GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Fakultät für
Physik

Bachelor's Thesis

Klassifizierung von Signal und Untergrund in Verbindung mit $t\bar{t}H$ Produktionen mittels eines künstlichen neuronalen Netzwerks.

Classification of signal and background in associated $t\bar{t}H$ production via a multi-class neural network.

prepared by

Konrad Helms

from Osnabrück

at the II. Physikalischen Institut

Thesis number:	II.Physik-UniGö-BSc-2021/05
Thesis period:	5th April 2021 until 12th July 2021
First referee:	Prof. Dr. Arnulf Quadt
Second referee:	Prof. Dr. Stan Lai

Abstract

In dieser Bachelorarbeit werden die Klassifizierungsfähigkeiten von künstlichen neuronalen Netzwerken anhand des Produktionsmechanismus eines Top-Antitop Quarkpaares in Kombination mit einem Higgs-Boson untersucht. Es wird gezeigt, dass künstliche neuronale Netzwerke dazu in der Lage sind, diese Produktionsmechanismen besser zu identifizieren, als die bisher genutzten gewichteten Entscheidungsbäume. Die Klassifizierungsleistung ist um etwa 1% verbessert.

Des Weiteren werden die künstlichen neuronalen Netzwerke erweitert und zur weiteren Klassifizierung der auftretenden Untergrundprozesse von Top-Antitop-Paaren in Assoziation mit weiteren Jets genutzt. Die Untergrundprozesse werden anhand der darin enthaltenen leichten, Charm-, Bottom- oder Top-Quarkjets klassifiziert. Hier wird beobachtet, dass diese Netzwerke nicht in der Lage sind, anhand der gegebenen Daten eine zuverlässige Klassifizierung der Prozesse zu ermöglichen. Dies wird auf die beschränkte Auswahl an Eingangsvariablen, welche ursprünglich zur binären Klassifizierung des Produktionsmechanismus produziert wurden, zurückgeführt.

Stichwörter: Physik, Teilchenphysik, Top Quark, Higgs Boson, Maschinelles Lernen, künstliches neuronales Netzwerk

Abstract

In this Bachelor's thesis, the classification performance of artificial neural networks in the context of Higgs-associated top anti-top quark pair production is analysed. Binary classification neural networks are employed to separate the Higgs-associated top anti-top quark pair production mode from the top anti-top and jets background events. When compared to boosted decision trees used in previous analyses, a performance increase of about 1% is observed.

Multi-class neural networks are then used to split up the background further. A classification is made based on the light, charm, bottom or top quark jets within the background events. Here, it is observed that the neural networks are not able to improve classification performance significantly compared to a random classification. This is caused by the biased selection of input variables, which was initially designed for the binary classification of Higgs-associated top anti-top quark pair processes and background events.

Keywords: Physics, Particle Physics, Top Quark, Higgs Boson, Machine Learning, Neural Network

Contents

1. Introduction

Particle physics is a fundamental field of physics investigating the constituents of our universe. Currently, it uses the Standard Model as a quantum field theory to describe the interacting particles as well as their interactions with each other.

The rapid technological advances in recent years allowed for testing the theoretical predictions of the Standard Model in the high energy regime. As a result of the ever increasing energies that can be used in particle accelerators, the top quark was discovered by the DØ and CDF collaborations in 1995 [1, 2]. Then, in 2012, the Higgs boson was discovered by the ATLAS and CMS collaborations [3, 4]. The Higgs boson is a fundamental piece in the Standard Model and is part of the Higgs mechanism giving rise to the masses of the fermions and gauge bosons in the Standard Model [5–7].

In 2018, the Higgs boson coupling to the top quark was probed by observing Higgs-associated top anti-top quark pair production $t\bar{t}H$ with ATLAS and CMS [8, 9]. Measuring the top-Higgs Yukawa coupling can be a precise test of the Standard Model and finding a disagreement with theory could indicate new, yet unknown physics.

To identify the $t\bar{t}H$ production mode in the recorded data, machine learning and event simulation are used. In previous analyses, boosted decision trees were employed to identify the production mode. As technology advances, neural networks become increasingly more powerful and focus shifts to implementing the more sophisticated neural networks in upcoming analyses instead of boosted decision trees.

In this Bachelor's thesis, the underlying Standard Model physics and neural networks are introduced in Sections 2 and 3, respectively. Then, previous analyses results and the classification performance of binary and multi-class neural networks in $t\bar{t}H$ production modes are presented in Section 5. Furthermore, the ATLAS detector is presented in Section 4. The results are analysed and compared to the ones of the previously used boosted decision trees in Section 6. A conclusion of the presented results is given in Section 7.

2. The Standard Model of Particle Physics

The Standard Model of Particle Physics (SM) is, as of now, the most precise theory in science, with observables measured to a relative precision of up to $\sim 10^{-13}$. It describes three of the four known fundamental forces in the universe [10–13]. With underlying relativistic quantum field theories it delivers a theoretical framework to describe the fundamental interactions of particles as an exchange of gauge bosons.

2.1. Elementary Particles and their Interactions

The SM consists of fermions and the force-carrying gauge bosons. Figure 2.1 shows the gauge bosons and the three generations of fermions, categorised as quarks and leptons. Quarks, carrying a fraction of the elementary charge e, can interact with the strong force, whereas leptons, carrying integer numbers of the elementary charge e, can not interact with the strong force. The generations separate the fermions roughly by mass. The quarks are then further divided into two types of quarks. The up-type quarks carry an electric charge of $+\frac{2}{3}e$, the down-type quarks carry an electric charge of $-\frac{1}{3}e$. The leptons are separated by electric charge, too, neutrinos ν are not charged whereas the electron e, muon μ and tau τ are negatively charged, carrying a charge of $-e$.

The gluon g is the gauge boson of the strong interaction, the photon γ is used to mediate the electromagnetic (EM) interaction, and the electrically charged $W^{\pm}$ bosons and neutral Z boson mediate the weak interaction. The gravitational force, modelled via the exchange of gravitons, is currently not part of the SM. Often, the EM and weak interaction are described as a unified electroweak interaction.

Each fermion in the SM, except for the neutrinos, has one corresponding anti-particle with the same mass, lifetime and spin, but opposite additive quantum number such as lepton number, baryon number, and electric charge. In the seesaw mechanism, an extension to

the SM, it is required, that a neutrino is its own anti-particle [14].

While the quarks can interact via all three fundamental forces of the SM, the electrically charged leptons only couple via the EM and weak interactions. Neutrinos only interact via the weak interaction.

An additional quantum field, the Higgs field, was added to the SM to explain the masses of the gauge bosons of the weak interaction, as these bosons a priori are not massive in the gauge invariant SM. In addition to that, the Higgs mechanism gives rise to the masses of the fermions of the SM via Yukawa couplings. The interaction strength with the Higgs boson H is proportional to the fermion's mass. Unlike the spin-$\frac{1}{2}$ fermions and other gauge bosons, which are vector bosons having a non-zero spin, the Higgs boson is a scalar boson and can be seen as an excitation of the Higgs field.

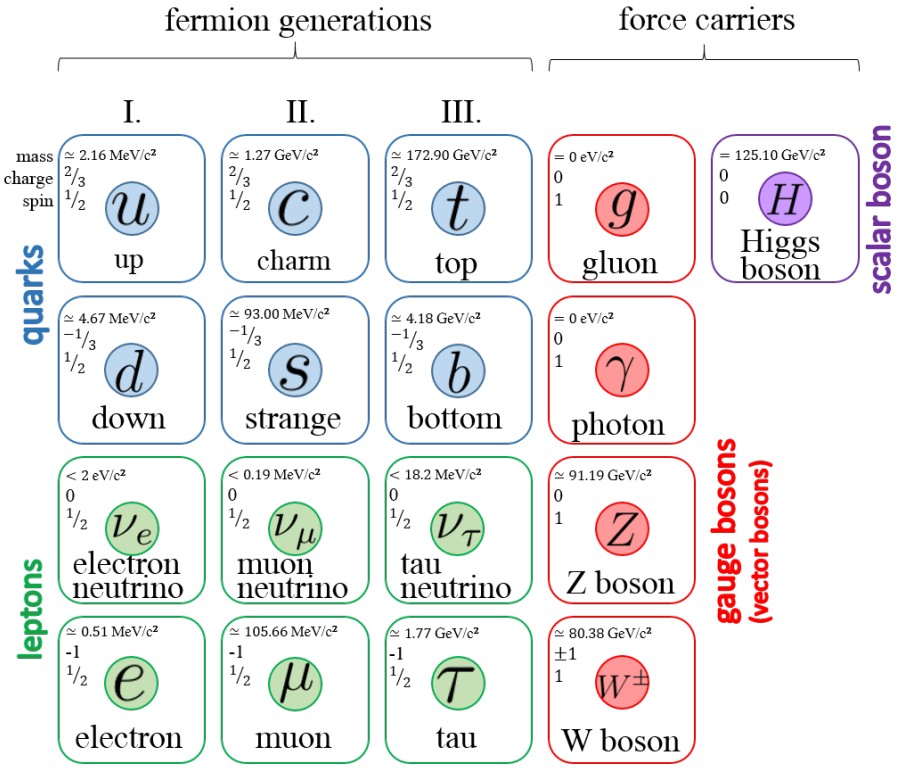

Figure 2.1.: Standard Model particles with their mass, electric charge, and spin [15].

2.2. Limitations of the Standard Model

Although most of the predictions of the SM match results from experiments within measurement and calculation uncertainties, the SM is not a complete description of our uni-

verse. It does not explain the gravitational force, the existence of dark matter, and the large discrepancies between energy scales of many fundamental forces and the electroweak force, known as the hierarchy problem, thus hinting at physics beyond the standard model (BSM) [16–18]. There are several approaches on BSM theories. Supersymmetry (SUSY) theories predict supersymmetric particles, the sparticles, which differ by half an integer in spin to the known SM particles. These sparticles could be a solution to the hierarchy and dark matter problems [19, 20]. Other more conceptional approaches to explain the phenomena are string theories and grand unified theories.

2.3. $t\bar{t}H$ Production Mode

In 2012, the Higgs boson H was discovered by the ATLAS and CMS collaborations [3, 4]. Since then, there have been several measurements of the Higgs boson's properties to find constraints on the couplings. The Higgs-associated top anti-top quark pair production mode, $t\bar{t}H$ was first observed in 2018 by the ATLAS and CMS collaborations [8, 9]. Three tree level Feynman diagrams for the $t\bar{t}H$ production mode can be seen in Figure 2.2.

