(c_1, c_2)$, $r_{N=1}^{(\beta=1)}$ and width. Making jet identifications from the associated but distinct jet topology will allow the mass-decorrelated classifier to perform at a greater range of energies.

The relationship of the mass with other features is clearly complex and, therefore, requires a high-capacity network architecture to achieve mass invariance and a high level of signal accuracy. This contrasts the binned pseudorapidity and transverse momentum features, which are not strongly correlated to many, if any, in the case of pseudorapidity.

2.3.2 Mutual Information

In the previous section, the correlation between features was determined; in this section, another metric to rank the features by a single score is employed. Such a method for measuring the relative importance of different features is mutual information. This metric determines relevance and redundancy between features and the target class [50]. Mutual information is a nonparametric measure that is useful in nonlinear settings [51], such as this complex ATLAS detector data scenario. The metric has a range $[0, \infty)$ and is only 0 if and only if the two random variables in question are independent [51]. The equations for both random discrete and continuous variables are presented in equations 2.1 and 2.2 respectively [51]. Where $p(x,y)$ is the joint probability density function (pdf) and, $p(x)$ and $p(y)$ are the marginal pdfs of random variables X and Y.



	$\Delta R(c_1, j)$	$\Delta R(c_2, j)$	$r_{N=1}^{(\beta=1)}$	E/p	f_{EM}	$\max(E_{Layer}/E_{Jet})$	Planar Flow	Width	Balance	m	$\Delta R(c_1, c_2)$	$\lvert\eta\rvert$ Bin	p_T Bin	$N^{Const.}$	N^{Trks}
$\Delta R(c_1, j)$	1														
$\Delta R(c_2, j)$	0.42	1													
$r_{N=1}^{(\beta=1)}$	0.7	0.35	1												
E/p	0.1	0.34	-0.07	1											
f_{EM}	0.32	0.55	0.17	0.68	1										
$\max(E_{Layer}/E_{Jet})$	0.28	0.55	0.14	0.64	0.87	1									
Planar Flow	-0.37	-0.58	-0.17	-0.58	-0.63	-0.59	1								
Width	0.9	0.53	0.78	0.04	0.29	0.24	-0.34	1							
Balance	-0.37	0.15	-0.31	0.25	0.08	0.12	0.06	-0.29	1						
m	0.55	0.45	0.83	0.09	0.28	0.29	-0.32	0.64	-0.14	1					
$\Delta R(c_1, c_2)$	0.63	0.81	0.51	0.43	0.53	0.5	-0.49	0.67	0.11	0.47	1				
$\lvert\eta\rvert$ Bin	0.02	-0.01	0.06	-0.13	0.05	-0.12	0.05	0.05	-0.01	-0.05	0.02	1			
p_T Bin	-0.32	-0.15	0.03	0.12	0.04	0.07	0.02	-0.42	0.1	0.29	-0.23	-0.14	1		
$N^{Const.}$	-0.04	-0.43	0.17	-0.83	-0.68	-0.66	0.66	0.06	-0.12	-0.05	-0.34	0.12	-0.18	1	
N^{Trks}	-0.22	-0.5	-0.02	-0.73	-0.79	-0.67	0.69	-0.16	-0.04	-0.11	-0.45	-0.05	0.02	0.78	1

Figure 2.3: A correlation coefficient matrix using the Spearman rank correlation coefficient between features.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{2.1}$$

$$I(X;Y) = \int_X \int_Y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \, dx\, dy \tag{2.2}$$

Overall, the MI values in Table 2.2 are somewhat low. This suggests that the features are not highly predictive in isolation. Figure 2.3 certainly suggests that the features possess many relations to one another, suggesting that there are likely complex relationships amongst the features and then with the target variable. Another possible interpretation is there is an inherent complexity to the target variable. The target variable is the decay process caused by strongly boosted particles in the LHC; this is subject to a range of physical processes, meaning the target variable is inherently complex.

If there were significantly more features, the MI score could be used as a feature selection method to remove the least relevant features. As previously stated, no features will be pruned due to the relatively low number. This, combined with the analysis

Table 2.2: Mutual information between the input feature variables and the target classification variables. This allows an interpretation of the importance of the features to the type of jet found in the training dataset.

Feature	MI Score	Feature	MI Score
N^{Trks}	0.63	$\Delta R(c_2, j)$	0.24
f_{EM}	0.54	$r_{N=1}^{(\beta=1)}$	0.21
$\max \frac{E_{layer}}{E_{jet}}$	0.43	$\|\eta\|$ Bin	0.12
$N^{Const.}$	0.39	$\Delta R(c_1, j)$	0.11
E/p	0.34	Width	0.10
m	0.34	p_T Bin	0.06
Planar Flow	0.30	Balance	0.02
$\Delta R(c_1, c_2)$	0.29		

suggesting that there are complex relationships between feature variables and the target variable, means that it is prudent not to prune these features from the set of input variables for the neural networks.

The next section will outline the network architectures and the training procedure under which the classifiers will learn to tag the highly boosted $H \to \gamma\gamma$ decays.

2.4 Network Architectures and Training Procedure

This thesis aims to produce a dedicated jet-tagging algorithm for the strongly boosted $H \to \gamma\gamma$ decay process as measured and recorded in the ATLAS detector in the LHC. In section 2.4.1, the methodology for developing this algorithm is provided, along with its architecture and training procedure. Section 2.4.2 contains the same information for developing a mass-decorrelated jet-tagging algorithm and highlighting the key differences between the two classifiers.

It should be noted that the q/g-jets have also been removed from the data for both classifiers. This provides the jet-tagging algorithms with a classification task between four jets as performed in previous related work [23]. This will make comparison to previous work more relevant and meaningful. These jets are selected for removal due to their relative lack of importance for confusion with the signal jet and the fact that it is the most numerous class; removing q/g-jets aids in addressing the class imbalance. Five-class jet taggers are developed and presented in Appendix E.

All networks are constructed using the PyTORCH [52] package and are defined using the associated framework. All training uses an Nvidia A100 GPU with 20Gb of memory and 32 CPU cores [53].

2.4.1 DNN

The architecture for this network is based on a recent work dedicated to another heavy boson decay at the same collision energy in the LHC using the ATLAS detector. This jet-tagging algorithm was developed for $Z \rightarrow e^+e^-$ decays [23]. This decay has been identified in the exploratory data analysis as the background class most likely to be confused for the signal jet $H \rightarrow \gamma\gamma$. Therefore, the classifying algorithm for the boosted Z-boson decay is a good architecture to base the H-boson decay classifier, as it has been proven to perform well for a similar decay process. This architecture is a fully-connected feedforward deep neural network [23, 54].

The network uses all features as input variables; thus, it has 15 input variables. These data pass through the hidden layers that are composed of nodes with four outputs corresponding to the four target classes, the signal $H \rightarrow \gamma\gamma$ decays and the three backgrounds, $Z \rightarrow e^+e^-$ decays, e/γ and $(\tau)\tau$-jets. These output nodes are the calculated probabilities for the jet in question to be attributed to one of the classes. Each hidden layer has an activation function; the two functions that are tested are the Sigmoid activation function [55] and the rectified linear unit (ReLU) [56]. The sigmoid activation function is the best-performing function in one of the previous works [23] and ReLU in another relevant work [36]. Thus, these activation functions will be trialled for the DNN jet tagger. The weights in the network are initialised with the widely used Glorot initialisation technique [57]. This initialisation method draws the initial weights from a zero-mean Gaussian distribution [57]. It has particular utility when used with symmetric activation functions around zero, such as the sigmoid activation function applied to each layer. Batch normalisation is also applied to each hidden layer [58]. This normalises layer inputs, reducing internal covariate shift. This makes the networks less sensitive to the specificities of random weight initialisation and accelerates the learning rate as convergence becomes easier [58].

For training the multiclass classification loss function, cross-entropy loss is used [59]. This function is weighted by the proportions of each jet to address the class imbalance within these data. The loss function is additionally weighted so that the p_T and $|\eta|$ bins have a flat distribution. The method for stochastic optimisation used is the popular ADAM optimizer [60].

To mitigate the risk of overfitting, the dropout method for regularisation is applied to the hidden nodes [61]. During training, a fraction of the nodes are randomly switched off so that the network becomes less sensitive to the specific weight of the nodes [61]. At inference time, the output of each node is scaled down by the dropout rate to account for the missing activations during training [61]. Additionally, early stopping is employed to reduce overfitting [62]; training is stopped after five epochs with the loss decreasing by less than 10^{-4}.

To find a set of optimal hyperparameters, ranges are defined for each hyperparameter, and then a grid search is performed over all combinations of the hyperparameters [63]. The optimisation range for each hyperparameter is seen in Table 2.3, where possible, the computational expense is reduced by narrowing ranges by utilising found values in the related work [23]. The top five models are found based on their cross-entropy loss on the validation dataset and are compared to one another using the evaluation methods presented in section 2.5 to find the best-performing DNN jet tagger.

Table 2.3: Hyperparameter optimisation table for the DNN classifier. In total, 768 hyperparameter combinations are used for trained and evaluated models.

Hyperparameter	**Optimisation Range**
Training epochs	[30, 50]
Optimiser	Adam
Dropout probability	[0.01, 0.1]
Learning rate	[0.001, 0.01]
Hidden layers	[3,4]
Hidden nodes	[100, 200]
Activation function	[Sigmoid, ReLU]

2.4.2 Mass-Decorrelated Adversarial Network

The second jet-tagging algorithm is developed in much the same way as the DNN classifier. The fundamental difference comes from having a classifying network and an adversary network. The classifying network architecture is essentially the same as the DNN classifier; the classifier is presented with the input feature data and then attempts to predict the class the jet belongs to. The softmax activation function [64] is then applied to the outputs; these outputs are passed through the adversary network, which attempts to predict the mass. The loss function is then a linear combination of the loss functions for both the classifier and adversary networks [36], as seen in equation 2.3. The coefficient λ is a positive constant and is a hyperparameter that must be determined in the hyperparameter optimisation. The λ value found in the previous work is $\lambda = 100$ [36]; this conveniently narrows the range required for hyperparameter optimisation. This type of neural network is termed an adversarial neural network (ANN) training setup.

$$L_{tagger} = L_{classification} - \lambda L_{adversary}, \quad \lambda \in \mathrm{R}^{+} \tag{2.3}$$

The complex relationships explored in the feature analysis in section 2.3 motivate using a flexible neural network that can learn these relationships implicitly between the class and mass and mass-related features. This contrasts the transverse momentum and pseudorapidity features that do not correlate with other features to anywhere near the same degree. Producing jet taggers decorrelated to these features can be managed by weighting the loss function to have a flat distribution with respect to both these features.

The network architecture and training implemented extends previous work for mass-decorrelated jet taggers [36]. Following this approach, both networks use the cross-entropy loss function [36]. The mass feature is transformed from a continuous feature variable to a categorical one. Ten mass bins are created, each with an equal number of samples. The classifier loss function is again weighted to have flat distributions with respect to the transverse momentum and pseudorapidity bins. The adversary loss is weighted so that signal jets have weight zero, thus becoming invisible to the adversary network [36]. This leads to a more challenging situation for the classifier network as it attempts to outperform the adversary without directly relying on the mass feature.

Nearly all hyperparameters are identical to the ones used in the DNN classifier, except there are now two networks with hyperparameters to determine. However, there is an additional hyperparameter, the aforementioned λ. It should also be noted that the adversary network is less complex than the classifier, as in the related work [36]. This allows the classifier to outperform the adversary as the training progresses so that the classifier is what is developed and not a highly competent mass predictor [36]. The hyperparameter space is seen in Table 2.4.

The adversary is also pre-trained [36] for five epochs to learn an initial representation of the data distribution from the training dataset. Additionally, to ensure that all mass bins contain sufficient signal jets for the adversarial setup to function across mass values, the training dataset is oversampled [65]. This functionally increases the size of the training dataset and, therefore, the computational cost. This occurs in tandem with the added cost of an additional network to optimise; the resulting impact is that the resource constraint reduces the number of hyperparameter combinations that can be searched over compared to the DNN architecture.

The top five model hyperparameter configurations are found by calculating the mean and standard deviation of the signal efficiency across the mass bins. A maximum threshold for the validation loss is set at 0.4, and a maximum threshold for the standard deviation of the signal efficiency is set to 0.2. The top five models then satisfy these threshold criteria and have the highest mean signal efficiency across the mass bins.

In the next section, the criteria by which the DNN and adversarial jet taggers are evaluated are presented.

Table 2.4: Hyperparameter optimisation table for the adversarial network classifier. (C) refers to the classifying network that is the $H \to \gamma\gamma$ decays jet tagger and (A) refers to the adversary network. In total, 576 hyperparameter combinations are used for trained and evaluated models.

Hyperparameter	**Optimisation Range**
λ	[0.01, 100]
Training epochs	[30, 50]
Optimiser	Adam
Dropout probability (C)	0.01
Learning rate (C)	[0.001, 0.01]
Hidden layers (C)	4
Hidden nodes (C)	200
Activation function (C)	[Sigmoid, ReLU]
Learning rate (A)	[0.01, 1]
Hidden layers (A)	4
Hidden nodes (A)	[100, 200]
Activation function (A)	[Sigmoid, ReLU]

2.5 Evaluation methods

Various evaluation methods are employed to determine the performance of the jet taggers in both a qualitative and quantitative manner. As mentioned in section 2.4, the cross-entropy loss is calculated on the validation dataset so as to have an online evaluation of the networks during hyperparameter optimisation. The aim is to minimise this loss; the five models in each hyperparameter optimisation with the smallest validation loss values are compared using this section's remaining criteria. The offline evaluation criteria use the held-out test dataset, which is the same size as the validation dataset.

The first evaluation metric is the confusion matrix. This enables a comparison between the model's performance and the actual data values, clearly visualising which classes the model tends to misclassify and what those mistaken classes are. Ideally, the jet taggers are confident and correct in all classes. However, it is signal class accuracy that is primarily desired. In the event of background class confusion, this is not an issue. Additionally, the confusion matrix will provide information about misclassifications, aiding interpretation as to the cause of confusion.

The area-under-curve (AUC) value for the receiver-operating characteristics (ROC) curves will also be calculated and tabulated. The ROC curve represents the true positive rate (TPR) against the false positive rate (FPR) for different classification thresholds. The AUC metric computes the area under the ROC curve. It provides a scalar value that quantifies the overall ability of the model to discriminate between the positive and negative classes. A perfect classifier will have an AUC of 1, while a random classifier will have an AUC of 0.5. These are useful metrics to assess the jet taggers holistically.

The next metric allows the assessment of the jet tagger under different physical assumptions, providing further flexibility in analysis to an end user. This is done by combining the probabilities of signal and background jets to allow further optimisation of the algorithm [23]. Thus, a single discriminant function for evaluating the jet tagger's performance is determined. This is a modification to the discriminant function from $Z \rightarrow e^+e^-$-jet tagger to create a similar discriminant for $H - \gamma\gamma$-jet tagging [23, 66], this is shown in equation 2.4. p_i is the softmax probability generated by the jet tagger for each jet class, and f_i, in this thesis, are the jet fractions in the test data sample [23]. The distributions of this discriminant for the signal and background jets should show a clear separation between the signal and backgrounds.

$$D_{H\gamma\gamma} = \ln\left(\frac{f_{H\rightarrow\gamma\gamma} \cdot p_{H\rightarrow\gamma\gamma}}{f_{Z\rightarrow e^+e^-} \cdot p_{Z\rightarrow e^+e^-} + f_{e/\gamma} \cdot p_{e/\gamma} + f_{(\tau)\tau} \cdot p_{(\tau)\tau}}\right) \tag{2.4}$$

The variable f_i is the relative importance of each jet class, and this parameter can be varied depending on the physical analyses an end user of these jet taggers is performing [23].