In the past, the Higgs coupling to other particles has been measured via the Higgs decay into a particle anti-particle pair, for example the decay to a bottom anti-bottom quark pair, $H \to b\bar{b}$, and the decay to a positively and a negatively charged W boson pair, $H \to W^+W^-$ [21–23]. The top quark's coupling to the Higgs boson is the largest Yukawa coupling. Measuring the Higgs decay into two top quarks is not possible as the decay does not exist due to energy conservation.

Another way of measuring the interaction strength of the top quark t and Higgs boson H is the $t\bar{t}H$ production. It is the most favourable production mode for a direct measurement of the coupling [24–27]. This coupling is not only sensitive to the properties of the Higgs boson H, but also to possible yet undiscovered particles [28]. Current measurements show a top-Higgs interaction $\approx 25\,\%$ stronger than the theoretical prediction, but with $\approx 25\,\%$ measurement uncertainties [8, 9]. Reducing systematic uncertainties could test the agreement between the experiment and the theory and potentially indicate the discovery of new physics.

This Bachelor's thesis on the separation of the $t\bar{t}H$ and tH production modes and classification of background events provides an approach to improve the $t\bar{t}H$ classification mechanisms. The $t\bar{t}H$ production mode is sensitive to the absolute value of the top-Higgs coupling, whereas the tH production mode is sensitive to the sign of the top-Higgs cou-

pling. Thus, it is desirable to separate both production modes to find constraints on this Yukawa coupling.

Figure 2.2 shows three tree level Feynman diagrams for possible $pp \to t\bar{t}H$ production modes, which can not be observed separately as they interfere with each other. The proton p is made up of two up u and one down d quarks, the valence quarks, and other virtual quark anti-quark pairs resulting from gluon splitting, the sea quarks. The left diagram shows the associated $t\bar{t}H$ production via the t-channel. The other Feynman diagrams in Figure 2.2 show Higgs boson radiation of one of the final state top quarks in either a gluon-gluon gg fusion or quark anti-quark $q\bar{q}$ annihilation.

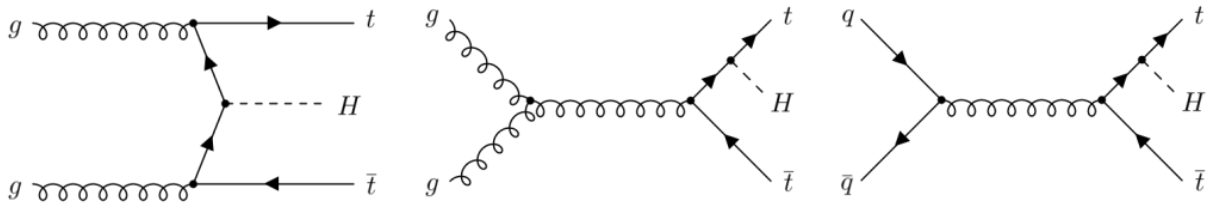

Figure 2.2.: Three possible tree level Feynman diagrams in proton-proton collisions for the $t\bar{t}H$ production mode.

2.4. tH Production Mode

In addition to the $t\bar{t}H$ production mode, the Higgs-associated top production, tH, can occur in proton-proton collisions at the Large Hadron Collider, LHC.

Higgs boson production in association with a single top quark at leading order can be categorised into three groups based on the virtuality of the W boson: t-channel (space-like W boson) and s-channel (time-like W boson) production and the associated production with an on-shell W boson, as can be seen in Figure 2.3 and Figure 2.4.

The tH production with an on-shell W boson in the final state, Figure 2.4, is unlikely to come out of proton-proton collisions at the LHC as both production mechanisms start with an initial state bottom quark b which is suppressed at the LHC, indicated by the proton parton distribution function [29]. On the left in Figure 2.4, an initial state gluon g is captured by a bottom quark b which then decays weakly forming a top quark t and a W boson from which the Higgs boson H is emitted. On the right in Figure 2.4, an initial state gluon g interacts with a top quark t stemming from the weak decay of an initial state bottom quark b. This results in the desired top-Higgs pair tH in addition to an on shell W boson. The two processes interfere with each other.

In the t-channel in Figure 2.3, an arbitrary initial quark q interacts with a bottom quark b from an initial state gluon g via the weak interaction to form a top quark t, which radiates off a Higgs boson H. In the s-channel in Figure 2.3, a quark anti-quark pair $q\bar{q}'$ annihilates to form a virtual W boson, which produces the top-Higgs pair tH in association with an anti-bottom quark $\bar{b}$.

In comparison with $t\bar{t}H$ production mode, the tH production mode has a smaller cross section, but provides the opportunity to measure the sign of the top-Higgs coupling [15]. The Higgs boson H decays further blending in with background events as described in Section 2.5.

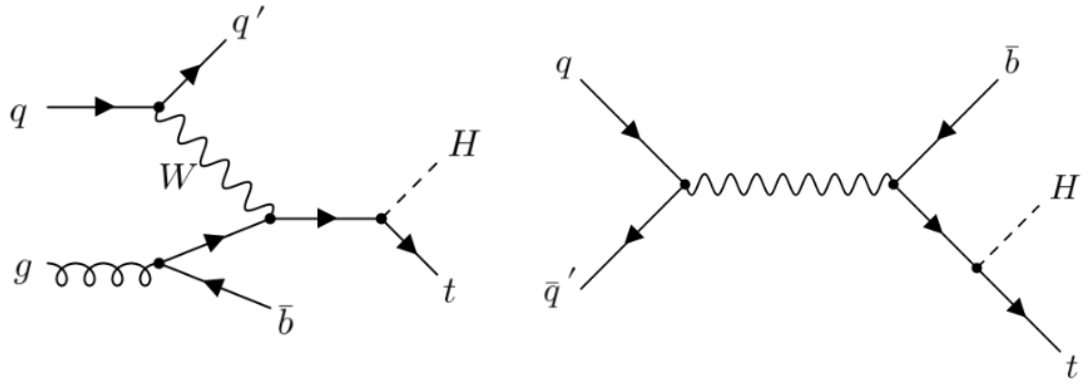

Figure 2.3.: Representative leading order Feynman diagrams for the tH production via the t- and s-channel.

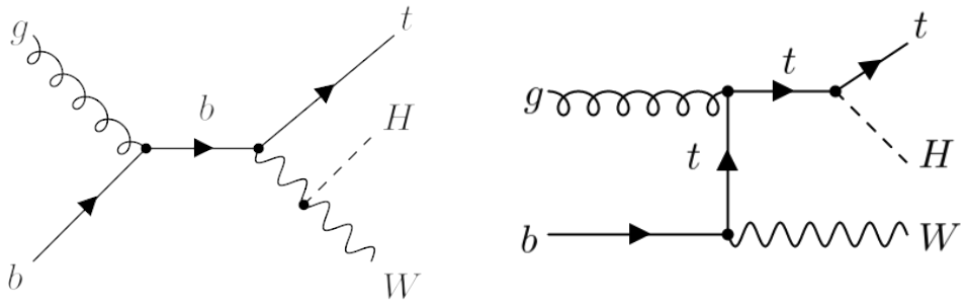

Figure 2.4.: Representative Feynman diagrams for the tH production with an on-shell W boson in the final state. Both channels can have Higgs radiation of the top quark t or W boson.

2.5. Background Processes and Jet Classification

The most prominent background processes to be considered are $t\bar{t}$ plus b-jet background processes, $t\bar{t}+b$. In addition to that, $t\bar{t}+$ light jets, and $t\bar{t}+c$ background processes are

also considered. Here, c indicates charm quark c-jets and light jets describe quark jets originating from the three light quarks: up u, down d, and the strange quark s, as their mass is significantly lower than the charm c, bottom b or top t mass, see Figure 2.1.

Feynman diagrams for possible $t\bar{t} + b\bar{b}$, $t\bar{t}$ + light jets and $t\bar{t} + c$ processes can be seen in Figure 2.5. Here, the quark anti-quark pairs are created via the interaction with a gluon g. Jet measurements performed at any particle detector are limited by detector coverage, pseudorapidity η and the transverse momentum p_{T} of the jets, see Section 4.3, thus not every bottom b-, charm c-, or light jet is detected or tagged properly. One can observe an odd number of jets. In practice, some b-jets might be mistagged and recognised as a different type of jet. Similarly, one might tag a non-b-jet as a b-jet. The b-jet identification, called b-tagging, is of particular interest for research on charge-conjugation-parity-symmetry (CP) violation and reconstruction of the Higgs boson H decays.

The Higgs boson H resulting from the $t\bar{t}H$ and tH production modes, see Section 2.3 and Section 2.4, has a broad spectrum of decay channels which depend on the Higgs boson's mass. At the measured Higgs mass $m_H \simeq 125\,\mathrm{GeV}$, the most dominant decay mode is the decay to a bottom anti-bottom quark pair $H \to b\bar{b}$ with a branching ratio of about 58% [30].

In addition to that, the top quark t decay is dominated by the two-body decay channel $t \to Wb$ with the W boson subsequently decaying further to a quark anti-quark pair $q\bar{q}'$ or a charged lepton and neutrino $\ell\nu$. The resulting jets create additional background in the $t\bar{t}H$ and tH production mode measurements.

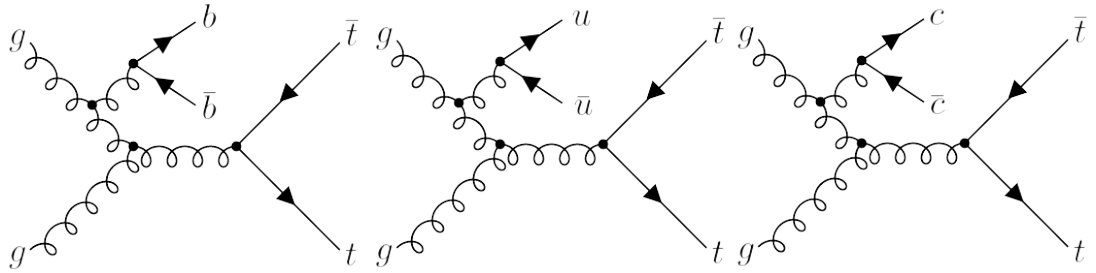

Figure 2.5.: Representative Feynman diagrams for the $t\bar{t} + b\bar{b}$, $t\bar{t}$ + light jets and $t\bar{t} + c$ background processes.

In the following, the $t\bar{t}H$ production mode will be referred to as ttH, and the considered background processes will be referred to as $tt + b$, tt + light jets, and $tt + c$.

3. Machine Learning and Neural Networks

The field of machine learning (ML), which can be considered a subfield of artificial intelligence, originated in the 1950s with Alan Turing proposing to consider the question "Can machines think?" [31].
Since the beginning of machine learning, numerous types of ML were invented, such as supervised, unsupervised, semi-supervised, and reinforcement learning. In this thesis, multi-class neural networks (NNs), which are part of the supervised learning category, are used for the classification of events into two, three, four or five classes, presented in Sections 2.3, 2.4 and 2.5. Each input event being classified, corresponds to four years of data collecting during Run II of the LHC. The event yields represent the statistical number of observed processes during one event. In Section 6.1, the performance of the NN is compared to that of boosted decision trees (BDTs) used in previous ATLAS analyses [32]. The multi-class NN can be thought of as a function, taking data as an input and giving a classification of the input data, corresponding to multiple output classes, as an output. An algorithm is trained to find the function that classifies ideally all of the input data correctly. In practice, 100% accuracy, meaning every input is classified correctly, will not likely be reached.

3.1. Structure of Neural Networks

The structures of NNs are, as the name suggests, inspired by the brain. Figure 3.1 shows an example of an NN structure. They are made up of neurons, referred to as nodes, and the connections of the nodes with each other. Nodes hold an activation a, which can be any real number, and are arranged in layers with each node in a layer being connected to each node in the previous and following layer. The nodes in the last layer, for probability normalisation reasons, hold an activation value a between zero and one.

In the first layer, the input layer, each node takes data as an input-activation. In the last layer, the output layer, each node corresponds to one output class. The output activations represent how much the NN "thinks" that the given input is the respective class. An activation equal to one in an output node means that "it is very certain" that the input belongs to this class. Descending values of activation, until zero, signify a decrease in the NN's certainty. The classes considered in this thesis are orthogonal to each other, so the NN is expected to yield an output layer with one node having an activation close to one and the other four having activations close to zero. The layers in between the first and last layer are referred to as hidden layers.

Analogous to the biological anatomy, neurons from different layers affect the ones in other layers. In artificial NNs, a weight w, which can be positive or negative, is assigned to each connection of nodes. Via the weight assignment, the amount of influence of nodes from one layer to the layer afterwards can be changed.

As a measure for the activation $a_n^{(j+1)}$ of node n in layer $j+1$, the weighted sum, of the activations $a_i^{(j)}$ from the nodes $i = 1, ..., N$ in the previous layer, layer j, can be calculated. Here, a bias b_n can be included to influence how large the weighted sum needs to be, before the node exceeds a certain activation level:

$$a_n^{(j+1)} = \sum_{i=1}^{N} w_{n,i}^{(j)} \cdot a_i^{(j)} + b_n^{(j)} = w_{n,1}^{(j)} a_1^{(j)} + w_{n,2}^{(j)} a_2^{(j)} + ... + w_{n,N}^{(j)} a_N^{(j)} + b_n^{(j)}. \tag{3.1}$$

$w_{n,i}^{(j)}$ refer to the weights, $i = 1, ..., N$, of the N connections from the nodes in layer j to the n-th node in layer $j+1$. Different biases $b_n^{(j)}$ can be chosen for the calculation of the weighted sum corresponding to different nodes n in different layers j. In practice, the activation of the node $a_n^{(j+1)}$ is calculated using Equation 3.1 and an activation function.

3.2. Activation Functions

The weighted sum of activations from the second to last layer in a NN, including a bias, can be any real number but needs to be compressed into the interval $[0, 1]$ for it to become a valid value for an activation of a node from the output layer m, as these are often interpreted as probabilities. This is done by activation functions. The activation functions applied on hidden layers do not need to compress the weighted sum into the

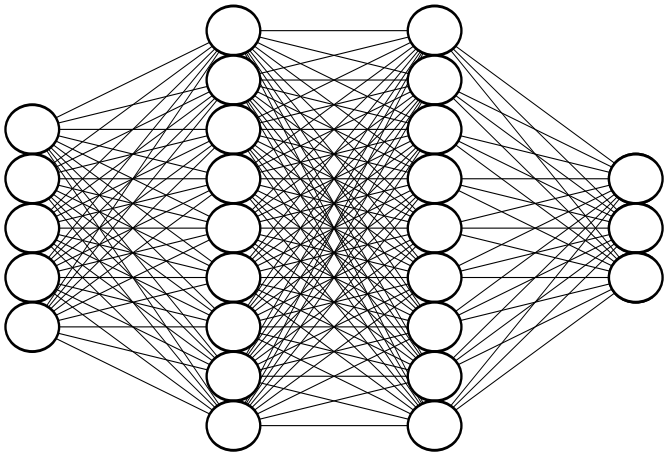

Figure 3.1.: Structure of a multi-class neural network with an input layer, two hidden layers and an output layer.

given interval. In this thesis, the sigmoid,

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.2}$$

and Softmax

$$\hat{\sigma}(\boldsymbol{a}^m)_n = \frac{e^{a_n}}{\sum_{i=1}^{K} e^{a_i}} \tag{3.3}$$

activation functions are used, each shown in Figure 3.2. Here, n corresponds to one of the K output classes of the NN. The Softmax activation function $\hat{\sigma}$ is appended to the output layer of multi-class NNs as this function is a multidimensional generalisation of the sigmoid function σ. The Softmax function takes a vector of activations from the last m-th layer

$$\boldsymbol{a}^m = \begin{bmatrix} a_1^m \\ a_2^m \\ \vdots \\ a_K^m \end{bmatrix}, \tag{3.4}$$

which is then normalised to a probability distribution with each entry in the resulting vector being in the interval $[0, 1]$. In this thesis, the multi-class NNs have up to five output probabilities, $K \leq 5$. For a binary-classification NN having only one node in the output layer, the sigmoid σ function in Equation 3.2, is appended to the weighted sum of

the N nodes from the last m-th layer

$$\sigma\left(\sum_{i=1}^{N} w_{1,i}^{(m)} \cdot a_i^{(m)} + b_1^{(m)}\right) = \sigma\left(w_{1,1}^{(m)} a_1^{(m)} + w_{1,2}^{(m)} a_2^{(m)} + ... + w_{1,N}^{(m)} a_N^{(m)} + b_1^{(m)}\right). \tag{3.5}$$

This yields a normalised binary classification output. Within the NN, the ReLU activation function, see Figure 3.2, defined as

$$R(x) = \max(0, x) \tag{3.6}$$

is applied to hidden layers. This function does not compress the input value, the value of the weighted sum, into the interval $[0, 1]$, but is used to introduce a non-linearity, such that the NN can not be collapsed to a single layer. Using ReLU can cause problems like the exploding gradient problem [33]. The NNs considered in this thesis are not deep enough, meaning that they do not have enough hidden layers, for this problem to occur.

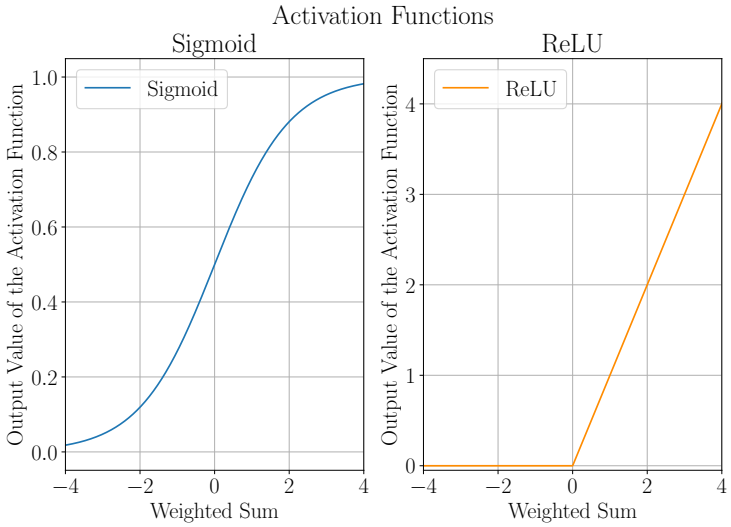

Figure 3.2.: The activation functions sigmoid and ReLU.

3.3. Cost Functions

The initially randomly initiated weights and biases in a NN need to be adjusted, so that the NN classifies its input correctly. To evaluate if the NN classifies its input in many training events correctly, a cost function is defined. This function takes all weights and biases of the NN as an input and outputs a positive real number, the cost. It is parameterised by the NN's behaviour over a number of training events.

The cost function is a measure of error between the correct classification of the input data and the classification of the NN over many classifications of events during training.

A loss function is used to evaluate a single training example. The categorical cross entropy loss in a single training example is defined as

$$\text{loss} = -\sum_{i=1}^{K} y_i \cdot \ln(a_i), \tag{3.7}$$

where y_i represents the activation of the i-th output node in a vector of a correct classification, so either zero or one for orthogonal output classes, and a_i is the activation of the i-th node in the NN's output layer. The activations in the output layer depend on all weights and biases of the NN. The cost function C is the average value of the loss function over a batch of training examples.

There are other methods of calculating the loss in a single training example, such as the mean squared error loss (MSE). Using the mean squared error loss in, for example, the cost function of a multi-class NN is not optimal, because the value of the cost is not directly linked to the classification error. A confidently wrong classified event should be penalised greater than an event that is vaguely wrong classified into more output classes. This is solved by using the categorical cross entropy loss [34]. The methods that are used to evaluate the training process are called metrics. Another example for a metric is the accuracy, defined in Section 3.7

3.4. Learning Process and Gradient Descent

The learning process of a NN consists of minimising the cost function, which, in this thesis, has a number of inputs in the order of $\sim 10^3$.

Determining the global minimum of the cost function would result in the best performance of the NN, but can be more CPU-intensive than approaching a local minimum. The direction of the local minimum of the cost function can be approximated using the negative gradient $-\nabla C$ of the cost function. The negative gradient of the cost function $-\nabla C$ points in the direction of the steepest descent of the cost function with respect to its input parameters. The algorithm for finding minima of the cost function is referred to as gradient descent.