Next, the signal efficiency ($\varepsilon_{H\rightarrow\gamma\gamma}$) is computed; this is the fraction of correctly identified $H \rightarrow \gamma\gamma$ decays [23].

$$\varepsilon_{H\rightarrow\gamma\gamma} = \frac{True\ positives}{Positives}, \qquad JRR = \frac{Negatives}{False\ Positives} \tag{2.5}$$

Signal efficiency is calculated as a function of background jet rejection rates (JRR) and jet mass. Plotting JRR against $\varepsilon_{H\rightarrow\gamma\gamma}$ will allow direct comparison with related work at reducing the same background jets. Signal efficiency as a function of jet mass is the metric by which the mass decorrelation in the adversarial jet tagger will be assessed. A mass-decorrelated classifier will possess a relatively flat and stable signal efficiency curve.

Finally, it should be noted that out of the large number of hyperparameter combinations trialled for both the DNN and ANN top five models. These models will overall be close to one another in performance. Thus, another useful criterion shall be applied. For the DNN, this will evaluate the signal efficiency as a function of the transverse momentum and pseudorapidity bins, as flattening this curve has been attempted within the training procedure by weighting the loss function. This metric can also be applied to the ANN, although the more important metric will be evaluating the background jet rejection rate as a function of the mass. The best ANN model shall be the one with the smallest variation in jet rejection rates with respect to the mass.

The optimal jet taggers selected using these evaluation methods will also be interpreted. The mutual information between the test dataset features and the discriminant $D_{H\gamma\gamma}$ will be calculated. This will assess which feature variables contain the most useable information for the jet taggers to learn [67, 23].

In this chapter, the methodology was presented and justified, from the origin of the simulated data to the final analysis of the information utilised by the neural networks. In Chapter 3, the results of the hyperparameter optimisation for both jet taggers are presented using the evaluation methods outlined in this section.

Chapter 3

Results

This chapter presents the results of experimentation with two deep neural network architectures for classifying highly boosted particle data. The results for the DNN are reported first; this functions as a baseline due to its focus on maximising the identification of $H \to \gamma\gamma$ decays.

Secondly, the adversarial neural network classifier results are detailed. This network architecture balances dual aims of maximising jet tagging performance and mitigating systematic biases introduced by a dominant mass feature in the datasets. The degree to which this mitigation is successful is under particular scrutiny, and to what degree, if any, this degrades classification performance. The results are presented using the evaluation methods outlined in Section 2.5.

3.1 DNN

This section presents the results from the tests performed on the best-performing DNN jet tagger. The hyperparameters for this classifier model are contained within Table 3.1. The signal and the backgrounds classification performance was broadly the same across all top five models. Therefore, the main criteria for differentiation of these models was the effectiveness of the transverse momentum and pseudorapidity bin weighting. The best model of these top five models was the model with the smallest variation in signal efficiency for both the transverse momentum and pseudorapidity bins. The best model's signal efficiency curves with respect to these features can be seen in Figure 3.1. The best model was trained again with rounded hyperparameter values; all figures contained in this section are from this final training procedure.

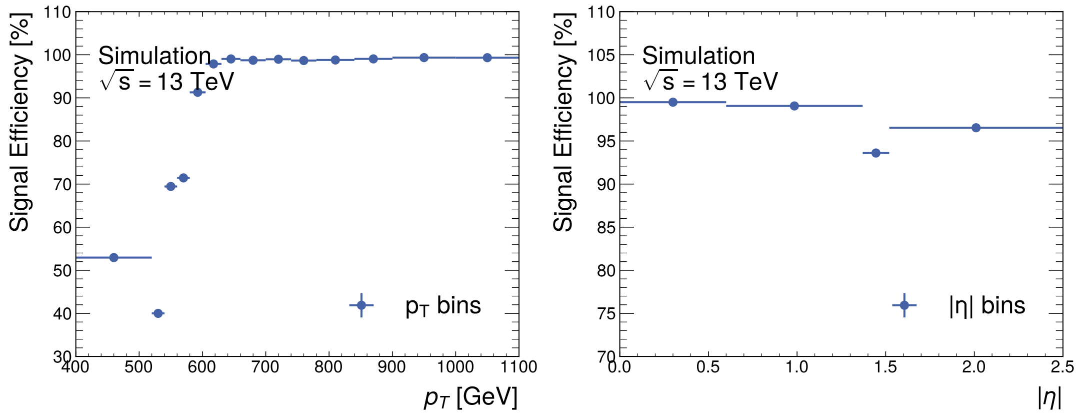


Figure 3.1: Signal efficiency for the best DNN jet tagger with respect to the transverse momentum bins (left) and the pseudorapidity bins (right). At p_T values up to 630 GeV, the jets are not fully efficient in capturing the entire $H \rightarrow \gamma\gamma$ decay. There are cables in the detector between $|\eta| \in [1.37, 1.52]$ which explains the reduced efficiency in the third $|\eta|$ bin.

Table 3.1: Optimal found DNN hyperparameters. The top five models were compared to one another using the evaluation methods in Section 2.5. These are the hyperparameters of the top model found during this evaluation. Dropout probability is rounded to the nearest 2 decimal place value.

Hyperparameter	**Value**
Training epochs	30
Early stopping	27
Optimiser	Adam
Dropout probability	0.03
Learning rate	0.001
Hidden layers	4
Hidden nodes	200
Activation function	ReLU

The training curves for the DNN are displayed in Figure 3.2. The left-hand side plot shows the cross-entropy loss as a function of the training epochs on both the training and validation datasets, whereas the right-hand side plot shows the signal accuracy as a function of the training epochs. The training curves are typical for DNN training, with training performance better than validation performance. The loss curves exhibit the classic shape with little fluctuation. When examining the gap between training and validation loss curves, it should be noted that the difference between the training loss at the first and last epochs is small at 0.025. On the other hand, whilst signal performance shows a trend of increased performance, it fluctuates more significantly. It should be noted that the signal accuracy is always greater than 98.8%; however, this fluctuation motivated the use of the cross-entropy loss as the hyperparameter optimisation criterion.

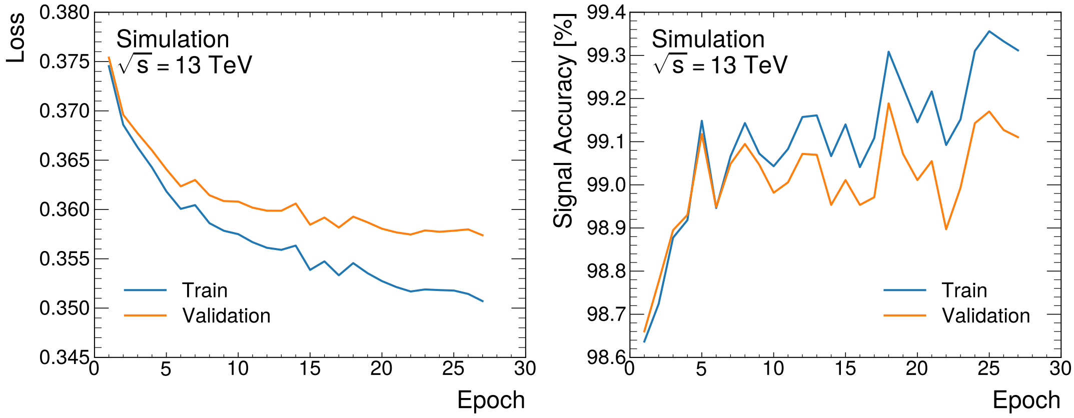


Figure 3.2: Training curves for the best performing DNN jet tagger. The cross-entropy loss is plotted (left) as a function of the training epochs, and the signal accuracy is also presented as a function of the training epochs (right). All curves are computed for both the training and validation datasets.

Of the evaluation methods, per jet classification performance is examined first, using the confusion matrix in Figure 3.3. Each jet in the input data is assigned to its highest softmax probability value produced by the DNN. This is then compared to the true jet classification to produce a percentage value. There is a high degree of signal separation with 99.09% of the signal jets correctly identified. There is little confusion between the signal and background classes; the most confusion occurs with the e/γ-jets with the DNN incorrectly predicting these jets 0.53% of the time when the true class is the signal. False positives are slightly more common, with the signal class being predicted 0.29%, 0.92% and 0.15% for each background $Z \rightarrow e^+e^-$, e/γ and $(\tau)\tau$ respectively.

The confusion between the background jets is significantly greater, especially between e/γ and $(\tau)\tau$-jets. The highest degree of confusion between the DNN outputs occurs when the DNN incorrectly predicts a jet is a $(\tau)\tau$-jet when it is, in fact, a e/γ-jet. This particular instance of confusion occurs at a rate of 33.63%.

Table 3.2 contains the DNN AUC values. The performance of the DNN over all classes is extremely high, with the $H \rightarrow \gamma\gamma$ and $Z \rightarrow e^+e^-$-jets possessing the maximum or near-maximum AUC score. No AUC score is lower than 0.91, indicating high capability by the DNN classifier.

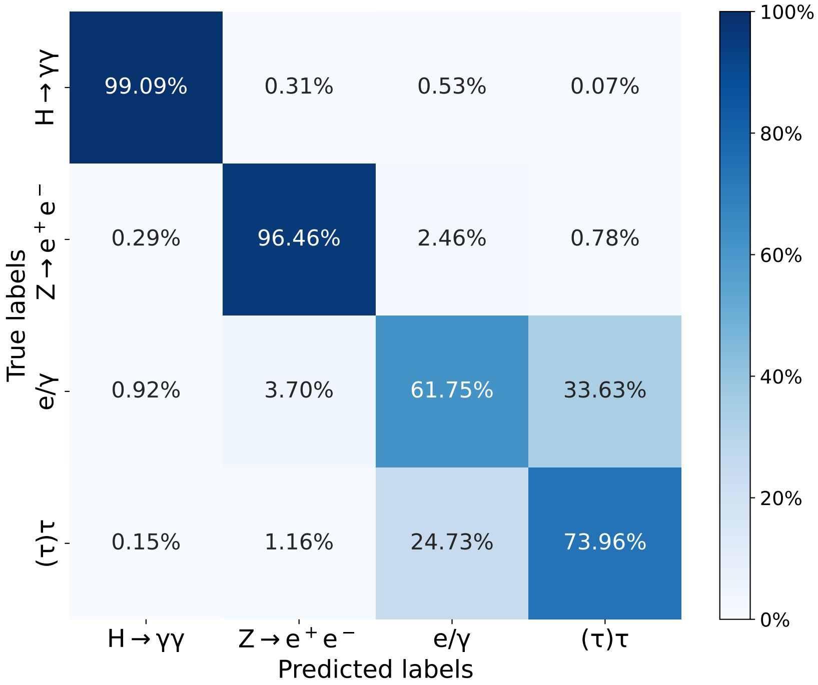


Figure 3.3: Confusion matrix of the $H \rightarrow \gamma\gamma$ decays with the three background classes. There is a high degree of separation between the signal jet and all background jets. The DNN jet tagger is also highly capable of distinguishing the $Z \rightarrow e^+e^-$-jets from the other background jets.

Table 3.2: AUC Scores for different jets using the DNN and ANN jet taggers. Performance is very similar but the DNN performs 0.01 better for e/γ and $(\tau)\tau$-jets.

Jet	DNN AUC Score	ANN AUC Score
$H \rightarrow \gamma\gamma$	1.00	1.00
$Z \rightarrow e^+e^-$	0.99	0.99
e/γ	0.93	0.92
$(\tau)\tau$	0.91	0.90

In Figure 3.4, the distribution of the scalar discriminant $D_{H\gamma\gamma}$ score is presented. This distribution is with respect to the fraction of the jets in the data sample for each jet class. A high degree of separation between the signal and the background jets can be observed. The distributions of the backgrounds are relatively similar to one another with peaks of $D_{H\gamma\gamma} \in (-17,-10)$ whereas the signal $D_{H\gamma\gamma}$ peak is around 5. Additionally, it should be noted that the $D_{H\gamma\gamma}$ score is most similar between the e/γ and $(\tau)\tau$-jets with the peaks around -17. The $D_{H\gamma\gamma}$ is around -12 for the $Z \rightarrow e^+e^-$ decays. This is similar to the other experimental results that indicate the ability of the DNN to discriminate between the backgrounds is weakest between e/γ and $(\tau)\tau$-jets.

The scalar discriminant is also used with mutual information to interpret the importance of different feature variables to the DNN output discriminant. This information is contained in Table 3.5 and provides the features and their mutual information scores with $D_{H\gamma\gamma}$. Importantly, the mass feature has the greatest mutual information score and is thus the most important feature of the DNN jet tagger.

Figure 3.5 displays the jet rejection rate per background jet as a function of the signal efficiency. As is typical for these curves, performance is degraded at higher signal efficiencies as more background jets are accepted as signal jets. The less smooth curves produced for the $Z \rightarrow e^+e^-$ decays and $(\tau)\tau$-jets indicate a lack of MC statistics for these background jets at lower signal efficiencies.

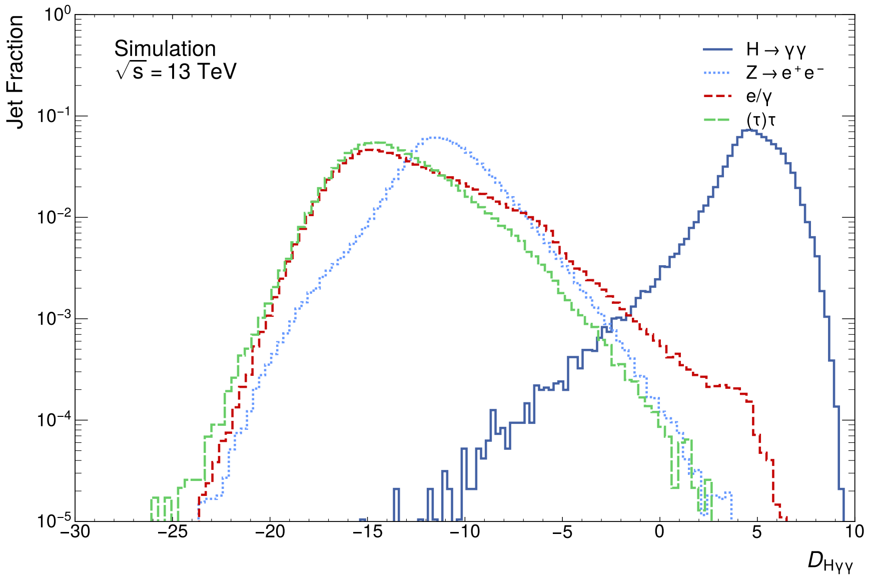


Figure 3.4: Distribution of the scalar discriminant $D_{H\gamma\gamma}$ score for all jet classes Clearly visible is a separation between the signal jet $D_{H\gamma\gamma}$ score distribution and that of the background jets.