First, the negative gradient of the cost function $-\nabla C$ is computed. Then, in the high dimensional space of the cost function, a small step is taken in the direction of the negative

gradient $-\nabla C$. This process is repeated until a local minimum is reached. The algorithm for computing the negative gradient of the cost function $-\nabla C$ efficiently is referred to as backpropagation, see Section 3.5. The amount of change to the NN during each step is called learning rate.

In practice the gradient descent algorithm is heavily optimised to reduce CPU usage. Steps are taken proportional to the absolute value of the gradient of the cost function to prevent the NN from passing by a local minimum, and not every component of the gradient vector is considered when taking a step. Only the ones that decrease the value of the cost function the most are taken into account. In this thesis, the stochastic gradient descent optimiser "Adam", provided as a class in `keras`[1], is used. Adam is well suited for problems that are large in terms of data and optimisable parameters [36]. In stochastic gradient descent methods, the training dataset is divided into batches and, using backpropagation, a gradient is calculated for each batch. This gradient is an approximation to the actual gradient of the function computed in a less CPU-intensive way.

3.5. Backpropagation

Backpropagation and its usability in the field of ML was introduced in 1986 and is a tool for calculating the gradient of the cost function ∇C in an efficient way [37]. Unlike in the naive calculation of a gradient vector with respect to each weight and bias individually, in backpropagation this is done via the mathematical chain rule.

Each activation value of a node is the output result of many activation functions chained together, all having different inputs. In the backpropagation algorithm, the gradient for a fixed input-target-output example is calculated with respect to the weights and biases via the chain rule. This results in a list of nudges, namely the values of the gradient vector, that should be applied to the weights and biases according to the one fixed input-target-output example. This is done for many different training input-target-output examples. Then these nudges, the components of the gradient vector, are averaged to yield an approximation for a gradient of the cost function. Here, to save computational expense, the stochastic gradient method can be used.

The calculations are carried out in an efficient way by avoiding calculating duplicates for example duplicate derivatives, by computing the gradient of the weighted input to each layer at a time and iterating backwards from the last layer to the first.

[1]Keras is software library that provides an interface for the TensorFlow library for artificial neural networks [35]

3.6. Hyperparameters of Neural Networks

In ML, hyperparameters are used to control the learning process.

In addition to the number of nodes per layer and number of hidden layers, other hyperparameters can be introduced. In the evaluation process, different behaviours, such as under- and overfitting, are analysed and hyperparameters can then be optimised for the NN to perform best on the given input datasets.

Setting the learning rate, introduced in Section 3.4, influences how quickly a neural network is able to learn, in other words, how quickly it updates its weights and biases. Low learning rates slow down the learning process, which then may converge smoothly towards the minimum of the cost function, whereas high learning rates result in faster learning of the NN but can overshoot a (local) minimum of the cost function, thus resulting in a learning process that does not converge.

The patience of a model can be seen as a delay for the early stopping mechanism, see Section 3.7. It can be used to prevent the model from over- or underfitting the training dataset. Overfitting is referred to, if a NN learns the structure of the training data in too much detail, for example learns the noise in training data, thus performing worse in the generalisation of the data. Underfitting is referred to, if a NN is not able to learn the underlying structure of the training data nor generalise to new datasets due to a lacking flexibility in the model. Since the training of a NN using, for example, a stochastic gradient descent algorithm can be noisy, the performance on the validation dataset may fluctuate. The patience is introduced to the model to not stop the training immediately if its performance on the validation dataset decreases slightly, as it may increase again. Performance improvements have a lower bound, $\Delta_{\text{min.}}$. It is referred to as the minimum change in a metric that qualifies as an improvement.

Deep learning NNs, networks that have a number of hidden layers, tend to overfit the datasets, but not so ensembles of NNs with different hyperparameter configurations. The downside of running ensembles of different NNs is that this requires additional computational resources. Using only a single NN, running a large number of different NN architectures can be simulated by randomly dropping out nodes during the training process. The layer number, at which dropout layers are added, and dropout probability of the nodes can be optimised to improve the generalisation capabilities of NNs.

Underfitting a dataset can be prevented by passing the same dataset through the network multiple times. Feeding the dataset forwards and backwards through the NN once is called one epoch. During multiple epochs, the NN can improve in fitting the dataset, i.e.

approaching smaller values of the cost function.

3.7. Evaluation of Training

To evaluate a NN's performance, its accuracy, classification outputs, receiver operating characteristic (ROC) curve, loss, and the relevant statistical uncertainties need to be taken into account. These evaluation parameters are defined in the following section.

The available datasets are divided up into two subsets, the training and testing dataset. The NN is trained using the training dataset and its performance is then tested using the testing dataset. In this thesis, k-fold cross validation with either $k = 5$ or $k = 4$ folds is used. Here, the NNs are trained on a fraction of the Monte Carlo events and tested on the rest of the events. The Monte Carlo events are randomly assigned to the training or testing subsets. Then the NNs are trained on all folds except for one fold, which is used for testing. The training process of NNs needs to be monitored to prevent overtraining or undertraining. Overtraining is generally referred to if the NN memorises its input training examples and increases performance with each training process on the given training dataset, but exhibits no improvement on the validation dataset. Undertraining can be observed if the NN increases in performance, for example in accuracy, on both datasets, but the training is stopped too early, such that no convergence of the accuracy to a specific value is observed.

The accuracy of an NN is calculated by dividing the number of correctly classified events by the total number of events classified. A low accuracy, meaning an accuracy just above an accuracy that can be statistically achieved by randomly classifying the input events, can be a result from over- or underfitting in the NNs. Overfitting can appear in highly flexible models, for example non-linear models. Here, non-signal events that are particular to some specific training events are picked up and learned by the model. This may result in the NN failing to classify other input data reliably. Non-signal events are referred to as background events. Underfitting can occur if the NN is not able, due to lacking flexibility, to adequately capture an underlying structure in the training dataset. More flexibility can be achieved, for example, by adding hidden layers and nodes to the model, which come with additional sets of weights and biases that can be optimised by the gradient descent algorithm.

The classification outputs are analysed in Kolmogorov-Smirnov (KS) tests [38, 39]. The KS test is sensitive to differences in both location and shape of the distributions and

is used to determine if the classifier output signal and background distributions are significantly different from each other. The KS tests used in this thesis, see for example Figure 5.3, output the P-value for the signal and background distributions. The P-value is the probability that the two tested distributions are as different as the observed ones, or more different than observed. By analysing the classifier output distributions for training and validation events, possible overtraining can be identified. Since the KS test was initially designed to evaluate continuous distributions, although it can be used for binned distributions too, in this thesis it is used with care.

A ROC curve, which shows the sensitivity against the specificity, is plotted for each output class, see for example Figure 5.8. The sensitivity, or true positive rate, measures the proportion of events that belong to a certain output class and are correctly identified. The specificity, true negative rate, refers to the proportion of events that do not belong to a certain output class and are correctly identified. ROC curves are used to evaluate the trade-off between the sensitivity and specificity for a classification model using different classifier thresholds. The area under the ROC curve (AUC) can be seen as a measure of how capable the model is in distinguishing between the signal and background.

Loss curves, see for example Figure 5.6, show the NNs learning progress. It represents changes in learning performance in terms of experience, gained through training on the training dataset. The loss curve can be used to determine if the model is underfitting or overfitting, similar to analysing the model's accuracy. It can also be used to stop the training when performance is best, specifically, if it reaches a minimum in loss on the validation dataset. The patience of the model affects this early stopping mechanism. The minimum improvement hyperparameter, $\Delta_{\text{min.}}$, can be used to trigger the early stopping mechanism. In addition to that, using the loss curve, the training and validation datasets can be analysed to determine if they represent problem domain properly.

Lastly, the relevant statistics need to be taken into account to evaluate the NNs performance. NNs need a large number of training events to minimise the cost function to a degree that results in the desired performance of the model. Since, in the used datasets, the fraction of tH and ttH events is small compared to that of the background events, the signal-over-square-root-of-background (SoSB) curves need to be considered to evaluate if a NN is capable, in the applied pre-selection region, of separating signal and background events and classifying them further. Here, the SoSB values are plotted against the classifier threshold. The SoSB plots are used to determine the NN's classifier threshold cuts, which are recommended to be applied at the peak of the SoSB ratio. Additionally, the SoSB ratio indicates if the desired signal can be distinguished from background noise.

4. The LHC and ATLAS Experiment

The LHC, run by the European Organisation for Nuclear Research (CERN) is, as of now, the world's largest and most powerful particle accelerator [40]. Here, proton beams are brought to collision at four locations, at the four different experiments ATLAS, CMS, ALICE, and LHCb, around the accelerator ring [41–44].

4.1. LHC and HL-LHC

The LHC first started up on 10 September 2008 and re-uses the approximately 27 km long tunnel, 100 m below the surface, first constructed in the 1980s for the Large Electron-Positron Collider (LEP). It is a proton-proton collider and therefore the synchrotron radiation losses ΔE at such high velocities are lower than those of the electrons and positrons used at LEP, $\Delta E \propto m^{-4}$, because of the larger masses m of the colliding particles. The LHC consists of two rings in which two high-energy particle beams travel at velocities close to the vacuum speed of light in an ultrahigh vacuum and are kept on track via superconducting magnets. The particles are accelerated by a separate accelerator complex [40].

As of 2021, the LHC is currently in the 'Long Shutdown 2'-phase. Prior to the shutdown, it achieved centre-of-mass energies of 13 TeV, the highest to date. The upgraded High-Luminosity Large Hadron Collider (HL-LHC) is expected to start in mid-2027 and achieve centre-of-mass energies of around 14 TeV with an instantaneous luminosity increased by a factor of five and integrated luminosity increased by a factor of ten to $3\,000\,\mathrm{fb}^{-1}$, compared to the original design value [45].

4.2. ATLAS Experiment

The ATLAS (A Toroidal LHC ApparatuS) experiment, shown in Figure 4.1, along with the CMS (Compact Muon Solenoid) experiment are general purpose particle detector

experiments at the LHC. Both experiments are used to probe the SM and search for physics beyond the standard model.

The ATLAS detector consists of different concentric detector layers, built with a cylindrical, forward-backward symmetry around the interaction point of the colliding protons. The detector, 25 m in height and 46 m in length, is described via a cylindrical coordinate system with an azimuthal angle ϕ, a polar angle θ and the Cartesian z-axis orientated along the beam line [41].

Often, the rapidity y, dependent on jet energy E and z-component of the jet momentum p_z, is used to express jet angles, which can be measured in the detector [41]:

$$y \equiv \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right). \tag{4.1}$$

Differences in rapidities Δy are invariant for Lorentz boosts along the z-axis. Therefore, the often unknown longitudinal parton boost does not affect rapidity distributions. In the high-energy approximation, where the jet mass is small compared to its energy, the pseudorapidity η is used instead of the rapidity with

$$\eta \equiv - \ln \left(\tan \frac{\theta}{2} \right), \tag{4.2}$$

and jet masses are neglected [41]. Other kinematic variables used as input parameters for the neural network are the transverse momentum p_{T} of the interacting particles, which is defined in the xy-plane perpendicular to the beam line, and the distance in pseudorapidity-azimuthal angle space [41]:

$$\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}. \tag{4.3}$$

4.3. ATLAS Detector

The detector layers are concentric around the beam pipe and consist of tracking detectors as the innermost layer. These are followed by electromagnetic and hadronic calorimeters. Muon detectors are installed as the outermost layer, see Figure 4.1 [41].

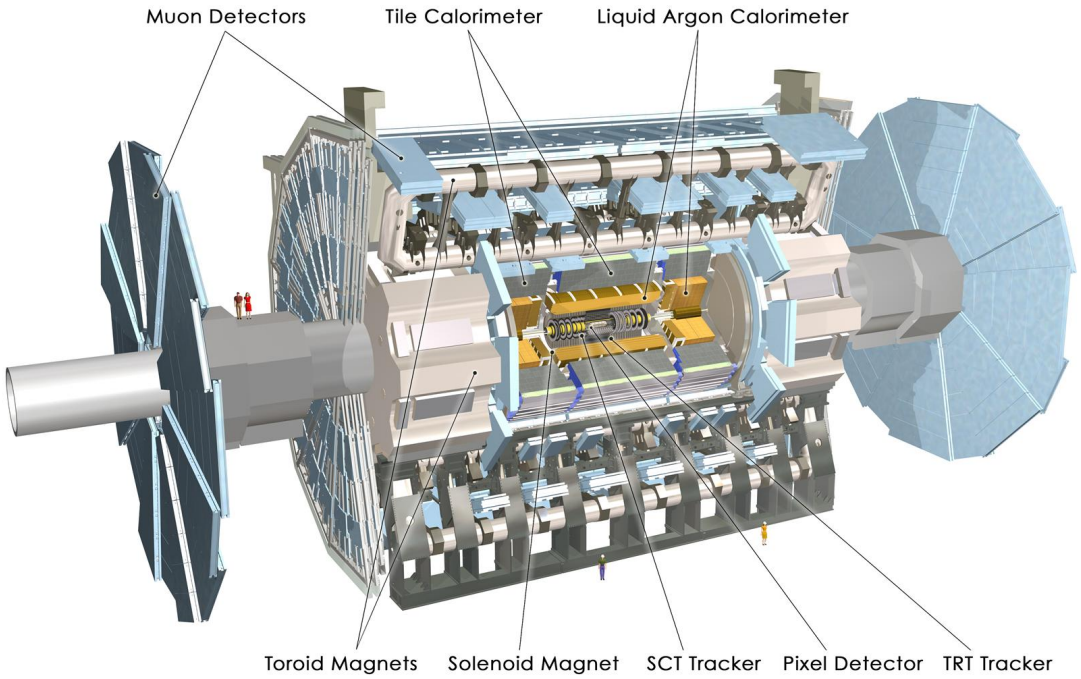

Figure 4.1.: A computer generated image showing the whole ATLAS detector, including the concentric detector layers, end-caps and toroidal magnets. ATLAS Experiment © 2008 CERN.

4.3.1. Inner Detector

The inner tracking detector is immersed in a 2 T solenoidal magnetic field and has three main subcomponents: the silicon-based Pixel detector, semiconductor tracker, and gaseous straw-tube Transition Radiation Tracker (TRT) [46, 47]. The inner detector provides charged-particle tracking over a pseudorapidity range $|\eta| < 2.5$ [48]. Charged particles that pass through the magnetic field inside the detectors experience a Lorentz force. As a result, their tracks are bent. In addition to the particle tracking, the curvature of the tracks is used to reconstruct the transverse momentum p_{T} of the charged particles.

The Pixel detector, closest to the beam line, contains approximately 100 million read-out channels, with the insertable B-layer, and provides a maximum spatial resolution of around 6 µm in the short direction and 65 µm in the long direction of the rectangular pixel. It is used to reconstruct secondary vertices, vertices outside the beam profile in a collider experiment, from the decay of particles or for b-tagging of jets. It is also used to reconstruct primary vertices, the location of an individual particle collision, coming from the proton-proton interaction region and the transverse momenta p_{T} of the charged particles [48, 49].

The semiconductor tracker, containing about 6.3 million semiconducting strips, has a spatial resolution of about 17 µm in the cross-strip direction and approximately 580 µm longitudinally due to the way of construction [50]. The particles that pass through the detector create charges in the form of electrons and holes in the material by ionisation. These are then separated by an electric field and detected at electrodes connected to the ionised material. Together with the Pixel detector, the semiconductor tracker can precisely measure the paths of the particles resulting from the collision.

Both the Pixel and semiconductor tracker use semiconducting silicon for particle tracking. In the first detector it is arranged as pixels, in the second one it is arranged as strips.

The outermost layer of the inner detector consists of the TRT, which contains about 3 million straw-tubes filled with a Xenon-gas mixture, each being 4 mm in diameter. The TRT has a spatial resolution of approximately 130 µm for charged particle tracks fulfilling $|\eta| < 2$. It detects transition radiation, caused by highly-relativistic charged particles crossing through transitions between two media with different indices of refraction of electromagnetic radiation [47]. Thus, the TRT can be used for particle identification of highly-relativistic particles.

The inner detector has a relative momentum resolution of $\frac{\sigma_p}{p} = (4.83 \pm 0.16) \times 10^{-4}\,\mathrm{GeV}^{-1} \times p_\mathrm{T}$ [51].

4.3.2. Calorimeters

The calorimeter system, located outside the solenoidal magnetic field, is used to measure the energy of the incident particles and can be divided into electromagnetic and hadronic calorimeters [52]. The energy is measured based on absorption, partial and total, in the detector material. The sampling calorimeters used in the ATLAS experiment consist of alternating layers of active materials, such as liquid Argon, to provide a detectable signal and passive materials, for example combinations of Lead, Tungsten, and Copper, to degrade the particle energy [53]. By separating the media that produce the particle showers and the media that measure the deposited energies, materials can be chosen well-suited to their tasks to build, for example, more compact calorimeters.

In electromagnetic calorimeters, the incident particles start an electromagnetic shower cascade of secondary particles such as electrons, positrons, and photons, which deposit their energy mostly in the passive materials. The electromagnetic calorimeters have a relative energy resolution of $\frac{\sigma_E}{E} = \frac{10\%}{\sqrt{E}} \oplus 0.7\%$ in the pseudorapidity range of $|\eta| < 3.2$ [41].

In hadronic calorimeters, hadronic showers are generated. These showers are more complex than electromagnetic showers due to strong interactions with the nuclei in the calorimeter's material. The resulting cascade consists of hadronic and electromagnetic showers. Here, charged as well as neutral hadrons and jets can be absorbed. The hadronic calorimeters have a relative energy resolution of $\frac{\sigma_E}{E} = \frac{50\%}{\sqrt{E}} \oplus 3\%$ in the pseudorapidity range of $|\eta| < 3.2$ for the barrel and a relative energy resolution of $\frac{\sigma_E}{E} = \frac{100\%}{\sqrt{E}} \oplus 10\%$ for end cap in a pseudorapidity range of $3.1 < |\eta| < 4.9$ in the forward direction [41].

In the ATLAS experiment, liquid Argon-Lead sampling calorimeters are used as electromagnetic and hadronic calorimeters, except in the barrel region. Here, scintillator-steel tile calorimeters are used as hadronic calorimeters [53].

4.3.3. Muon Detectors

The muon spectrometers make up the outermost layer of the ATLAS detector structure, see Figure 4.1, and cover a pseudorapidity range of $|\eta| < 2.7$ [41, 54]. Here, particles that are not fully absorbed by the inner detector and calorimeters, mainly muons, are detected. Muons are not detected in the electromagnetic and hadronic calorimeters as they do not emit Bremsstrahlung as rapidly as, for example, electrons due to their higher mass compared to that of the electrons. Here, similar to the inner detector, the particle tracks are bent by a magnetic field, in this case applied by three air-core superconducting toroidal magnets, so the muon momenta can be measured. The muon spectrometer is made up of individual muon chambers using four different technologies: thin gap chambers, resistive plate chambers, monitored drift tubes, and cathode strip chambers, resulting in a transverse momentum resolution of approximately 10% for 1 TeV tracks [41, 55].

All detector types are frequently upgraded. As of June 2021, the new small wheel upgrade for the muon spectrometer is nearing completion and the ITk Pixel detector upgrade is currently in the final stages of research and development, with production scheduled to start in 2021, too [55–57].

4.3.4. Trigger System

Lastly, data taking at the ATLAS experiment would not be possible without the trigger system.

The ATLAS Level 1 trigger is a hardware-based system, designed to decide which subsection of the occurring events in the beam pipe is recorded. The selection algorithm is

based on predefined criteria, based on the hard-scatter events [58].

The high level trigger is a software based trigger mechanism and consists of the Level 2 trigger and Event Filter trigger. Events accepted by the two subsystems are then written to mass storage [59].

4.4. Monte Carlo Event Generators

Monte Carlo event generators are a tool to simulate full collision events as they would happen at the LHC. First, the collision events are simulated at matrix element level. Then, the subsequent showering and hadronisation processes of the particles are generated in showering-generators. Afterwards, the generated simulation data is fed into detector models, which simulate the passage of the resulting particles from the collision through the detector material [60, 61]. The GEANT 4 and ATLFAST-II detector simulations are used. Since the GEANT 4 detector simulation is more CPU-intensive, often the faster, but less precise ATLFAST-II detector simulation is used. After GEANT or ATLFAST-II, the readout of the electronics in the detector is simulated. Then, the data is fed into the event reconstruction algorithms. The simulations allow, for example, for studies on the impact of experimental cuts on data, detector effects and statistics of the collision experiment. These simulations generate results at hadron levels, which gives rise to the ability to compare them to experimental data and interpret them more precisely than at detector level, but also general enough to not be model-dependent at particle level. The most commonly used matrix element generator is POWHEG BOX [62]. HERWIG, PYTHIA, and SHERPA are used for showering simulation [63–65]. The data presented in this thesis is solely generated by Monte Carlo generators.