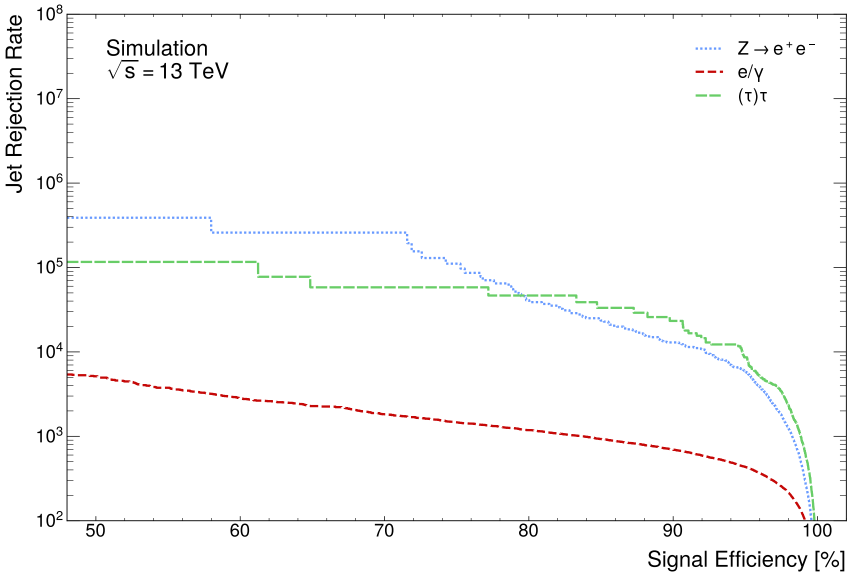


Figure 3.5: Jet rejection rates of the background classes as a function of the signal efficiency. The steps seen for $Z \rightarrow e^+e^-$ and (τ)τ backgrounds indicate a lack of statistics for these classes at lower signal efficiencies.

3.2 ANN

This section presents the results from the ANN jet tagger. The same evaluation figures and tables are used as in the previous section to allow direct comparison to the DNN jet tagger. Additionally, jet rejection rates for the background jets as a function of the mass bins are included. These are created for both DNN and ANN jet taggers to display the effect of adversarial training on mass correlation.

Table 3.3 contains the hyperparameters found during the grid search optimisation. A final ANN classifier was trained using rounded values from the hyperparameter optimisation. Notably, the two learning rates are orders of magnitude different from the learning rate found for the DNN. Additionally, the adversary was found to be a less complex and expressive network than the classifier.

Figure 3.6 contains the confusion matrix of the ANN classifier subnetwork. Performance is broadly similar to the DNN, with slightly more confusion occurring, but this is a relatively minimal degradation in performance.

Figures 3.7 and 3.8 are the scalar discriminant distributions and jet rejection rates, respectively. Both of these figures are highly similar to those produced for the DNN;

however, there is a greater difference in rejection rate at lower signal efficiencies for the ANN jet taggers, but these are small distinctions between the tagging networks. Table 3.2 shows that the DNN possesses mildly better AUC scores for e/γ and $(\tau)\tau$-jets than the DNN, although the DNN and ANN have the same AUC scores for the signal and $Z \rightarrow e^+e^-$-jets.

Table 3.3: Optimal found hyperparameters for the adversarial neural network setup. (C) the classifying network is the $H \rightarrow \gamma\gamma$ jet tagger, and (A) refers to the adversary network.

Hyperparameter	**Value**
λ	0.01
Training epochs	30
Early stopping	16
Optimiser	Adam
Dropout probability (C)	0.01
Learning rate (C)	0.01
Hidden layers (C)	4
Hidden nodes (C)	200
Activation function (C)	ReLU
Learning rate (A)	1
Hidden layers (A)	4
Hidden nodes (A)	100
Activation function (A)	ReLU

In Table 3.5, the mutual information score between the features and the scalar discriminant is presented. In absolute value, the greatest change between this metric for the DNN and ANN is the -0.10 change in the mutual information between the mass and $D_{H\gamma\gamma}$. This corresponds to a percentage change of -27.8%, the second-greatest percentage change decrease after the max (E_{layer}/E_{jet}) feature. Several features undergo much larger percentage change increases; the greatest is the mutual information score for the $|\eta|$ bin with $D_{H\gamma\gamma}$, which undergoes a 175.0% increase.

Figure 3.9 contains two subplots displaying the background jet rejection rates as a function of mass at the $\varepsilon_{H\rightarrow\gamma\gamma} = 95\%$ operating point for both the DNN and ANN jet taggers. Only three mass bins satisfy this criterion: those closest to the signal mass. It can be seen in this figure that the distributions of both $Z \rightarrow e^+e^-$ and e/γ-jets are slightly flatter, and the jet rejection rate for the $(\tau)\tau$-jets are greater for the ANN. Total jet rejection rates derived from figures 3.5 and 3.8 at the same operating point are contained in Table 3.4. The DNN outperforms the ANN in reducing background jets at this operating point in this more general metric.

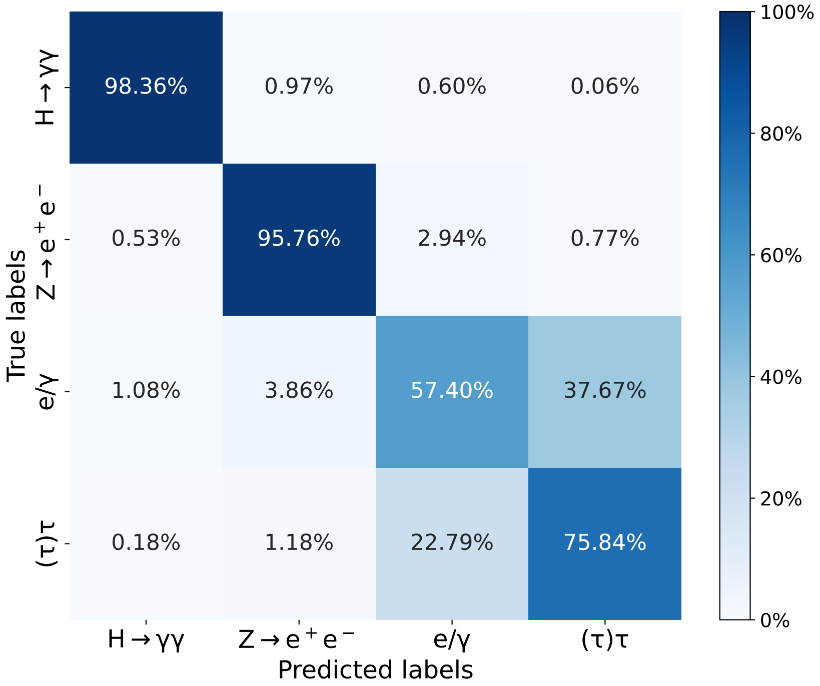


Figure 3.6: Confusion matrix of the $H \rightarrow \gamma\gamma$ decays with the three background classes. There is a high degree of separation between the signal jet and all background jets. The ANN jet tagger performs similarly to the DNN.

Table 3.4: Comparison of rejection rates for the DNN and ANN jet taggers at the $\varepsilon_{H\rightarrow\gamma\gamma} = 95\%$ operating point for different background jets. Jet rejection rates are low relative to the number of jets because boosted $H \rightarrow \gamma\gamma$ decays are so rare.

	Rejection Rates at $\varepsilon_{H\rightarrow\gamma\gamma} = 95\%$	
Background Jet	DNN	ANN
$Z \rightarrow e^+e^-$	7401	1469
e/γ	448	275
$(\tau)\tau$	8955	4312

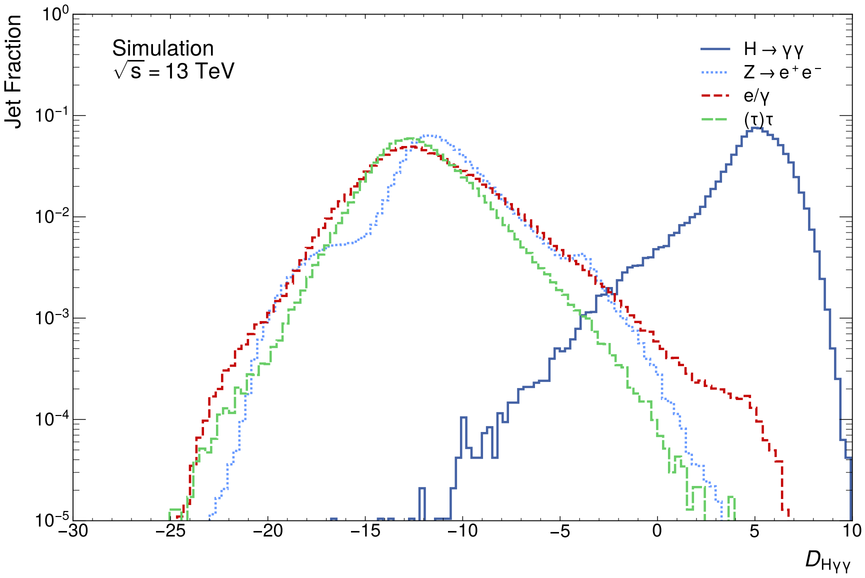


Figure 3.7: Distribution of the scalar discriminant $D_{H\gamma\gamma}$ score for all jet classes Clearly visible is a separation between the signal jet $D_{H\gamma\gamma}$ score distribution and that of the background jets.

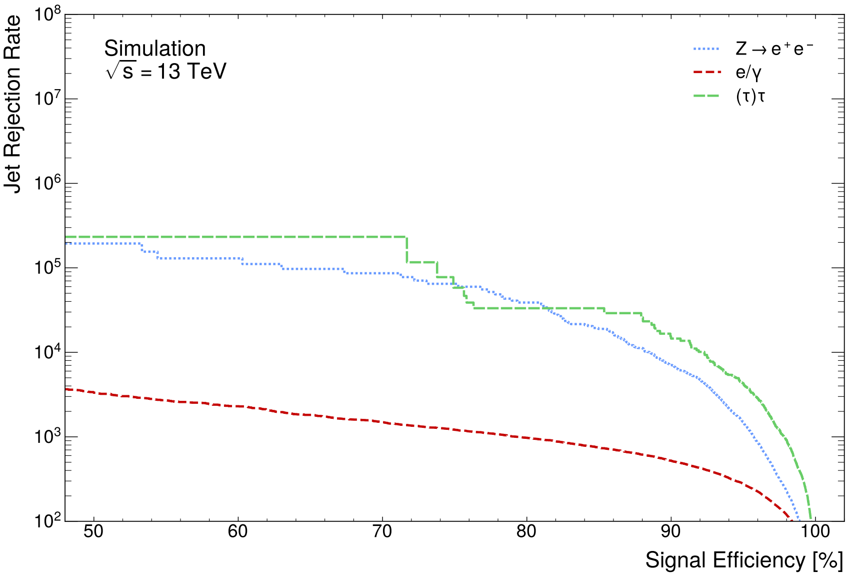


Figure 3.8: ANN classifier jet rejection rates of the background classes as a function of the signal efficiency. The shape of these curves are similar to those seen for the DNN, although the value of the jet rejection rates are reduced as seen in Table 3.4.

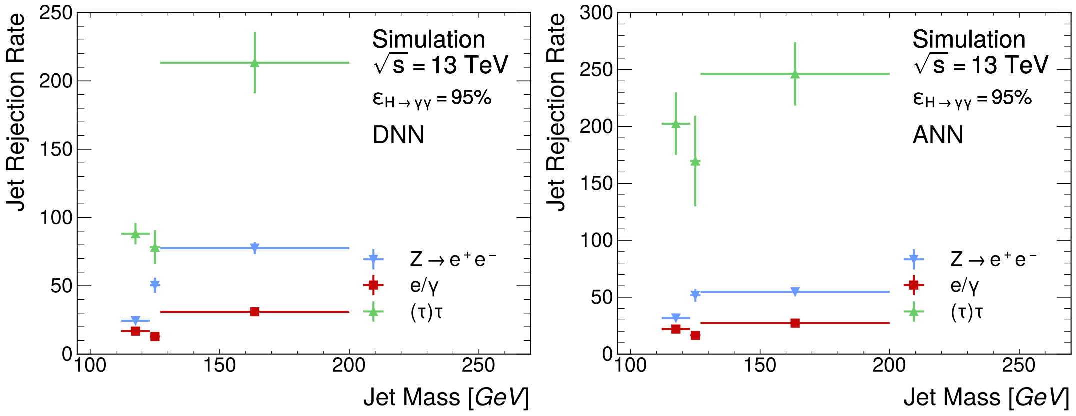


Figure 3.9: Background jet rejection rates as a function of mass at the $\varepsilon_{H\rightarrow\gamma\gamma} = 95\%$ operating point. Only three mass bins satisfy this criterion for both networks.

Table 3.5: Mutual information between the input feature variables and the scalar discriminant $D_{H\gamma\gamma}$ for the DNN and ANN jet taggers. Features are ranked in descending ANN MI score order.

Feature	DNN MI	ANN MI	ΔMI	ΔMI [%]
m	0.36	0.26	-0.10	-27.8
$r_{N=1}^{(\beta=1)}$	0.21	0.21	0.00	0.0
E/p	0.17	0.21	+0.04	+23.5
N^{Trks}	0.25	0.20	-0.05	-20.0
f_{EM}	0.18	0.15	-0.03	+16.7
$\max\left(\frac{E_{layer}}{E_{jet}}\right)$	0.21	0.14	-0.07	-33.3
p_T Bin	0.12	0.12	0.00	0.0
$\Delta R(c_1, c_2)$	0.14	0.12	-0.02	-14.3
$N^{Const.}$	0.14	0.11	-0.03	21.4
$\lvert\eta\rvert$ Bin	0.04	0.11	+0.07	+175.0
$\Delta R(c_2, j)$	0.11	0.10	-0.01	-9.1
Planar Flow	0.12	0.09	-0.03	-25.0
Balance	0.03	0.06	+0.03	+100.0
Width	0.06	0.04	-0.02	-33.3
$\Delta R(c_1, j)$	0.04	0.03	-0.01	-25.0

Chapter 4

Discussion

4.1 DNN

The overall performance of the DNN jet tagger is strong. It has a high signal classification rate at 99.09% and a high value for the other heavy boson $Z \to e^+e^-$ identifications at 96.46% seen in the confusion matrix in Figure 3.3. The DNN correctly predicts the $Z \to e^+e^-$ decays to a high degree, supporting this architecture's use for highly boosted, heavy boson identification. High AUC scores seen in Table 3.2 for the signal (1.00) and $Z \to e^+e^-$ decays (0.99) support this. It should be noted that the AUC scores are high for the other backgrounds, both above 0.90; for signal and $Z \to e^+e^-$ decays, they are near the possible maximum. The distribution of the scalar discriminant $D_{H\gamma\gamma}$ as seen in Figure 3.4 is very similar to the distribution of the related scalar discriminant D_{Zee} for $Z \to e^+e^-$ decays seen in previous work [23]. This indicates that the DNN jet tagger performs similarly to the jet tagging scenario in which it was previously implemented.

The DNN jet tagger achieves high background jet rejection rates for all backgrounds, as seen in Figure 3.5. The MC statistics are insufficient for the $Z \to e^+e^-$ decays and $(\tau)\tau$-jets, as evidenced by the curves possessing large steps at lower signal efficiencies. This motivates using a high signal efficiency as an operating point for the jet tagger as it remains on a smooth part of the curve. This reason, combined with its previous use as the highest (*tight*) operating point in related work [23] and the need for high efficiency for boosted Standard Model searches, motivates the choice of $\varepsilon_{H\to\gamma\gamma} = 95\%$. Future work may also generate more MC statistics so as to use lower operating points to assess the DNN, as related work uses multiple operating points [23]. The highly boosted $H \to \gamma\gamma$ decays DNN jet tagger developed in this work at the same operating point performs 39.5% worse on reducing e/γ-jets than the related highly boosted $Z \to e^+e^-$

jet tagger [23]. The rejection rate in Table 3.4 for e/γ-jets is in the same order as the previous result [23]. However, the DNN jet tagger developed here has an increased rejection rate for $(\tau)\tau$-jets of 1029.3% compared to the previous study [23]. Such a stark increase may reflect the difference in the distribution of MC-generated jets in the training dataset. Nonetheless, the DNN jet tagger is significantly more successful at reducing the $(\tau)\tau$-jets than the previous study. The DNN jet tagger is also highly successful at reducing the $Z \rightarrow e^+e^-$ decays; the rejection rate at the $\varepsilon_{H\rightarrow\gamma\gamma} = 95\%$ is of the same order as the rejection rates for the $(\tau)\tau$-jets which have been noted as high.