5. Boosted Decision Trees and Neural Networks in $t\bar{t}H$ Analyses

In the following sections, the classification performance of (binary) BDTs used in previous ATLAS analyses is presented, in Section 5.2, and compared to that of NNs trained on the same dataset, which is described in Section 6.1. Prior to this comparison, the training process and classification performance results of the NNs are documented, in Section 5.3. Lastly, in Section 5.4.2, the NN's performance on a larger dataset is outlined.

The NNs used in this thesis are a different approach to production mode selection from the previously employed BDTs. Here, the NNs identify production mode dependencies on variable distributions. Then a cut is applied to the classifier output.

Binary and multi-class NNs are used to classify the input events into 2, 3, 4, or 5 classes. The 3-class NNs are used to separate the ttH and tH signal modes from the tt + jets background, 4-class NNs are used to separate the two signal modes from the background, which is split further into tt + light jets and c, and $tt + b$. In the 5-class NNs, the two signal modes are separated from the $tt + b$, tt + light jets, and $tt + c$ background events. The binary classification NNs are trained to separate the ttH signal from the tt + jets background, analogous to the BDTs used in previous ATLAS analyses.

Using the ROC curves, confusion, Train/Test, and signal over square root of background plots, the optimal cut on the classifier threshold can be determined such that events that pass this threshold are, with a high certainty, the desired events.

5.1. Event Selection

The events from the smaller dataset, used in Sections 5.2, 5.3 and 5.3.3 and discussed in 6.1, are required to have exactly one electron or muon. In addition to that, they need to fulfil the electron and muon identification criteria defined in Ref. [66], based on Ref. [67–69]. Then, cuts for a minimum transverse momentum p_T and upper cuts on the

absolute value of the pseudorapidity $|\eta|$ and other impact parameters analogous to the cuts in Ref. [66] are applied. Furthermore, events are vetoed if they contain two or more hadronic tau candidates, which are distinguished from other jets using track multiplicity and a multivariate discriminant [66, 70].

In addition to the previously stated event selection criteria, leptons are further required to satisfy identification criteria based on a likelihood discriminant [66, 68].

The jet selection is analogous to previous analyses [66]. Jets are reconstructed from topological clusters of calorimeter depositions calibrated at the electromagnetic scale [71]. The average pileup energy contribution is subtracted according to the jet area. They are calibrated as described in [71]. Furthermore, jets are required to fulfil minimum cuts on transverse momenta $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ and upper cuts on the absolute value of the pseudorapidity $|\eta| < 2.5$ as defined in [66]. b-jets are tagged with the MV2c10 multivariate algorithm at 60%, 70% and 85% working point (WP) [72].

The applied pre-selection aims at selecting ttH events with semi-leptonic tt decays and the $H \to bb$ Higgs decay mode. This leads to the background mainly containing $tt + b$ events.

To take advantage of the higher jet and b-jet multiplicities of the ttH signal process, events are classified into non-overlapping regions based on the total number of jets and the number of b-tagged jets. The non-overlapping regions are shown in Table 5.1. Due to the tH production mode features, described in Section 2.4, a greater fraction of tH events is expected in the 5 jet region. Thus, in this thesis, a NN analysis is performed for each region separately and in the combined 6 jet inclusive regions, whereas the BDT was solely used on the combined 6 jet regions. The event selection in the larger dataset, used in Section 5.4.2, is presented in Section 5.4.1.

Region	#leptons	#jets	#b-tags			
			@60%	@70%	@77%	@85%
6 jet, high$^{\geq 6j}_{\geq 4b}$	=1	≥6	≥ 4	≥ 4	≥ 4	≥ 4
6 jet, low$^{\geq 6j}_{<4b}$			< 4			
5 jet, high$^{5j}_{\geq 4b}$		=5	≥ 4			
5 jet, low$^{5j}_{<4b}$			< 4			

Table 5.1.: The non-overlapping event regions. The regions are defined on the number of: leptons, jets and b-tagged jets at different working points (WPs).

5.2. Boosted Decision Tree Results from Previous Analyses

The classification BDTs used in previous analyses are trained to separate the ttH production from the tt + jets background, containing tt + light jets, $tt + c$ and $tt + b$ events in the 6 jet inclusive region [32]. The BDTs are trained on half of the total events and validated on the other half of events. A complete list of the monte carlo generated input variables to the classification BDTs and their definitions can be found in Table B.1.

The nominal ttH signal sample was generated with PowhegBox+Pythia8. The $tt + b$ background was generated using the PowhegBox+Pythia8 $ttbb$ four flavour scheme (4FS) and for the $tt + c$ and tt + light jets background, the PowhegBox+Pythia8 tt 5FS was used.

In the previous analysis, it was shown that training on the signal regions in Table 5.1, only including events with ≥ 4 b-jets at the 70% WP, performs better than the training on the looser 85% b-tagging WP regions. Table 5.2 shows the performance of different BDT trainings when applied on the full MC16 samples with loose b-tag cuts, requiring ≥ 6 jets with ≥ 4 b-jets at the 70% WP, and tight b-tag cuts, requiring ≥ 6 jets with ≥ 4 b-jets at the 60% WP. It is observed that the BDT trained at ≥ 4 b-jets at the 85% WP performs always worse than that trained at the 70% WP, but better than the BDT trained at the 60% WP. The overall best performance was shown by the BDT trained at ≥ 4 b-jets at the 70% WP and evaluated on the region requiring ≥ 4 b-jets at the 60% WP. It was concluded that the BDT trained at the 85% WP was too biased towards events missing in the 70% WP samples and the BDT trained at the 60% WP was limited in training events due to tight b-tag cuts.

BDT training region	BDT testing region	
	≥ 4 b-jets at the 60% WP	≥ 4 b-jets at the 70% WP
≥ 4 b-jets at the 60% WP	0.760	0.748
≥ 4 b-jets at the 70% WP	0.761	0.758
≥ 4 b-jets at the 85% WP	0.757	0.755

Table 5.2.: AUCs of the 2019 analysis BDT ROC curves from MC16 trainings applied to samples with ≥ 4 b-jets at the 60% WP, ≥ 4 b-jets at the 70% WP and ≥ 4 b-jets at the 85% WP [73].

5.3. Training of the Neural Networks on the small Dataset

In total, six binary classification NNs are trained on the small dataset, described in Section 5.1, using the same input variables as the BDTs from previous analyses, listed in Appendix B.1. The output variables of the *ttH* reconstruction BDT can be used as inputs for the NNs trained on the 5 jet region of the small dataset, as they are filled with data, i.e. the histograms are not empty. This is not the case for the 5 jet region of the larger dataset, thus they can not be used in the Training of the NNs on the larger dataset, see Section 5.4. In addition to that, a 5-class NN is trained on the small dataset using the input variables from Table B.2, see Section 5.3.2.

5.3.1. Binary classification Neural Networks

One NN is trained on each of the 4 regions shown in Table 5.1. NN training on the 6 jet inclusive high region is analogous to the BDT training at the 60% b-tag WP in terms of event selection criteria. In addition to that, two NNs are trained on a combination of the 6 jet inclusive high and low region, in the following referred to as 6 jet combined region. This represents the NN training on samples at the 70% WP.

The NNs share the same hyperparameters, except for one of the two NNs trained on the 6 jet combined region. It is in the following referred to as complex NN. The hyperparameters are shown in Table B.3. The more complex model is trained to check the simpler models for possible underfitting of the data.

If their performance in training and testing matches, the setup is not sensitive to the randomly assigned initial weights and biases and is able to determine the same minimum of the cost function.

The hyperparameters are initially chosen adequately for models of the given complexity. For both models, no over- or undertraining can be observed and the early stopping mechanism was triggered in each training.

The total number of Monte Carlo events in each region is shown in Table 5.3. The Monte Carlo events are re-weighted by dividing the total number of Monte Carlo events by the sum of the weights for normalisation reasons. Then they are multiplied with the cross section of the corresponding process.

The event yield of tH, ttH, $tt + c$ and $tt + b$ processes that can be detected with ATLAS

in Run II is shown in Table B.4. It can be observed that the event yield of tH events is significantly smaller than the event yield of ttH. Thus the tH background was neglected in the binary classification NNs, analogous to previous analyses [32, 66].

Class	Number of Monte Carlo events					
	5 jet low	5 jet high	6 jet low	6 jet high	5 jet	6 jet
	smaller dataset				larger dataset	
ttH	81,160	85,081	205,230	215,050	1,407,148	1,236,630
tH	3,270	2,853	7,336	6,978	83,478	47,006
tt + light jets	10,284	7254	27557	22063	1,600,000	3,607,537
$tt + b$	136,987	97,837	278,391	235,345	1,600,000	1,919,591
$tt + c$	3,204	382	6410	771	1,600,000	1,324,606

Table 5.3.: Number of Monte Carlo events in each class and region in the smaller and larger dataset.

5.3.2. Multi-Class Neural Network

In addition to the 6 binary classification NNs described in Section 5.3.1, one 5-class NN is trained on the small dataset. Here, in contrast to previous analyses, the tH background is included. To increase performance for the tH class with higher Monte Carlo event statistics, it was trained on the 6 jet inclusive combined region, corresponding to the BDT training in previous analyses at a b-tag WP of 70%.

Analogous to the binary classification NNs, this NN uses the same hyperparameters as the simple NNs.

In contrast to the binary classification NNs, different input variables than the ones used in the BDTs are considered, and can be found in Appendix B.2. In the variable selection, the separation power S of each variable in each of the 10 permutations of classes, for example tH vs. ttH, tH vs. $tt + b$ etc., was computed. It is defined as $S = \frac{1}{2}\sum_i \frac{(a_i - b_i)^2}{a_i + b_i}$, where a_i and b_i represent the bin contents of the i-th bin of the two variable distributions. Then, the two variables with the greatest separation power in each permutation are chosen as input variables. No duplicate variables are used, as this would not improve classification performance. Variables that are heavily correlated are included to improve the poor classification performance, see Section 5.3.3, as they still show a good separation power in the permutations.

The same variable selection procedure is used for the NNs trained on the larger dataset, described in Section 5.4.