The best-performing DNN was selected based on the flatness of the signal efficiency curves with respect to the transverse momentum and pseudorapidity bins. These data are shown in Figure 3.1; performance strongly degrades at low transverse momentum bin values and the highest pseudorapidity bin value. The signal efficiency distribution largely follows the data distribution seen in Figure 2.1. This indicates that weighting the loss function to have a flat distribution with respect to the classes has a limited effect as it does not overcome the inherent data distribution. This is a limitation of the DNN jet tagger as the method of weighting the loss function to be flat with respect to these features is proving limited in its success.

The DNN jet tagger's performance aligns with related work for a highly boosted $Z \rightarrow e^+e^-$ decays [23], showing this architecture effectively identifies different jets using the ATLAS detector. The effectiveness of weighting the loss function to have a flat distribution with respect to the p_T and $|\eta|$ bins has, however, been identified as questionable due to its mirroring of the data distribution when the signal efficiency is calculated for these feature bins. A more complex method may be required in future work to improve identification invariance with respect to the transverse momentum and pseudorapidity bins.

4.2 ANN

The performance of the jet tagging sub-network of the ANN setup is a modicum worse than the DNN. Comparing the confusion matrices between these two classifiers, the ANN jet tagger is -0.73% less accurate on the signal classification. The ANN is -0.70% worse performing on $Z \rightarrow e^+e^-$ decay classification and -4.35% for the e/γ class. It performs slightly better on $(\tau)\tau$-jets with a $+1.88\%$ increase. All class confusion is approximately similar between the DNN and ANN. The $D_{H\gamma\gamma}$ distributions share the same attributes, and a clear separation between the background and signal jets is easily observed. The jet rejection rates of the background jets at the $\varepsilon_{H\rightarrow\gamma\gamma} = 95\%$ are about half for the e/γ and $(\tau)\tau$-jets when compared to the DNN, as seen in Table 3.4. The most significant degradation in rejection rate is for the $Z \rightarrow e^+e^-$ decays, which is -80.2% for the ANN compared to the rejection rate of the DNN jet tagger. Therefore, attempting to reduce the bias and sensitivity to the mass feature has reduced the ability of the jet tagger to reduce the background jets. However, in Figure 3.9, the effect on jet rejection rate as a function of mass between the DNN and ANN can be seen. In this metric, the distribution of rejection rates for all the background jets is somewhat flattened, and there is a clear increase in the rejection rate of the $(\tau)\tau$-jets. Unfortunately, only the three mass bins closest to the signal mass satisfy the operating point criterion. The case for both the DNN and ANN is due to the availability of signal MC statistics in the data and could be mitigated by a differently generated dataset.

The ability of the jet tagger to classify the signal and reduce backgrounds is comparable to the ability of the DNN. How it uses the available information in the features is of significant interest as reducing dependence on one feature may reasonably suggest that other features are utilised more. Table 3.5 contains the mutual information between the feature variables and the scalar discriminant of the jet taggers. As reported in Table 3.5, the greatest absolute change in mutual information score occurs with the mass feature. This strongly indicates that this feature's adversarial training targeting decorrelation has occurred as intended. It should be noted that the mass feature still possesses the highest mutual information with the discriminant. Four features increased their mutual information score with the discriminant, E/p, f_{EM}, $|\eta|$ bin and the balance. These features are not strongly associated with the mass, so it is reasonable that the network would learn more from these features than the DNN did to compensate for the reduced importance of the mass feature. The increased dependence on the $|\eta|$ bin feature is relatively large in absolute and percentage terms. This is an issue as it is attempted to make the network decorrelated to this feature by weighting the loss function during the training. Whilst this is a concern and may warrant more attention to the invariance of this feature in the future, the $|\eta|$ bin dependence is now in line with the transverse momentum bin feature for which the mutual information score has remained unchanged between the DNN and ANN. Dependence on f_{EM} has increased; this result is understandable due to the di-photon decay products of the signal jet, which has a relatively distinct distribution from the e/γ and $(\tau)\tau$-jets, as seen in Figure 2.2. However, its similarity to the $Z \rightarrow e^+e^-$ decays may explain why the ANN jet tagger has not utilised this feature even more. Interestingly, there is no change in the mutual information score between the jet substructure feature $r_{N=1}^{(\beta=1)}$ and the discriminant, comparing the DNN and the ANN. It does, however, become the second most important feature due to the decrease in the importance of the multiplicity of the tracks per jet. So, relatively, the substructure feature of the jet becomes more influential in the decision process of the jet tagger.

As mentioned, the ability of the ANN jet tagger to learn to decorrelate from the mass is highly influenced by the availability of examples of the jets in each mass decile bin. This has limited the ability of the ANN to decorrelate from mass evenly in different regions of the mass due to the availability of MC statistics from signal events. As MC-generated signal jets are clustered around the Higgs boson mass at 125.09 GeV [12], examples at lower masses, particularly, are limited. The same datasets for training, validation and testing were used for the ANN and the DNN. In future efforts employing the ANN architecture for feature decorrelation, a significant focus should be on generating MC events that are more suitable for this task. Allowing a broader distribution of signal masses around the peak of 125.09 GeV should immediately improve the distribution of the examples in the data and improve the ability of the ANN to learn to identify jets and decorrelate from the mass more easily across a greater mass range.

It should also be noted that related work in mass-decorrelated jet tagging has been done in binary classification scenarios, a signal class and a combined backgrounds class [36, 68, 69]. Thus, this multiclass classification task that has been set is inherently more complex than previous work. This may explain why the degree of mass invariance seen in related work is greater than what has been determined in this thesis.

In summary, the ANN jet tagger's performance is qualitatively and quantitatively similar to the DNN network, which has been shown effective to the same order as previous work on another highly boosted heavy boson [23]. This suggests that the ANN jet tagger is at least a viable architecture in classification ability compared to another functioning jet tagger architecture. There has been a 27.8% reduction in the dependence on the mass variable as measured using the mutual information between the feature and the discriminant $D_{H\gamma\gamma}$. Combined with the flattened distributions of the background jet rejection rates for the three operating point satisfying mass bins, there is good evidence that this approach has somewhat decorrelated the jet tagger's identifications from the mass feature data. This decorrelation is limited, with the most likely culprit for the limitation identified as the distribution of the MC-generated training data with respect to mass. As the datasets used are MC generated, engineering this data to tailor the training scenario to the ANN setup is likely the single most effective recommendation to improve the mass invariance of the identifications made by the ANN jet tagger.

4.3 Directions for Future Research

The previous sections of this chapter discussed the capabilities and limitations of the developed jet taggers. The found limitations motivate directions for future research as mitigating or entirely avoiding these limitations come from three main areas. The first route is better hyperparameter optimisation, thus searching for a better model within the described architecture. The second is to extend the ANN feature decorrelation technique to other features for which reduced sensitivity and bias are desirable.

4.3.1 Hyperparameter Optimisation

The hyperparameter grid search spaces were defined based on the results found in previous and related work [23, 36]. However, a different hyperparameter optimisation technique may be more appropriate for both networks, especially the ANN jet tagger, due to its significantly greater number of hyperparameters.

One such method is random search; a hyperparameter space is defined, and the hyperparameter combinations are then randomly selected from this space [63]. This comes with significant benefits, especially when dealing with many hyperparameters. First, computational efficiency: random search usually requires fewer trials to find a good set of hyperparameters compared to grid search [63]. Secondly, random search can explore a larger hyperparameter space in a given amount of time than grid search [63]. Random search also does not waste resources computing very similar hyperparameter combinations. Finally, random search can converge to an optimal solution faster than grid search for larger hyperparameter spaces [63]. In this thesis, the hyperparameter space was restricted to make the research compatible with the computational and time resources available; even with this limitation, a proof-of-concept ANN model was found. A greater hyperparameter space can be searched using random search to improve hyperparameter optimisation efficiency, and a better-performing ANN $H \to \gamma\gamma$ decays jet tagger will likely be found.

4.3.2 Extending ANN to Other Features

The jet tagger developed using adversarial training has measurably reduced systematic uncertainties by reducing the bias on the mass feature variable compared to the DNN jet tagger. This was done using a regularising term in the loss function for the mass feature that was determined by the powerful and flexible adversary DNN. However, there are concerns about the jet tagger being relatively sensitive to the transverse momentum and pseudorapidity of the decay products. This was handled by weighting the loss function with respect to these features. This method is less effective than desired, as seen for the DNN. Applying the adversarial network architecture and training to the transverse momentum and pseudorapidity may be quite effective, especially as these features are much less strongly correlated to other features than the mass feature. This can be seen in the correlation plot in Figure 2.3.

One could create one ANN setup per feature from which the jet tagger output is to be decorrelated, then create an ensemble of these networks [70]. The ensemble approach would simplify training significantly as a single jet tagger per decorrelated feature variable would be required instead of regularising all N features in a single complex training setup. Each training scenario would then be on par with what has been developed and shown to have the desired effect in this thesis. At inference time, the N jet taggers would vote on the jet class [70]. The weighting of the jet taggers in the ensemble can then be tuned to a user's specific requirements. Ensembles are generally better than single classifiers as they tend to be more accurate than their constituent parts [71]. This is because they can capture different representations in the data and pool this information into the final output classification [71]. Training could also be highly parallelised as each ANN could be trained independently, reducing the computational expense of training a single highly complex adversarial training scenario.

Developing further ensemble ANN jet taggers would allow the taggers to be more general and would likely be effective for more boosted decays that result in two photons. This would allow the legitimate use of such a general jet tagger for searches of hypothetical particles such as additional Higgs bosons [72], axion-like particles [73], and various dark matter candidate particles [74].

In the next and final chapter, conclusive remarks highlight this thesis's key contributions and outcomes.

Chapter 5

Conclusions

In this thesis, two jet tagging algorithms have been developed and presented to identify highly boosted $H \rightarrow \gamma\gamma$ decays using the ATLAS detector at the LHC. The first jet tagging algorithm based on the DNN architecture was found to perform to a similar level as a related jet tagger used for highly boosted $Z \rightarrow e^+e^-$ decays. Noise from the background jets was reduced to a similar degree. Additionally, some performance improvements were observed in the $H \rightarrow \gamma\gamma$ jet tagger, which was found to reject $(\tau)\tau$-jets at a higher rate than the related jet tagger. Thus, the research aim of creating a robust jet tagging algorithm for highly boosted $H \rightarrow \gamma\gamma$ decays has been achieved.

The second jet tagger developed using adversarial training was found to have a slightly degraded performance compared to the DNN jet tagger. However, this performance reduction was overall relatively mild, meaning the ANN jet tagger could still be reliably used to detect $H \rightarrow \gamma\gamma$ decays. A 27.8% reduction in mutual information between the mass and scalar discriminant was found between the DNN and ANN jet taggers. Indicating a reduction in the importance of the mass feature to the jet tagger's decision-making. There is also evidence that the rejection rates for the background jets have a flatter distribution compared to the DNN at the 95% operating point.

The evidence for improved mass-decorrelated classification predominantly indicates what can be achieved; several recommendations have been made to move beyond this proof-of-concept mass-decorrelated jet tagger to an actual tool for LHC experiments. The most important of these to improve the ANN jet tagger is to MC-generate the jet statistics purposefully with downstream ANN training in mind. This would involve generating $H \rightarrow \gamma\gamma$ decays with a significantly broader mass peak. Additional recommendations for the ANN are to change the hyperparameter optimisation method to random search to increase the explored hyperparameter space and to create ensembles of these jet taggers. These are the methods for further developing a mass-decorrelate ANN jet tagger and reducing the limitations found in this thesis.

Thus, the second research aim to produce a mass-decorrelated jet tagger is shown to be possible. Still, considerable work and further research must be done to make this model sufficiently insensitive to mass to reduce the systematic bias caused by the mass to a significant degree. This is especially the case due to the much higher complexity of the ANN than the DNN jet tagger, which is much cheaper to develop.

To conclude, the research aims have either been fully achieved or robust progress has been made to achieve them. Jet tagging algorithms of sufficient sensitivity have been developed to improve the detection of highly boosted Higgs boson di-photon decays in the ATLAS detector. Thus furthering the search for new physics in discovering properties beyond the Standard Model and the search for discrepancies within the Standard Model.

Bibliography

[1] S. L. Glashow, "Partial-symmetries of weak interactions," *Nuclear physics*, vol. 22, no. 4, pp. 579–588, 1961.

[2] M. Veltman *et al.*, "Regularization and renormalization of gauge fields," *Nuclear Physics B*, vol. 44, no. 1, pp. 189–213, 1972.

[3] S. Weinberg, "Electromagnetic and weak masses," *Physical Review Letters*, vol. 29, no. 6, p. 388, 1972.

[4] S. Weinberg, "Physical processes in a convergent theory of the weak and electromagnetic interactions," *Physical Review Letters*, vol. 27, no. 24, p. 1688, 1971.

[5] N. Svartholm, *Elementary Particle Physics: Relativistic Groups and Analyticity: Proceedings of the Eighth Nobel Symposium Held May, 1968 at Lerum, Sweden*. Almquvist & Wiksell, 1968.

[6] P. W. Higgs, "Broken symmetries, massless particles and gauge fields," *Phys. Lett.*, vol. 12, pp. 132–133, 1964.

[7] P. W. Higgs, "Broken symmetries and the masses of gauge bosons," *Physical review letters*, vol. 13, no. 16, p. 508, 1964.

[8] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons," *Physical Review Letters*, vol. 13, no. 9, p. 321–323, 1964.

[9] ATLAS Collaboration, "The atlas experiment at the cern large hadron collider," *Jinst*, vol. 3, p. S08003, 2008.

[10] S. Weinberg, "A model of leptons," *Physical review letters*, vol. 19, no. 21, p. 1264, 1967.

[11] M. Carena and H. E. Haber, "Higgs boson theory and phenomenology," *Progress in Particle and Nuclear Physics*, vol. 50, no. 1, pp. 63–152, 2003.

[12] ATLAS and CMS Collaborations, “Combined measurement of the higgs boson mass in pp collisions at $\sqrt{7}$ and 8 TeV,” *Physical Review Letters*, vol. 114, no. 19, 2015.

[13] CMS Collaboration, “Measurements of properties of the higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV,” *Journal of High Energy Physics*, vol. 2017, no. 11, 2017.

[14] CMS Collaboration, “Measurements of higgs boson properties in the diphoton decay channel with 36 fb1 of pp collision data at $\sqrt{s} = 13$ tev with the atlas detector,” *Physical Review D*, vol. 98, no. 5, 2018.

[15] CMS Collaboration, “Measurement and interpretation of differential cross sections for higgs boson production at $\sqrt{s} = 13$ TeV,” *Physics Letters B*, vol. 792, pp. 369–396, may 2019.

[16] ATLAS Collaboration, “Measurement of the production cross section for a higgs boson in association with a vector boson in the $h \to ww^* \to l\nu l\nu$ channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector,” *Physics Letters B*, vol. 798, p. 134949, nov 2019.

[17] ATLAS Collaboration, “Measurement of the four-lepton invariant mass spectrum in 13 TeV proton-proton collisions with the atlas detector,” *Journal of High Energy Physics*, vol. 2019, no. 4, 2019.

[18] ATLAS Collaboration, “Study of the spin and parity of the higgs boson in diboson decays with the ATLAS detector,” *The European Physical Journal C*, vol. 75, oct 2015.

[19] ATLAS Collaboration, “Measurements of the higgs boson inclusive and differential fiducial cross-sections in the diphoton decay channel with pp collisions at $\sqrt{s}$ = 13 TeV with the atlas detector,” *Journal of High Energy Physics*, vol. 2022, no. 8, 2022.

[20] T. D. Lee, “A theory of spontaneous T violation,” *Physical Review D*, vol. 8, no. 4, p. 1226–1239, 1973.

[21] CMS Collaboration, “The CMS experiment at the CERN LHC,” *Journal of Instrumentation*.

[22] M. Cacciari, G. P. Salam, and G. Soyez, “The anti-k_t jet clustering algorithm,” *Journal of High Energy Physics*, vol. 2008, pp. 063–063, apr 2008.

[23] ATLAS Collaboration, “Identification of highly boosted $Z \to e^+e^-$ decays with the ATLAS detector using deep neural networks,” tech. rep., CERN, Geneva, 2022.

[24] ATLAS Collaboration, “Search for higgs boson pair production in the two bottom quarks plus two photons final state in pp collisions at,” *Physical Review D*, vol. 106, no. 5, 2022.

[25] ATLAS Collaboration, “ATLAS Insertable B-Layer Technical Design Report Addendum,” tech. rep., 2012. Addendum to CERN-LHCC-2010-013, ATLAS-TDR-019.

[26] ATLAS Collaboration, “Measurement of the photon identification efficiencies with the atlas detector using lhc run-1 data,” *The European Physical Journal C*, 2016.

[27] ATLAS Collaboration, “Electron and photon performance measurements with the atlas detector using the 2015–2017 lhc proton-proton collision data,” *Journal of Instrumentation*, vol. 14, p. P12006, dec 2019.

[28] F. Bishara, Y. Grossman, R. Harnik, D. J. Robinson, J. Shu, and J. Zupan, “Probing cp violation in h$\to$ $\gamma\gamma$ with converted photons,” *Journal of High Energy Physics*, vol. 2014, no. 4, pp. 1–41, 2014.

[29] J. C. Sharp, “Symmetry of the lorentz boost: the relativity of colocality and lorentz time contraction,” *European Journal of Physics*, vol. 37, no. 5, p. 055606, 2016.

[30] S. Fartoukh *et al.*, “LHC configuration and operational scenario for run 3,” tech. rep., 2021.

[31] M. A. Nielsen, *Neural networks and deep learning*, vol. 25. Determination press San Francisco, CA, USA, 2015.

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, p. 533–536, 1986.

[33] CMS Collaboration, “Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques,” *Journal of Instrumentation*, vol. 15, pp. P06005–P06005, jun 2020.

[34] ATLAS Collaboration, “Reconstruction and identification of boosted di-τ systems in a search for higgs boson pairs using 13 TeV proton-proton collision data in atlas,” *Journal of High Energy Physics*, vol. 2020, no. 11, pp. 1–47, 2020.

[35] J. Dolen, P. Harris, S. Marzani, S. Rappoccio, and N. Tran, “Thinking outside the rocs: Designing decorrelated taggers (ddt) for jet substructure,” *Journal of High Energy Physics*, vol. 2016, no. 5, 2016.

[36] C. Shimmin *et al.*, “Decorrelated jet substructure tagging using adversarial neural networks,” *Physical Review D*, vol. 96, no. 7, p. 074034, 2017.

[37] J. K. Basu, D. Bhattacharyya, and T.-h. Kim, “Use of artificial neural network in pattern recognition,” *International journal of software engineering and its applications*, vol. 4, no. 2, 2010.

[38] T. Gleisberg, S. Höche, F. Krauss, A. Schälicke, S. Schumann, and J.-C. Winter, “Sherpa 1., a proof-of-concept version,” *Journal of High Energy Physics*, vol. 2004, p. 056, mar 2004.

[39] T. Gleisberg, S. Höche, F. Krauss, M. Schönherr, S. Schumann, F. Siegert, and J. Winter, “Event generation with sherpa 1.1,” *Journal of High Energy Physics*, vol. 2009, p. 007, feb 2009.

[40] E. Bothmann *et al.*, “Event generation with sherpa 2.2,” *SciPost Physics*, vol. 7, no. 3, 2019.

[41] J. Alwall *et al.*, “Madgraph/madevent v4: the new web generation,” *Journal of High Energy Physics*, vol. 2007, p. 028, sep 2007.

[42] J. Bellm *et al.*, “Herwig 7.0/herwig++ 3.0 release note,” *The European Physical Journal C*, vol. 76, no. 4, 2016.

[43] T. Sjöstrand *et al.*, “An introduction to pythia 8.2,” *Computer Physics Communications*, vol. 191, pp. 159–177, 2015.

[44] F. James, “Monte carlo theory and practice,” *Reports on progress in Physics*, vol. 43, no. 9, p. 1145, 1980.

[45] C. Bierlich *et al.*, “A comprehensive guide to the physics and usage of pythia 8.3,” *SciPost Physics Codebases*, p. 008, 2022.

[46] V. Khachatryan *et al.*, “Transverse-momentum and pseudorapidity distributions of charged hadrons in p p collisions at s= 7 tev,” *Physical Review Letters*, vol. 105, no. 2, p. 022002, 2010.

[47] A. J. Larkoski, G. P. Salam, and J. Thaler, “Energy correlation functions for jet substructure,” *Journal of High Energy Physics*, vol. 2013, no. 6, 2013.

[48] ATLAS Collaboration, “Selection of jets produced in 13 TeV proton-proton collisions with the ATLAS detector,” tech. rep., CERN, Geneva, 2015.

[49] A. Tsanas, M. Little, and P. McSharry, “A methodology for the analysis of medical data,” in *Handbook of systems and complexity in health*, pp. 113–125, Springer, 2012.

[50] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, “Normalized mutual information feature selection,” *IEEE Transactions on neural networks*, vol. 20, no. 2, pp. 189–201, 2009.

[51] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[52] A. Paszke *et al.*, “Pytorch: An imperative style, high-performance deep learning library,” *Advances in neural information processing systems*, vol. 32, 2019.

[53] U. of Edinburgh, “Edinburgh compute and data facility web site.” `www.ecdf.ed.ac.uk`, 2023. Accessed: 25 August 2023.

[54] T. D. Sanger, “Optimal unsupervised learning in a single-layer linear feedforward neural network,” *Neural networks*, vol. 2, no. 6, pp. 459–473, 1989.

[55] D.E. Rumelhart *et al.*, “Learning internal representations by error propagation,” 1985.

[56] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

[57] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, JMLR Workshop and Conference Proceedings, 2010.

[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, pmlr, 2015.

[59] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[61] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[62] L. Prechelt, "Early stopping-but when?," in *Neural Networks: Tricks of the trade*, pp. 55–69, Springer, 2002.

[63] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization.," *Journal of machine learning research*, vol. 13, no. 2, 2012.

[64] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," *arXiv preprint arXiv:1704.00805*, 2017.

[65] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[66] J. Neyman and E. S. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypotheses*, pp. 73–108. New York, NY: Springer New York, 1992.

[67] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[68] T. Cheng and A. Courville, "Invariant representation driven neural classifier for anti-qcd jet tagging," *Journal of High Energy Physics*, vol. 2022, no. 10, 2022.

[69] T. Cheng, J.-F. Arguin, J. Leissner-Martin, J. Pilette, and T. Golling, "Variational autoencoders for anomalous jet tagging," *Physical Review D*, vol. 107, no. 1, 2023.

[70] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, p. 123–140, 1996.

[71] T. G. Dietterich, *Ensemble Methods in Machine Learning*, p. 1–15. Advances in Grid and Pervasive Computing, 2000.

[72] R. Barbieri *et al.*, “One or more higgs bosons?,” *Physical Review D*, vol. 88, no. 5, 2013.

[73] M. Bauer *et al.*, “Collider probes of axion-like particles,” *Journal of High Energy Physics*, vol. 2017, no. 12, 2017.

[74] F. Kahlhoefer, “Review of lhc dark matter searches,” *International Journal of Modern Physics A*, vol. 32, no. 13, p. 1730006, 2017.

Appendix A

Features with Default Values

Default values for the feature variables that possess them are shown in Table A.1. A default value is provided when the particles in question do not record a certain feature in the ATLAS detector; these values are distinct from the data distribution. Default values help the models learn representations that include information for instances in the data that do not have natural values for certain features.

Table A.1: Feature variables with default values and what these default values are. When a value for a certain feature cannot be MC-generated due to the physical process in question, a default value is provided to prevent generating incomplete data.

Feature	Default Value
$\Delta R(t_2, j)$	0.5
$\Delta R(t_1, t_2)$	-0.1
$r_{N=1}^{(\beta=1)}$	0.0
E/p	-1.0
Balance	-1.0

Appendix B

Enlarged Multiplicity Feature Bar Plots

In Figure B.1, enlarged bar plots for the N^{Trks} and $N^{Const.}$ features are presented. This makes these same important bar plots in Figure 2.1 more readable.

The distributions of these features discussed in Chapter 2 are more easily seen in these larger plots. The $H \to \gamma\gamma$ decays are only in relatively large numbers in the early bin indices for both features. Whilst all classes tail off at the higher bin indices for both features, all backgrounds are spread much more across the feature bins.

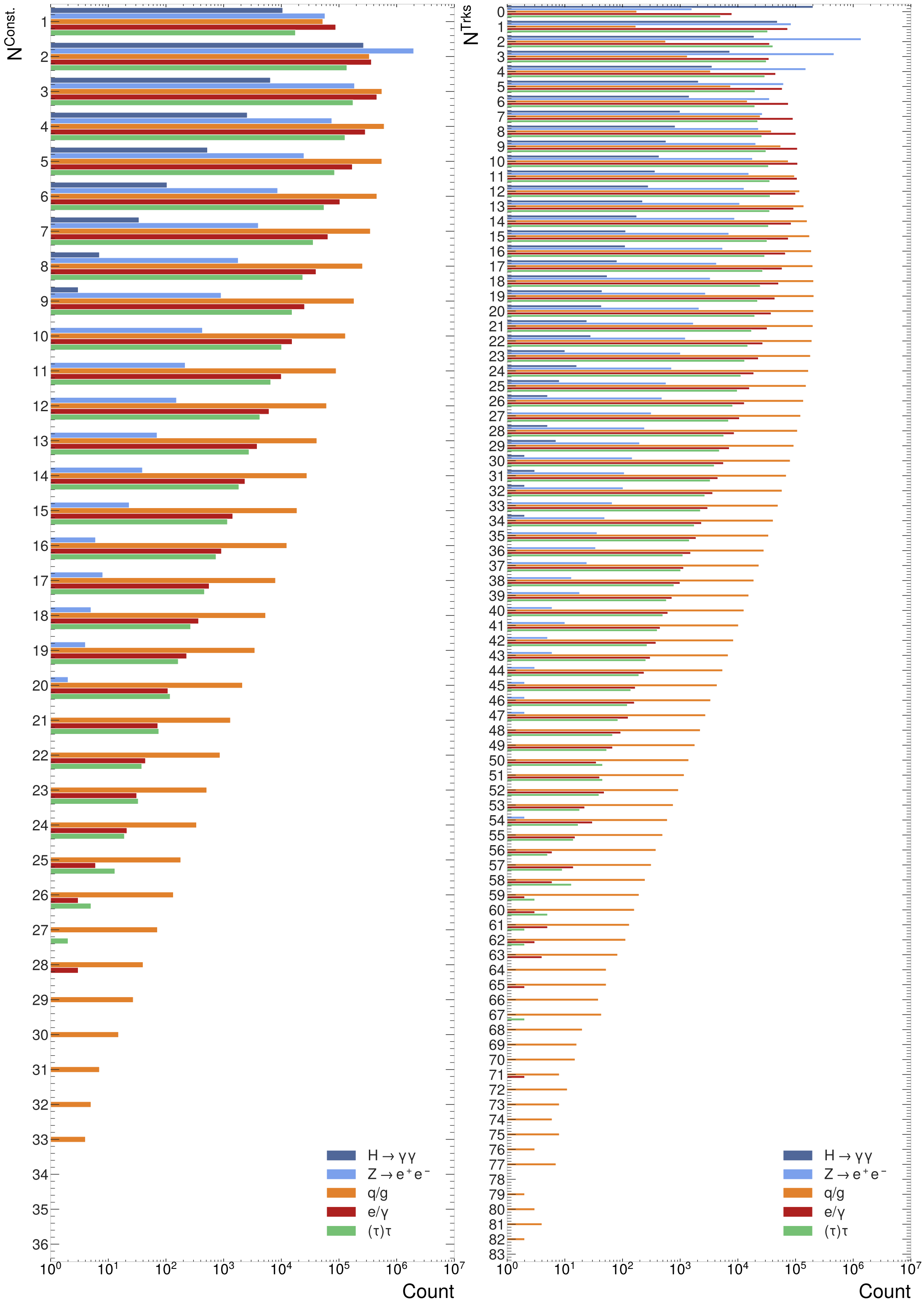


Figure B.1: Enlarged bar plots for the multiplicity of particle-flow objects per jet (left) and multiplicity of jet tracks (right).

Appendix C

Additional Network Figures

This appendix presents figures for the ANN setup, such as the training curves and the signal efficiency as functions of the transverse momentum and pseudorapidity bins. As well as the ROC curves for the DNN and ANN jet taggers.

The training curves seen in Figure C.1 show the optimisation of the ANN classifier over its training epochs. Fluctuation in these curves is greater than seen for the DNN, especially for the loss curves, which show much more instability during training, which was expected considering the more complex training scenario. A greater gap between training and validation curves indicates that overfitting is a more significant problem for the ANN training setup.

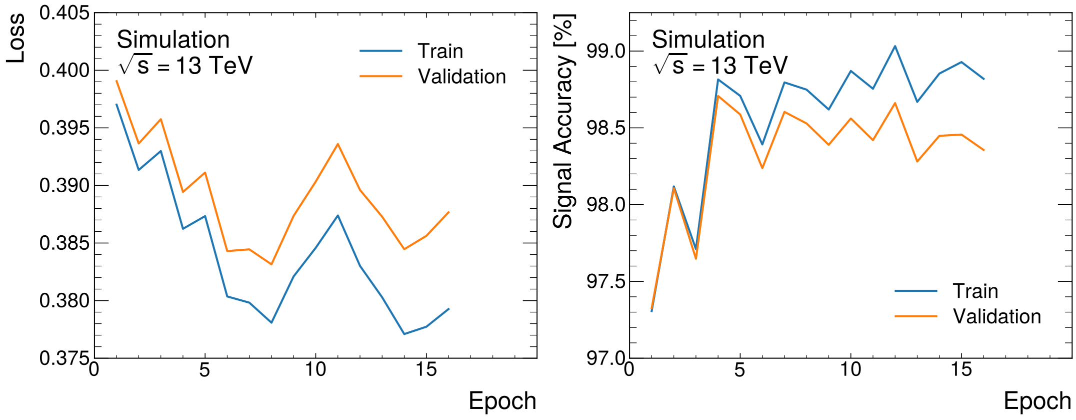


Figure C.1: Training curves for the four-class ANN classifier. There is a greater separation between training and validation curves than was seen for the DNN, suggesting a larger degree of overfitting. There is also more instability in the loss curve than in the DNN training.