5.3.3. Performance Results of the Neural Networks

In this Section, the performance results of the binary classification NNs and the multi-class NN on the small dataset are presented. The plots shown in this thesis are generated using 139 fb^{-1} monte carlo generated pp collision data at $\sqrt{s} = 13\,\mathrm{TeV}$.

Performance Results of the Binary Classification Neural Networks

The binary classification NNs use the same input parameters as the BDTs from previous analyses, listed and defined in Appendix B.1. Their correlations are similar in all regions and a representative correlation matrix for the 6 jet inclusive combined region is shown in Figure 5.1. Variables B^3_{jet} and B^4_{jet} carry no discriminating information in the 6 jet inclusive high region and the variable B^4_{jet} carries no discriminating information in the 6 jet inclusive low region. Thus they are excluded from the correlation matrix plot.

One can observe a significant correlation of the $m_{bb}^{\text{Reco, Higgs}}$ and $m_{\text{Reco, } b_{\text{lep. top}}}$ variables in all regions, which is expected as the invariant mass of the two b-jets, $m_{bb}^{\text{Reco, Higgs}}$, used to calculate the Higgs candidate influences the invariant mass of the Higgs candidate and the b-jet coming from a leptonically decaying top quark, $m_{\text{Reco, } b_{\text{lep. top}}}$.

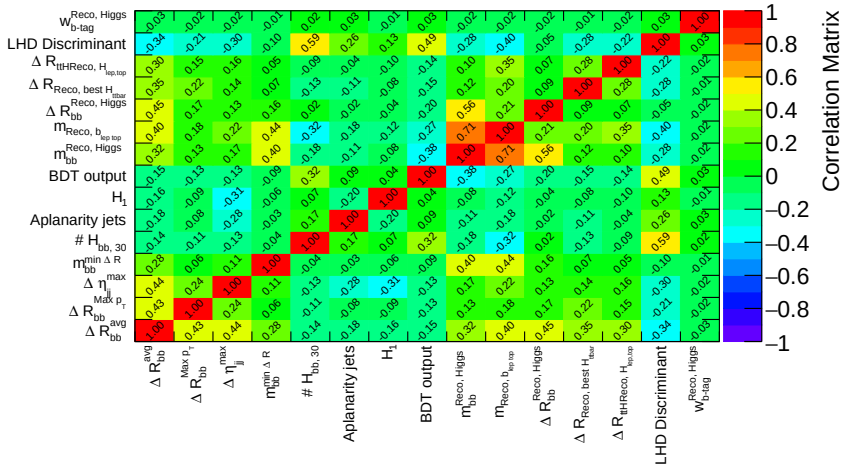

Figure 5.1.: Correlation matrix plot for the BDT and binary classification NN input variables, shown in the 6 jet inclusive combined region of the smaller dataset, excluding the tH samples.

The AUCs of the ROC curves show the best performance in the 5 jet high region. Overall, the performances in all regions are comparable without major differences in terms of AUCs, for an exact list see Appendix B.7. All AUCs are within 5% of the average of the best performing NN in the 5 jet high region. All ROC curves are well defined and their AUCs are significantly greater than 0.5, indicating a better performance than a random binary classification

In terms of AUCs of the ROC curves, the complex binary classification NN performs marginally better than the more simple model, cf. Table B.7. This trend can be seen throughout the performance comparison. Since the added complexity does not increase the NN's performance significantly, it is out-weighed by the additional computing time needed and increased sensitivity to hyperparameters. Thus the simpler model is used throughout this thesis as this is the more stable model. The region selection has a larger impact on the NNs performance than the fine tuning of the hyperparameters.

Confusion plots are shown in Figure 5.2 for a representative for a binary classification example and in Figure 5.4 for the remaining plots, are used to determine the NN's confidence of classifying the input data. They show similarities in input variable distributions within the input classes. Here, the fraction of events in each class classified as ttH is shown depending of the classifier output. The fraction of events is normalised to all the events in the corresponding class. Ideally, the binary classification NNs would classify 100% of the ttH events as such, with a classifier value of 1.0, and 100% of the events in the other categories with classifier values of 0.0.

The confusion plots of the simple and complex binary classification NNs trained and validated on the 6 jet inclusive combined region are shown in Figure 5.2. They indicate similar distributions to the ones of each individual region, in Figure 5.4. The significant confusion of ttH with the $tt + c$ and $tt + b$ classes in the NNs, trained and evaluated on the 6 jet inclusive combined region, influences the cuts on a classifier threshold minimum that have to be made on the classifier output.

Overall, the confusion plots for the NNs evaluated in the 5 jet high and 6 jet inclusive high regions indicate smaller fractions of background events being classified as ttH events with classifier thresholds ≥ 0.5, in comparison to the NNs trained and evaluated on the 5 jet low or 6 jet inclusive low regions. The classification of ttH signal events as such is of similar performance in the high and low regions.

For both NNs, the simple and complex NN trained and validated on the combined 6 jet inclusive region, no systematic over- or undertraining can be observed in the Train/Test

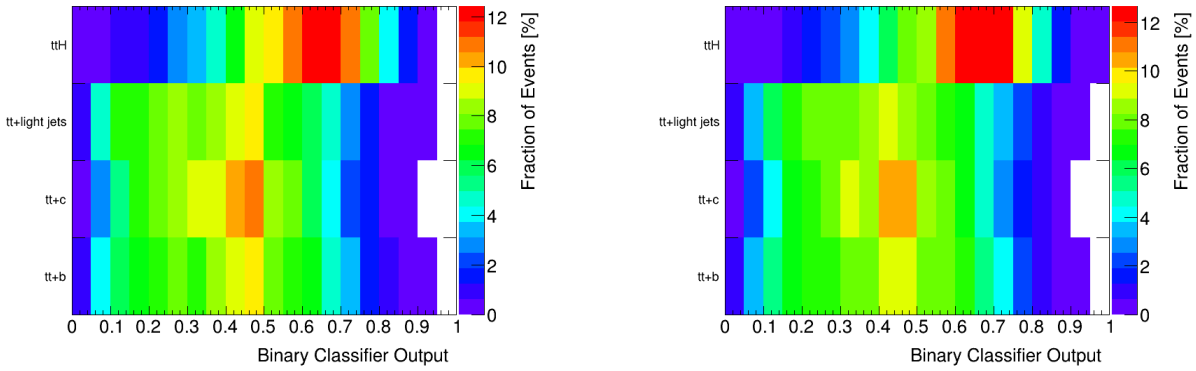

Figure 5.2.: Confusion plots of the simple (left) and complex (right) binary classification NNs trained on the 6 jet inclusive combined region.

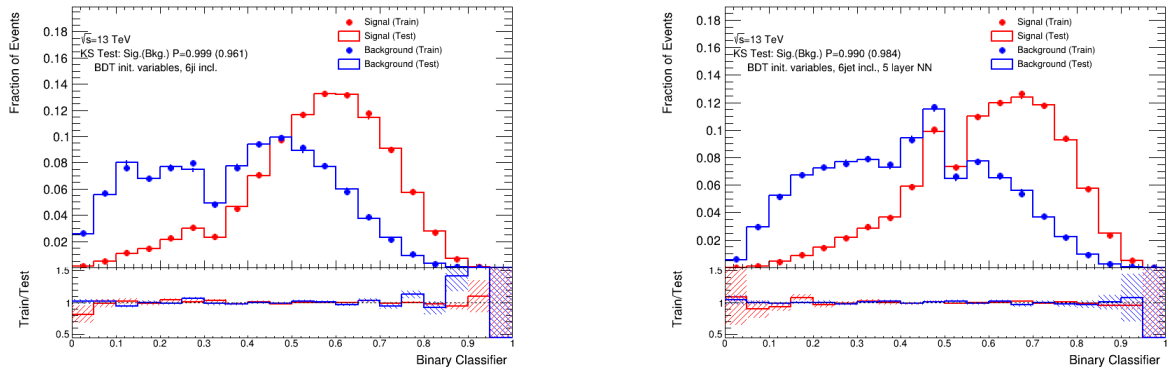

Figure 5.3.: Train/Test plots of the simple (left) and complex (right) binary classification NN, trained and tested on the 6 jet inclusive combined region. A representative fold is shown.

plots, as shown in Figure 5.3, since the observed Train/Test ratio includes 1 within its uncertainty levels in most bins. No under- or overtraining was observed for the NNs operating on the individual 5 jet high and low or 6 jet inclusive high and low regions. Here, the complex NN shows no significant improvement, too.

By taking the SoSB ratio plots into account, the threshold cuts can be determined. The threshold is recommended to be cut at the peak of the SoSB ratio.

A representative SoSB ratio, from the 6 jet inclusive simple and complex NN, for the binary classification NNs can be seen in Figure 5.5, with more in Figure A.1 in the Appendix. Here, the SoSB ratio is plotted against the binary classification threshold.

No major difference can be observed in the SoSB performance of the simple and complex binary classification NNs. Both curves peak, with a SoSB ratio of ~ 4.65 at a classifier