Figure C.2 displays the signal efficiency as a function of both the transverse momentum and absolute pseudorapidity. The signal efficiencies across the features of both plots are very similar to what was observed for the DNN jet tagger.

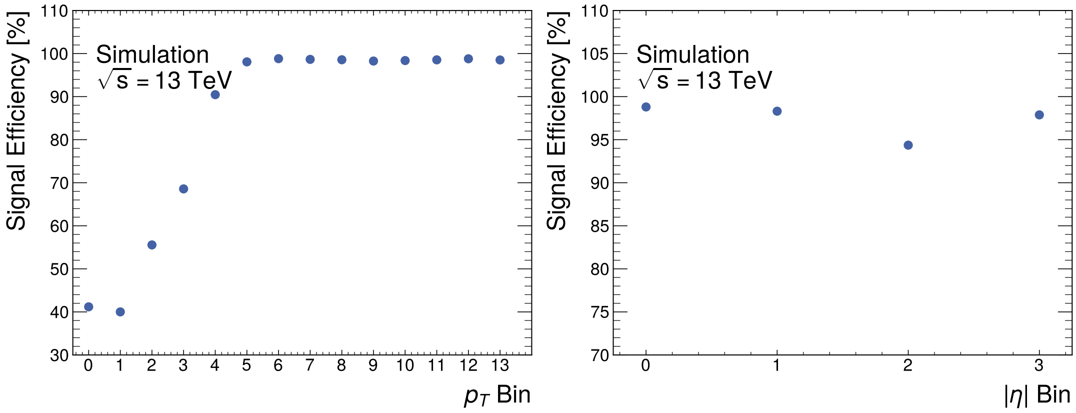


Figure C.2: Signal efficiency as a function of binned transverse momentum (left) and binned absolute pseudorapidity (right) for the ANN jet tagger. The signal efficiency across these features is similar to what was seen in for the DNN jet tagger.

Figures C.3 and C.4 show the ROC curves for the DNN and ANN jet taggers respectively. Both figures are visually extremely similar. AUC scores are the same for the signal and $Z \rightarrow e^+e^-$ decays remain the same and at or near the maximum value, whilst the AUC scores for the e/γ and $(\tau)\tau$-jets are very slightly reduced.

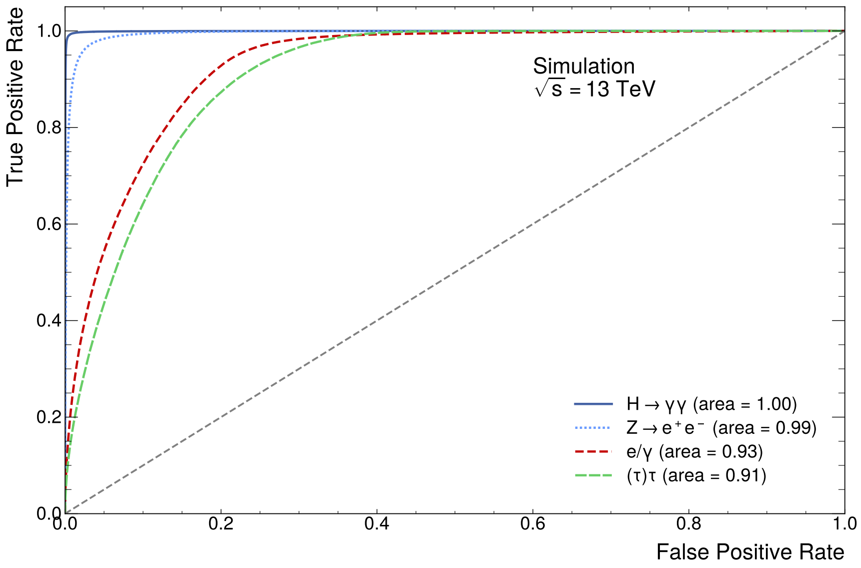


Figure C.3: DNN ROC curves for signal and background jets. AUC values for each jet are also displayed in the figure. The AUC values are high for all jet classes; the signal jet has the highest AUC value and has essentially achieved the possible maximum.

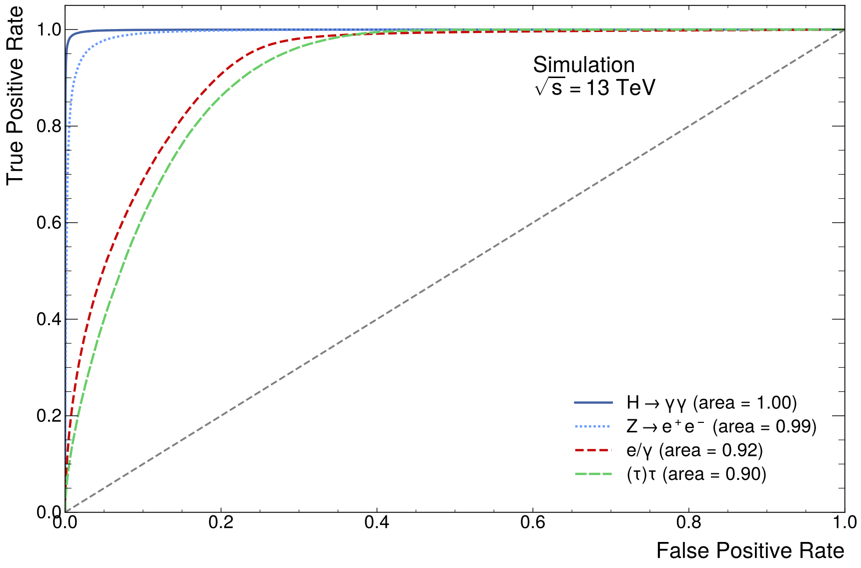


Figure C.4: ANN ROC curves for signal and background jets. AUC values for each jet are also displayed in the figure. Performance on the two heavy boson jets remains the same, whilst AUC scores on e/γ and $(\tau)\tau$-jets are both reduced by 0.01.

Appendix D

Scalar Discriminant for $Z \rightarrow e^+e^-$ decays

Appendix D presents the scalar discriminant for $Z \rightarrow e^+e^-$ decays for both four-class jet taggers. The scalar discriminant D_{Zee} for these jets is defined similarly to the scalar discriminant $D_{H\gamma\gamma}$ for the $H \rightarrow \gamma\gamma$ decays, D_{Zee} is given in equation D.1. This scalar discriminant for a background jet is examined in more detail in this appendix as it is the scalar discriminant used to assess performance for a previously developed $z \rightarrow e^+e^-$-jet DNN jet tagger [23].

$$D_{Zee} = \ln\left(\frac{f_{Z\rightarrow e^+e^-} \cdot p_{Z\rightarrow e^+e^-}}{f_{H\rightarrow\gamma\gamma} \cdot p_{H\rightarrow\gamma\gamma} + f_{e/\gamma} \cdot p_{e/\gamma} + f_{(\tau)\tau} \cdot p_{(\tau)\tau}}\right) \tag{D.1}$$

Figure D.1 displays the D_{Zee} distribution for the DNN jet tagger. Much like for the distributions of $D_{H\gamma\gamma}$ a clear separation between the $z \rightarrow e^+e^-$ and other jets is observed. The greatest separation exists between the $H \rightarrow \gamma\gamma$ decays and the $Z \rightarrow e^+e^-$ decays. This separation is somewhat reduced for the D_{Zee} distributions for the ANN jet tagger, as seen in Figure D.2. Thus showing from another metric that the ANN has a slightly degraded performance compared to the DNN.

In related work, there is a greater separation between what is the signal $Z \rightarrow e^+e^-$ decays and the background jets than is seen in these jet taggers, even the DNN [23]. This is likely due to a greater similarity between $Z \rightarrow e^+e^-$ decays and $H \rightarrow \gamma\gamma$ decays than the other background jets used in that research, which are q/g, e/γ and $(\tau)\tau$-jets [23].

Overall, the D_{Zee} distributions in the previous work are more similar in distribution at range to the $D_{H\gamma\gamma}$. Therefore, comparing signal performance between these three networks, despite one having a different signal jet, is relevant and strengthens conclusions made during these comparisons.

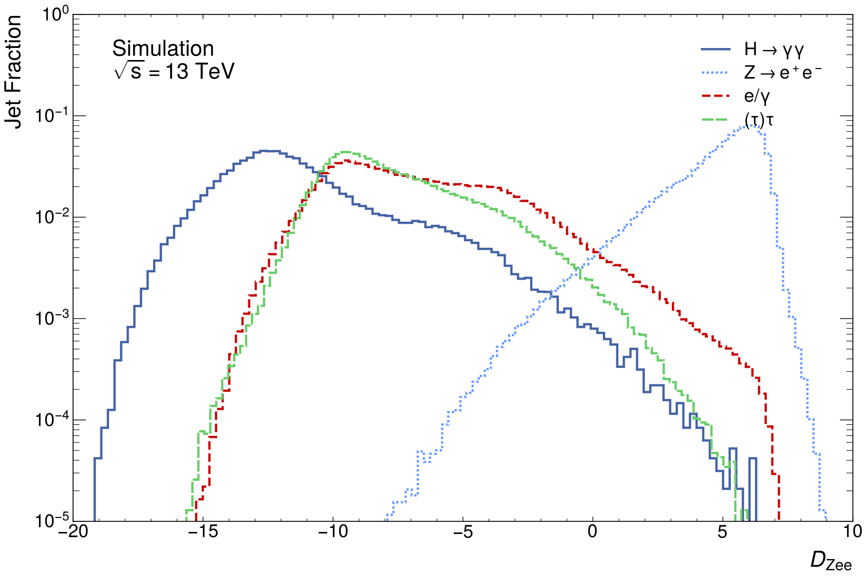


Figure D.1: Distribution of the scalar discriminant D_{Zee} score for all jet classes for the DNN. Clearly visible is a separation between the signal jet D_{Zee} score distribution and that of the other jets.

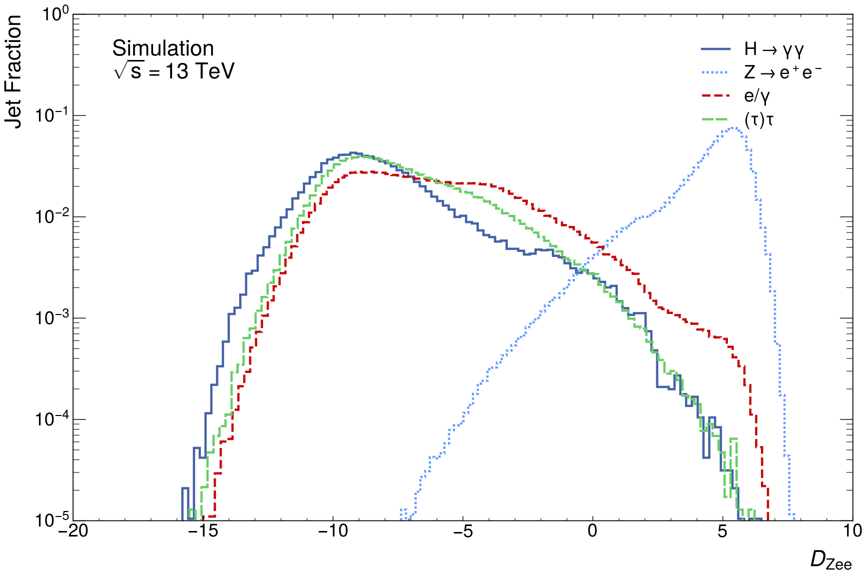


Figure D.2: Distribution of the scalar discriminant D_{Zee} score for all jet classes for the ANN. A separation between the signal jet D_{Zee} score distribution and that of the background jets is clearly visible.

Appendix E

Five-Class Jet Taggers

Both jet tagging algorithms are trained with MC-generated statistics for five jet classes. The additional background q/g-jets have been included in the training dataset. Both models are trained using the set of hyperparameters found for the four-class jet taggers. However, the networks are allowed to train for 50 epochs with early stopping implemented.

E.1 DNN

The DNN jet tagger is allowed to train for 50 epochs; early stopping was initiated at epoch 35. Thus, slightly longer training was beneficial to the five-class network compared to the four-class network studied in the main text of this thesis due to the increased complexity and amount of data by adding the most numerous jet class.

The training of the five-class DNN jet tagger proceeds very similarly to the four-class and is visually very similar. These curves are displayed in Figure E.1, both the decreasing loss function and high but fluctuating signal accuracy as a function of the epoch proceed as would be expected based on the four-class DNN jet tagger training seen in Figure 3.2. The only notable difference is the slightly longer training time required by the five-class jet tagger, with 35 epochs elapsing before the same early stopping criterion is met versus 27 epochs for the four-class jet tagger.

The signal efficiency as a function of the transverse momentum and pseudorapidity bins is observed in Figure E.2. The loss function is notably weighted during training to try and make the decision-making decorrelated to these features. This works quite well for the signal efficiency as a function of the pseudorapidity bins but is less effective for it as a function of the transverse momentum bins. However, the signal efficiency as a function of the transverse momentum bins is a bit flatter for the five-class DNN jet tagger versus the four-class model. The minimum signal efficiency is greater by over 10% meaning the signal efficiency is spread over a significantly smaller range across the transverse momentum bins.

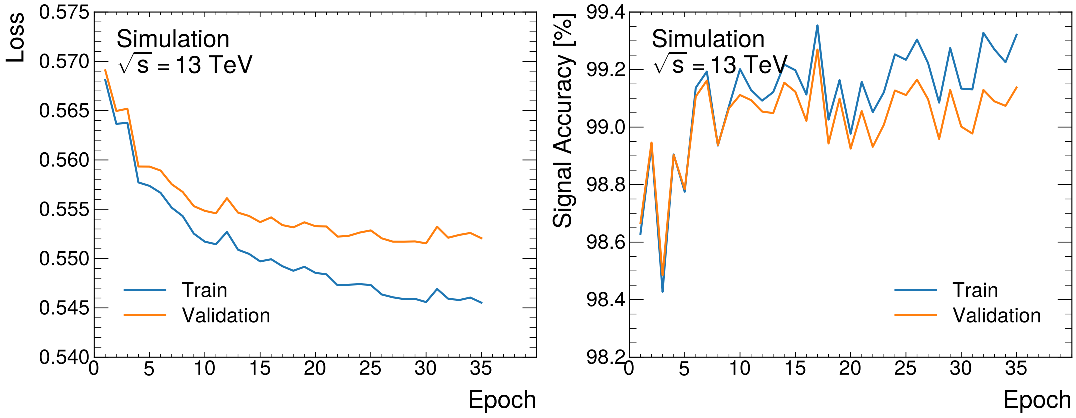


Figure E.1: Training curves for the best performing DNN jet tagger. The cross-entropy loss is plotted (left) as a function of the training epochs, and the signal accuracy is also presented as a function of the training epochs (right). All curves are computed for both the training and validation datasets.

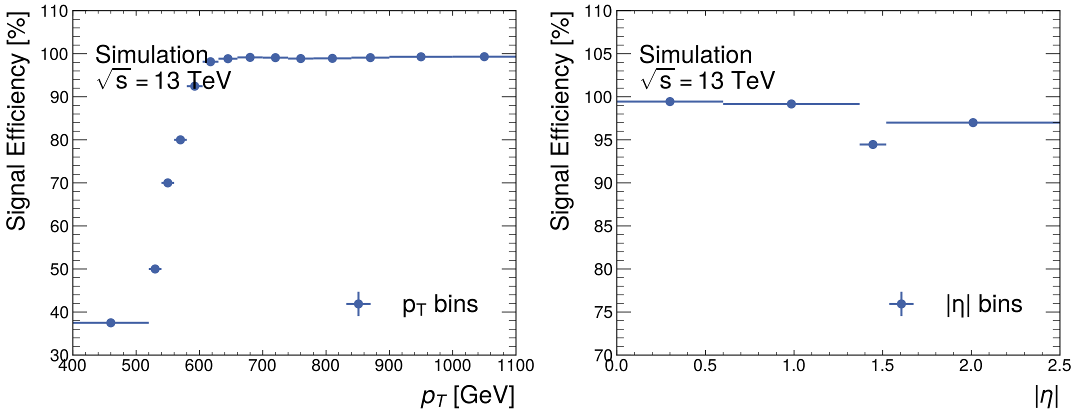


Figure E.2: Signal efficiency for the five-class DNN jet tagger with respect to the transverse momentum bins (left) and the pseudorapidity bins (right). The signal efficiency as a function of the $|\eta|$ bins is largely the same as for the four-class DNN, as seen in Figure 3.1. The curve's shape is very similar between the four and five-class models for signal efficiency against the transverse momentum bins, the minimum is about 10% greater for the five-class DNN than the four-class.

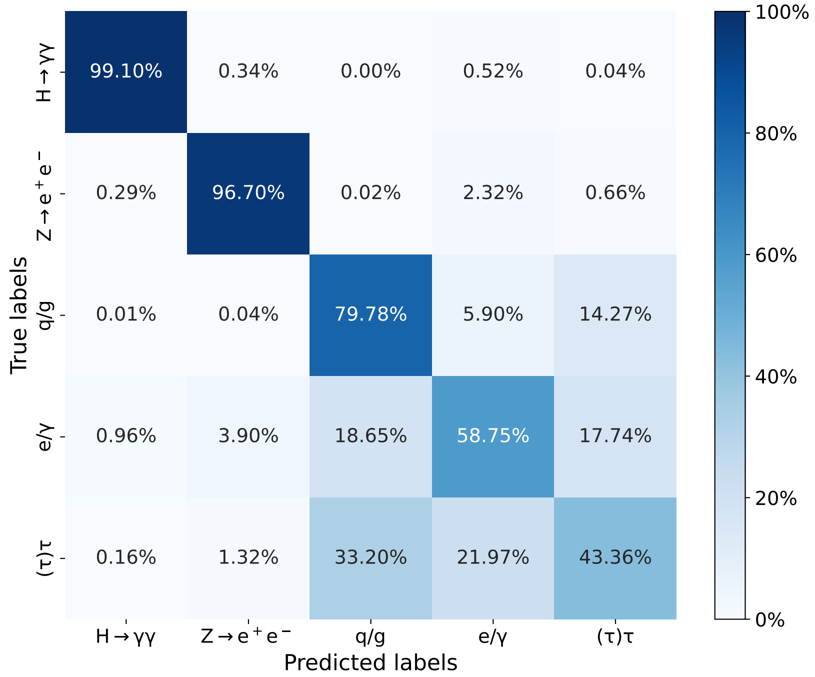


Figure E.3: Confusion matrix of the $H \rightarrow \gamma\gamma$ decays with the three background classes. There is a high degree of separation between the signal and all background jets. The added q/g-jets are highly separated from the signal jets, with the DNN never predicting q/g when the true label is the signal. It also only predicts the signal 0.01% of the time when the true class is q/g.

The confusion matrix for the five-class DNN jet tagger is displayed in Figure E.3. The signal accuracy is slightly better ($+0.01\%$) for the five-class model than for the four-class model. Adding the q/g-jets has almost no impact on signal identification as the confusion with q/g-jets when the true label is the signal is zero, and the confusion with the signal when the true label is q/g is 0.01%. This is the lowest level of confusion between the signal and any background jet. There is also a very meagre improvement in $Z \rightarrow e^+e^-$ classification, and there is also hardly any confusion between the Z boson jets and the q/g-jets. The most noticeable effect of adding the q/g-jets is reducing the proportion of predictions for the $(\tau)\tau$-jets. This occurs to a high degree when the true labels are both e/γ and $(\tau)\tau$; this indicates that the q/g and $(\tau)\tau$ classes share similar characteristics as measured by the ATLAS detector in these features.

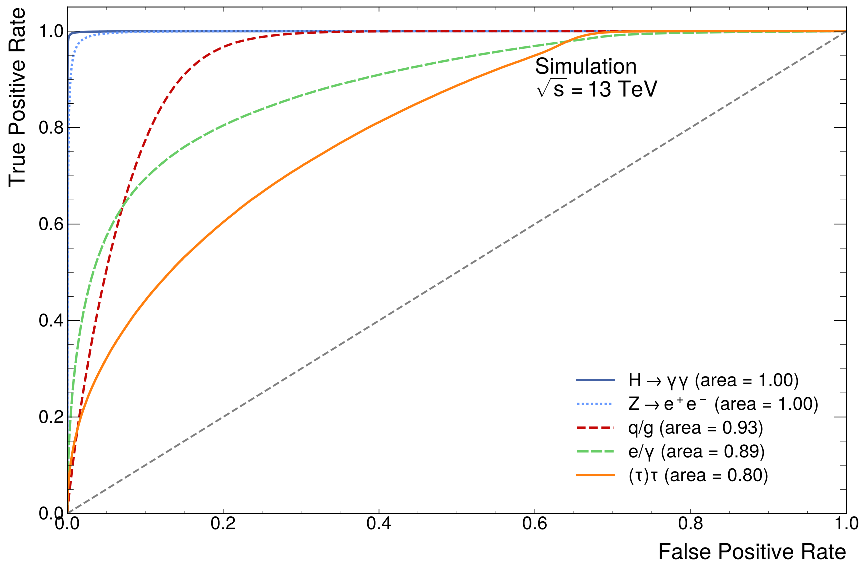


Figure E.4: Five-class DNN ROC curves for signal and background jets. AUC values for each jet are also displayed in the figure. The AUC values are high for all jet classes; the signal jet has the highest AUC value and has essentially achieved the possible maximum. There has been a significant decrease in the AUC score of the e/γ and $(\tau)\tau$-jets and the ROC curves for these classes are visually different to what has been observed in the four-class ROC curves.

Figure E.4 contains the ROC curves and AUC scores for the DNN five-class jet tagger. The $H \to \gamma\gamma$ and $Z \to e^+e^-$ decays possess textbook perfect ROC curves and maximum AUC scores. The new q/g ROC curve and AUC score are a shape and value for the e/γ and $(\tau)\tau$-jets for the four-class DNN jet tagger. Interestingly, the ROC curves are far from perfect curves for the e/γ and $(\tau)\tau$-jets. The AUC score is massively reduced for the $(\tau)\tau$-jets to 0.80. This is likely due to the confusion between the q/g and $(\tau)\tau$-jets seen in the confusion matrix in Figure E.3.

Both of the scalar discriminants, $D_{H\gamma\gamma}$ displayed in Figure E.5 and D_{Zee} seen in Figure E.6 are similar to the same distributions for the four-class DNN jet tagger. There is, however, somewhat less separation between the signal and the $Z \to e^+e^-$ decays. Suggesting that adding another jet class makes it more difficult for the DNN to learn to separate these two similar heavy boson jets as it is learning to separate an additional jet class.

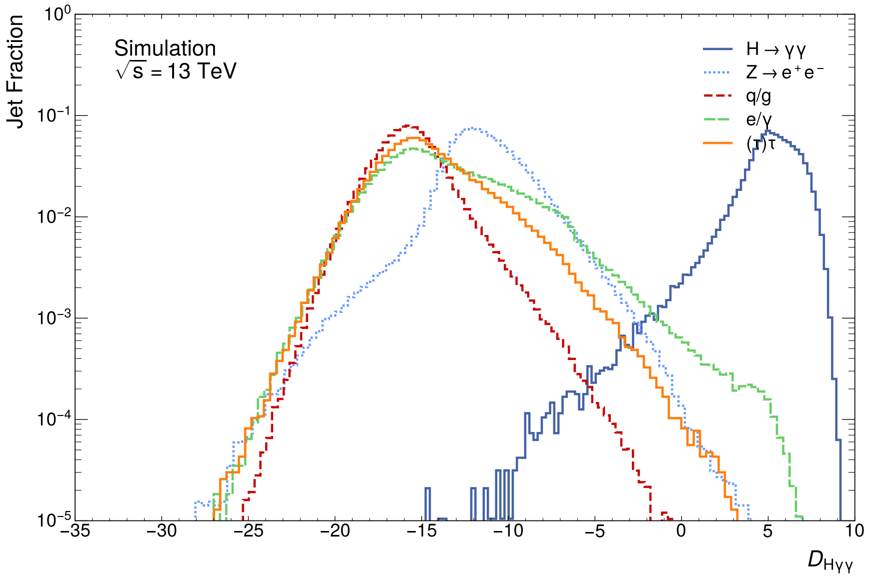


Figure E.5: Distribution of the scalar discriminant $D_{H\gamma\gamma}$ score for all jet classes Clearly visible is a separation between the signal jet $D_{H\gamma\gamma}$ score distribution and that of the background jets.

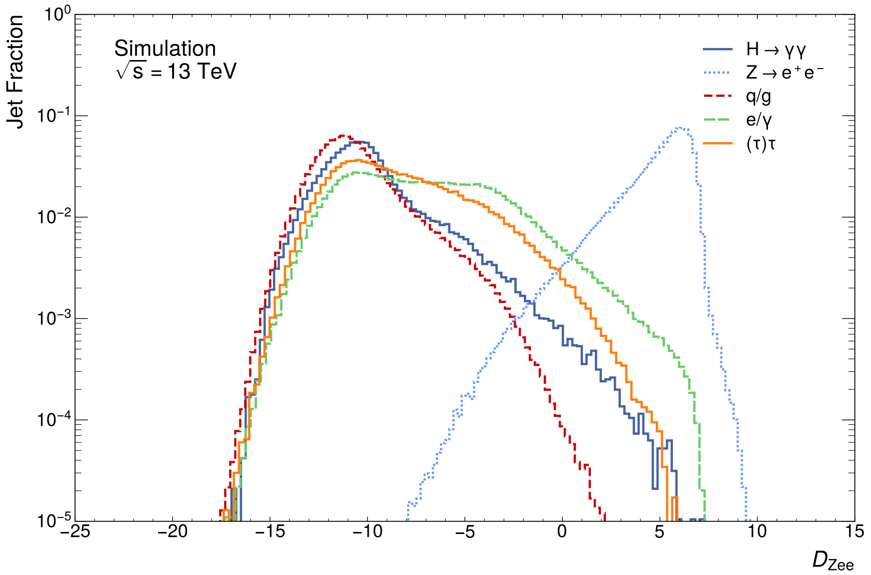


Figure E.6: Distribution of the scalar discriminant D_{Zee} score for all jet classes Clearly visible is a separation between the signal jet D_{Zee} score distribution and that of the background jets.

Table E.1 contains the background jet rejection rates at the 95% operating point for the DNN jet tagger. Compared to the rejection rates seen in table 3.4 for the four-class DNN jet tagger, there is some degradation in performance for $z \to e^+e^-$, e/γ and $(\tau)\tau$ rejection rates, although this is fairly minimal. When the q/g-jet rejection rate is compared to the rate in the related work for a $z \to e^+e^-$ jet tagger at this operating point, which is 88613, the five-class DNN jet tagger is two orders of magnitude better at rejecting these jets. There is little confusion, practically none, at many signal efficiencies between the signal and the q/g-jets. This is evidenced by the near zero confusion seen in Figure E.3 and that the value of the q/g-jet rejection rate is being determined by the offset value in Figure E.1.

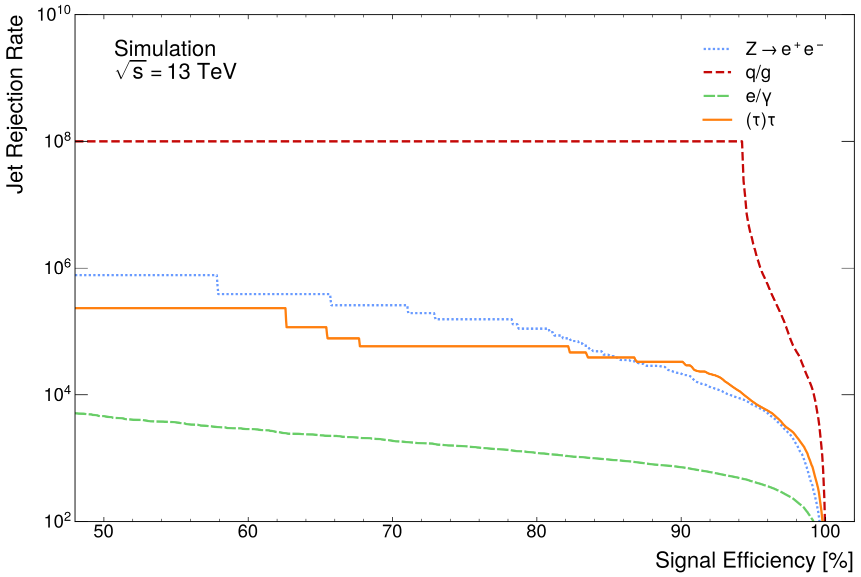


Figure E.7: Jet rejection rates of the background classes as a function of the signal efficiency. 10^8 is the offset value used to avoid division by zero; this indicates very high separation between the signal jets and the q/g-jets. Thus, the false positive rate is non-zero for a small range of the signal efficiency. This is supported by the very low confusion between these jets in Figure E.3.

E.2 ANN

The ANN jet tagger is also trained for 50 epochs and undergoes early stopping epoch 17. Figure E.1 displays the ANN training curves; the ANN classifier is visibly having

Table E.1: Comparison of rejection rates for the DNN and ANN jet taggers at the $\varepsilon_{H\to\gamma\gamma} = 95\%$ operating point for different background jets. q/g-jets have been added to these five-class jet taggers. For the DNN there is a limited decrease in the rejection rates of the other background jets. The q/g-jet rejection rate is 10^3 times greater than the highest other background jet rejection rate. For the ANN, the jet rejection rate has decreased to, in one case, the minimum possible positive value at the operating point.

	Rejection Rates at $\varepsilon_{H\to\gamma\gamma} = 95\%$	
Background Jet	DNN	ANN
$Z \to e^+e^-$	6877	1
q/g	2418936	5
e/γ	426	2
$(\tau)\tau$	7761	2

more difficulty minimising the cross-entropy loss. Fluctuations in the signal accuracy as a function of the epochs are also greater than in other networks developed in this thesis. The loss and signal accuracy difference between the first and final epoch is small, indicating reduced learning compared to other models.

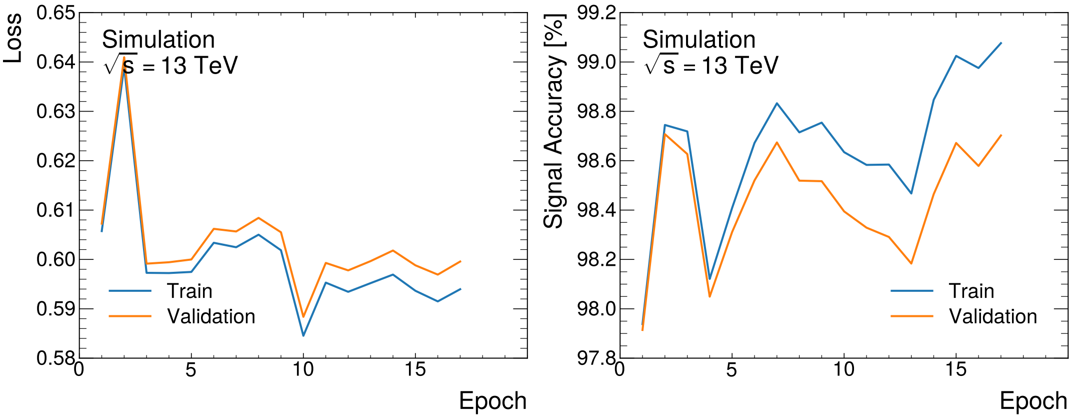


Figure E.8: Training curves for the five-class ANN jet tagging algorithm. These training curves are much flatter over the training epochs. This indicates that there is limited improvement in performance as the epochs progress.

The signal efficiency as a function of the transverse momentum and the absolute pseudorapidity bins, as seen in Figure E.9, is significantly degraded compared to all other developed algorithms. Entirely different distributions are seen; both plots have many points with the signal efficiency near zero. This is corroborated by the confusion matrix in Figure E.10. The percentage of true positive classifications for $H \rightarrow \gamma\gamma$ decays in the confusion matrix is a third below the values calculated for all other jet tagging algorithms. For the first time, the percentage for true positive classifications is not the greatest for the signal. It is the q/g-jets that has the greatest value, 2.83% greater than the value for the signal. Interestingly, there is still little confusion between the signal decays and the q/g-jets. The greatest increase in confusion with for prediction when the true class is $H \rightarrow \gamma\gamma$ is for the e/γ class, which makes up the bulk of the confusion, with some additional confusion occurring with the $Z \rightarrow e^+e^-$ decays. When the true class is the $Z \rightarrow e^+e^-$ decays, the predictions for the signal class are now greater than the true class's predictions. There is also a nearly equal number of predictions for the e/γ-jets. Generally, there is now much greater confusion between the classes with photons and electrons as the ATLAS detected particles. As ATLAS processes photons and electrons using the same components concurrently, this is a likely scenario that previous jet tagging algorithms have avoided.

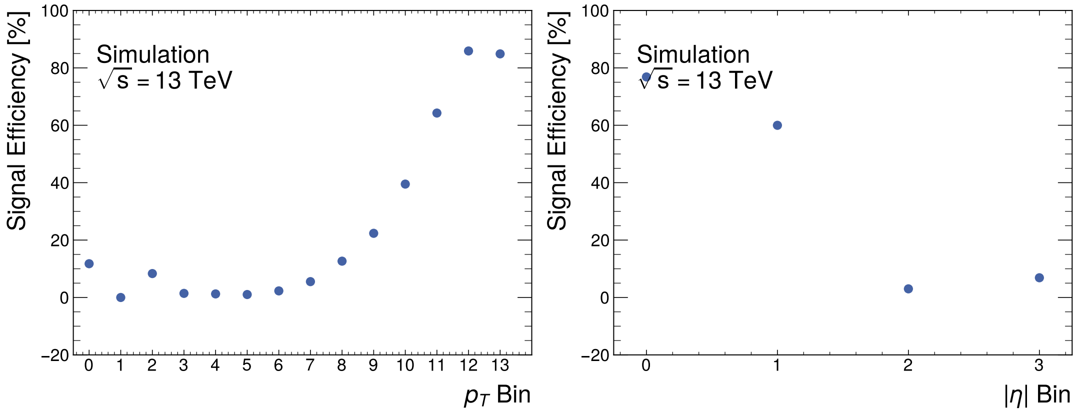


Figure E.9: Signal efficiency as a function of binned transverse momentum (left) and binned absolute pseudorapidity (right) for the five-class ANN jet tagger. The shape of the distributions of the signal efficiency is different to all other networks. There is a high proportion of bins for each plot for which the signal efficiency is near zero. This model has a performance that is much worse than previous jet taggers.

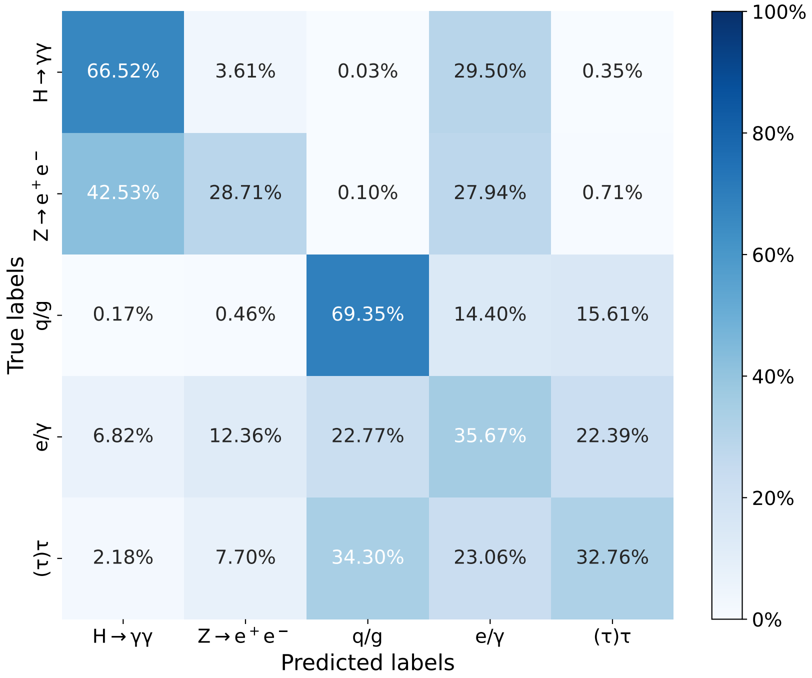


Figure E.10: Confusion matrix for the five-class ANN jet tagger. There is far more confusion between classes for this classifier than for all previously developed classifiers. The true positive classifications for the $H \to \gamma\gamma$ for the first time do not have the highest percentage value.

Figure E.11 presents the ROC curves and AUC scores; the curves are quite different, with no curves exhibiting the shape of a near-perfect classifier. The AUC scores have decreased across the board; for the first time, the AUC score for identifying $H \to \gamma\gamma$ decays is not the highest value at 0.86. The AUC scores for classifying $Z \to e^+e^-$ decays and q/g-jets now have the greatest scores at 0.90. The ROC curves and AUC scores show decreased performance for e/γ and $(\tau)\tau$-jets.

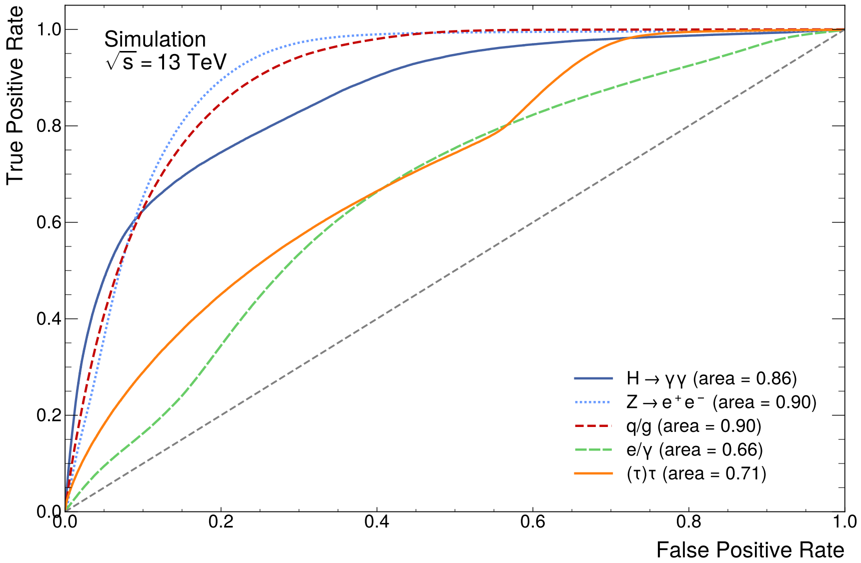


Figure E.11: ROC curves and AUC scores for the ANN jet tagger. No ROC curves exhibit the near-perfect shape observed for $H \to \gamma\gamma$ and $Z \to e^+e^-$ decays in previous figures. For the first time, the AUC score for the $H \to \gamma\gamma$ decays is not the highest value. All AUC scores are reduced compared to values for all other jet taggers.

The two scalar discriminants, the distributions of which are seen in Figure E.12 for $D_{H\gamma\gamma}$ and in Figure E.13 for D_{Zee}. Both distributions show far less separation between the signals and the backgrounds. Both distributions show far less separation than for previous scalar discriminant distribution plots. In particular, the $H \to \gamma\gamma$ and $Z \to e^+e^-$ decays are no longer separated. As previously seen in the confusion matrix in Figure E.10, this five-class ANN classifier performs poorly at separating these two decays, unlike all previous classifiers.

Different shapes are observed in the jet rejection rate as a function of signal efficiency displayed in Figure E.14. The curves are much smoother, indicating more jet rejection and confusion at lower signal efficiencies than seen for the other jet taggers. The jet rejection rate values at the 95% operating point tabulated in Table E.1 support this. At the operating point, all rejection rates are at the lowest possible order, with the rejection rate for $Z \to e^+e^-$ decays being at the lowest possible nonzero value.

From the evidence in this section for the five-class ANN jet tagger, simply using the found hyperparameter values for a four-class ANN jet tagger and using them to train a five-class jet tagger has yet to lead to a functional jet tagger. The DNN jet tagger is much more robust to the choice of hyperparameters, as the four-class DNN hyperparameters were sufficient to train a high-performing five-class jet tagger. This is likely due to the difference in complexity between the DNN and ANN training setups; the ANN is much more sensitive to the choice of hyperparameters. To develop a five-class ANN jet tagger, hyperparameter optimisation must be conducted again, making the ANN jet tagger a much more expensive multiclass jet tagger to develop.

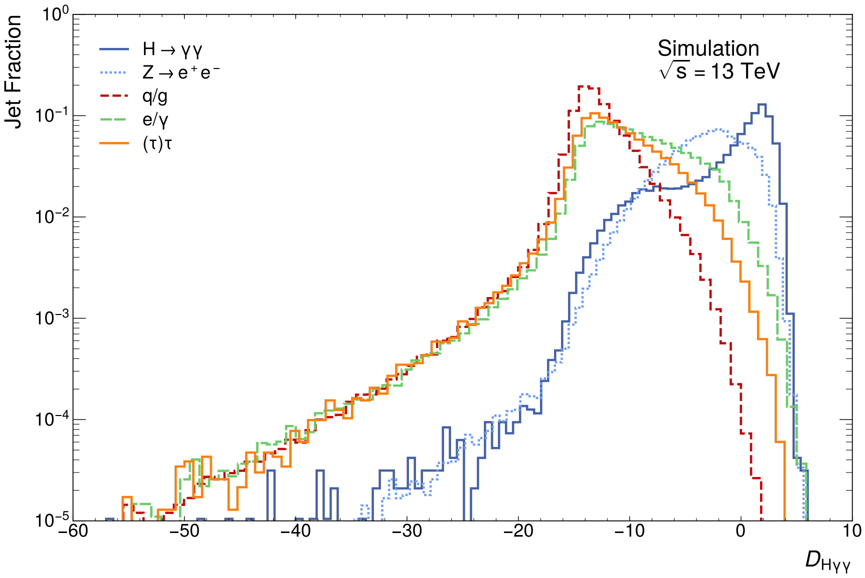


Figure E.12: Distribution of the scalar discriminant $D_{H\gamma\gamma}$ for the five-class ANN jet tagger. There is no clear separation between the $H \rightarrow \gamma\gamma$ and $Z \rightarrow e^+e^-$ decays. The separation between the signal and the other backgrounds is also reduced.

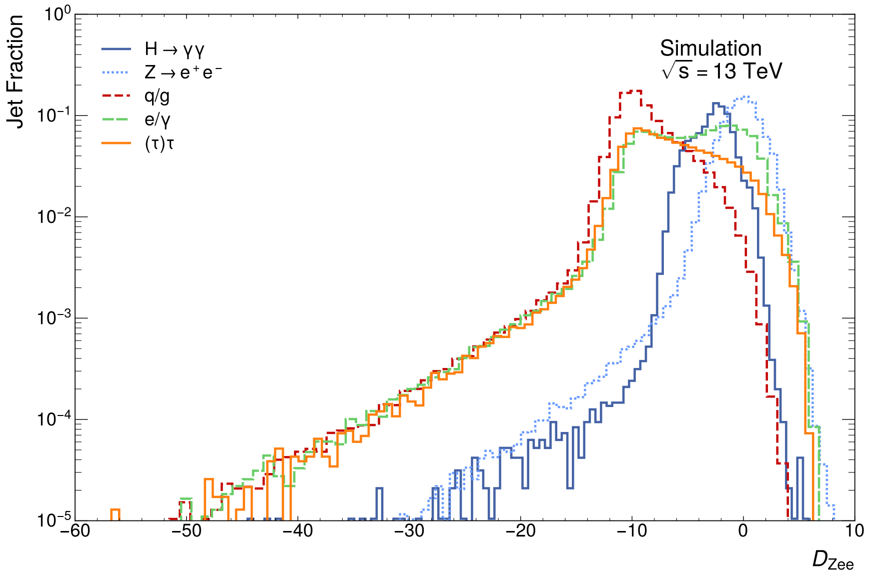


Figure E.13: Distribution of the scalar discriminant D_{Zee} for the five-class ANN jet tagger. The separation between $Z \rightarrow e^+e^-$ decays and the other classes is even less than observed in Figure E.10.

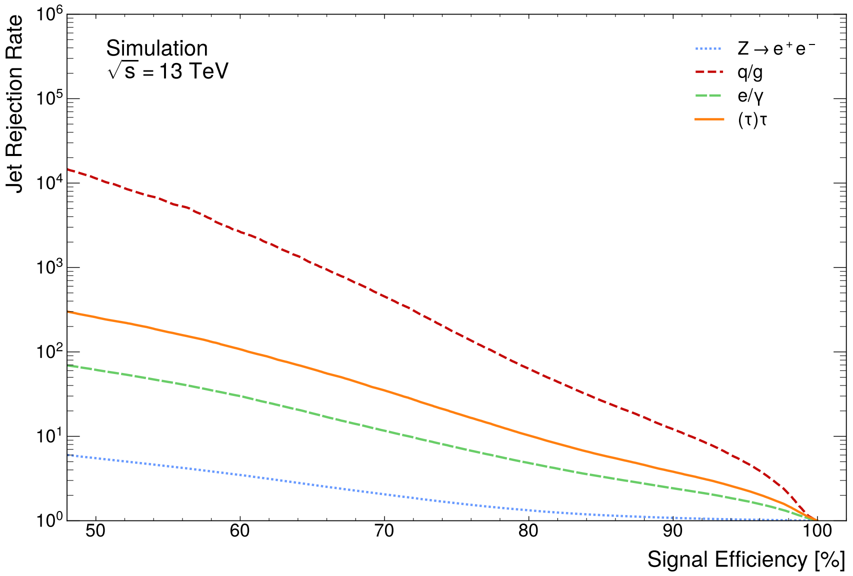


Figure E.14: Jet rejection rate as a function of signal efficiency for the five-class ANN jet tagger. Jet rejection rates at all signal efficiencies are much lower. The jet rejection rate curve for $Z \to e^+e^-$ decays has the lowest jet rejection rate at all signal efficiencies. For all previous jet taggers, this curve has been very similar to the jet rejection rate for $(\tau)\tau$-jets, so the ability of the jet tagger to reduce $Z \to e^+e^-$ backgrounds has degraded the most.