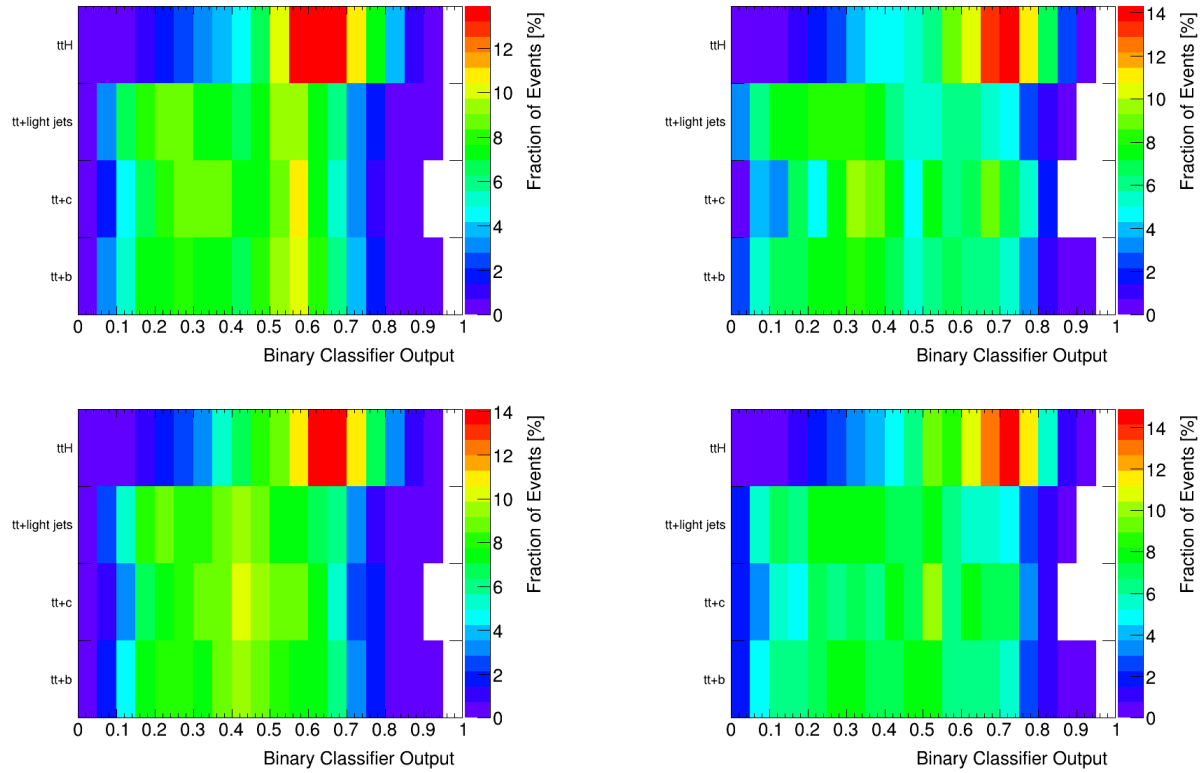

Figure 5.4.: Confusion plots of the binary classification NNs, trained and validated on the 5 jet low (top left) and high (top right) region and 6 jet inclusive low (bottom left) and high (bottom right) region.

threshold ≥ 0.5. The SoSB ratio of these NNs trained and tested on the 6 jet inclusive combined region is, on average, the highest compared to the NNs trained on the other regions. This is expected given the comparably small amount of ttH signal in each individual region. Given the event yields, the SoSB ratio of the NNs trained and tested on the 5 jet high region is, on average, larger than that of the NNs trained on the 5 jet low region. Analogously, the NN's SoSB ratio on the 6 jet inclusive high region is greater than that on the 6 jet low region. On average the SoSB ratio in the 6 jet region is increased compared to that of the NNs trained on the 5 jet regions individually.

Representative cross entropy loss and accuracy plots, as introduced in Section 3.7, for the simple NN operating on the 6 jet inclusive combined region, are shown in Figure 5.6. The loss and accuracy plots for the NNs, trained and evaluated on the other regions are analogous to the presented ones, with accuracy and loss functions converging to the same values. The early stopping mechanism was triggered in all learning processes.

The validation loss underestimates the training loss in all trained NNs. This is caused by `keras`' deactivation of the dropout layer, introduced in Section 3.6, during testing times

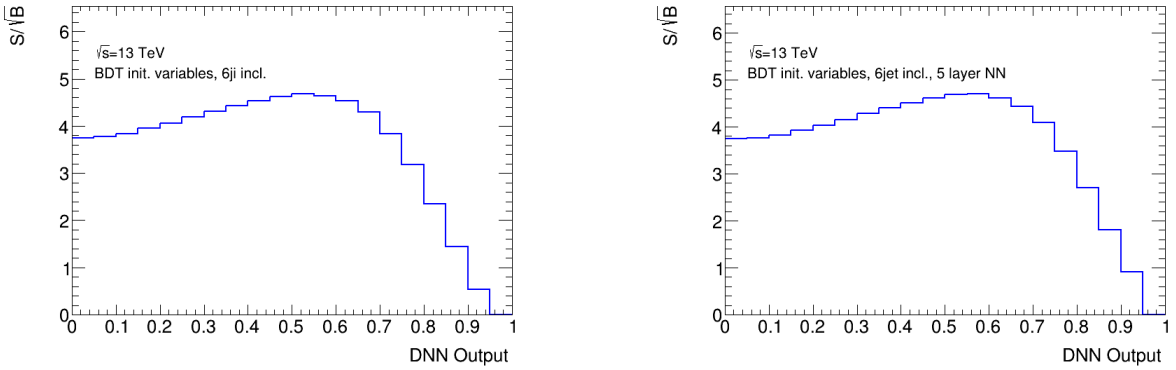

Figure 5.5.: SoSB plot for the simple (left) and complex (right) binary classification NNs, trained and tested on the 6 jet combined region.

but not during the training of the NN. Thus, the dropout layers are reflected in the loss during training time, but not during testing time [35].

In general, the loss curves decrease with the learning process of the NNs. No NN shows a noticeably rapid decrease of the loss function. All losses converge to the same loss of ~ 0.6.

The accuracy converges to ~ 0.66, in all trained NNs, which indicates that about 66% of events are classified correctly. Here, in the binary classification NNs a classifier output threshold of 0.5 is applied to evaluate if a signal event is correctly classified, threshold ≥ 0.5, vice versa with background events and classifier thresholds < 0.5.

For the best performance results, the complex binary NN, trained and tested on the 6 jet inclusive high and low regions is recommended to use, although the more simple model, trained and tested on the same regions, performs only marginally worse.

It is observed that all AUCs of the binary classification NNs indicate a better performance than the BDTs used in previous analyses.

Performance Results of the Multi-Class Neural Network

The binary classification NNs are able to minimise the loss function, thus increase their accuracies. Their ROC curves are well defined and have an AUC significantly greater than 0.5. The Train/Test plots and confusion plots show large fractions of events correctly classified. These performances can not be matched by the multi-class NN, trained and tested on the 6 jet inclusive combined region. Although the classification performance for the ttH signal is better than that of the other classes, it is not comparable to the

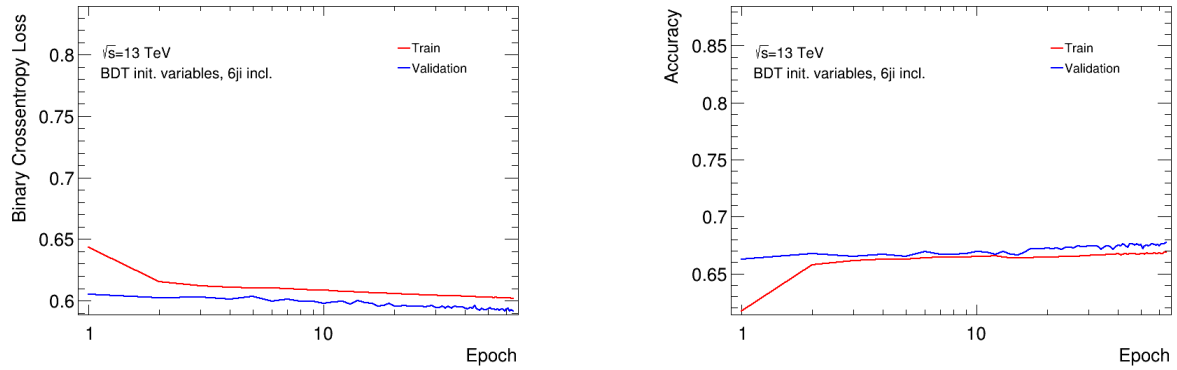

Figure 5.6.: The binary cross entropy loss function and accuracy for the simple binary classification NN, shown for the folds that correspond to the folds presented in the classification plots, trained and validated on the 6 jet inclusive combined region. The validation loss underestimates the training loss.

performance of the binary classification NNs.

Despite the AUC of the ROC curve for the ttH classification being comparable to the ones of the binary classification NNs, displayed in Table B.8 in the Appendix, the confusion and Train/Test plots, see Figure 5.7, show large differences in performance in comparison to the binary classification NNs.

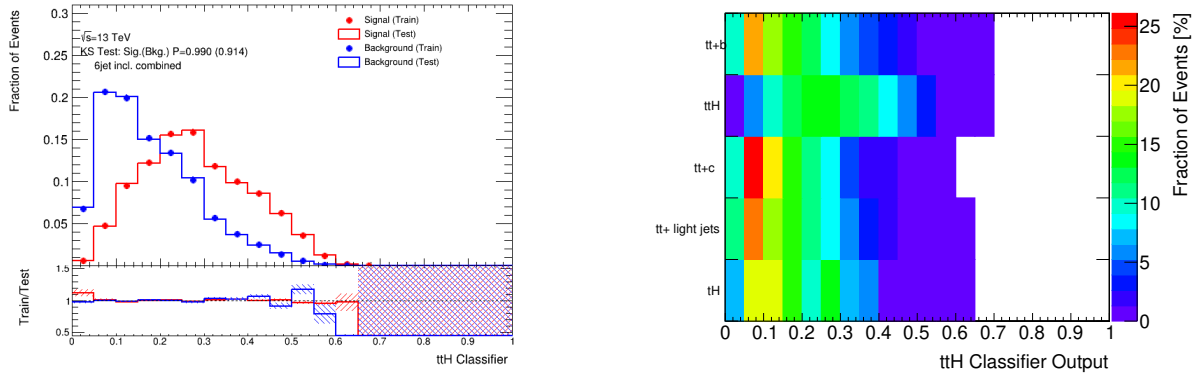

Figure 5.7.: Representative Train/Test (left), and confusion (right) plots of the multi-class NN for the classification of events into the $t\bar{t}H$ output class, trained and validated on the 6 jet inclusive high and low regions combined.

Most of the ttH signal events are not classified as such, having ttH classifier outputs < 0.5, see Figure 5.7. The performance in classifying all background events to the ttH signal as such, is better than the performance of the binary classification NNs. This leads to a comparable AUC of the multi-class NN's ttH ROC curve to the AUC of the binary classification NN.

An overall worse performance of the multi-class NN compared to its ttH classification performance can be seen for the $tt + b$, $tt + c$ and $tt +$ light jets categories. An AUC of ~ 0.5 can be observed for the ROC curves of the $tt + b$ and $tt +$ light jets classes, which indicates a classification performance close to that of a random classification of events.

In addition to that, the SoSB ratio of $tt + b$ and $tt +$ light jets is smaller than 1.0, at $139\,\mathrm{fb}^{-1}$ data, for classifier thresholds ≥ 0.5, indicating that the $tt + b$ and $tt +$ light jets signal events vanish in the background noise. Classification performance and SoSB ratios for the $tt + c$ classifier are higher than that of the $tt + b$ and $tt +$ light jets, see Figure A.2, but no confident classification can be made either.

tH events are also included in the multi-class NN. Here, the ROC curve fluctuates and the sensitivity can not be described as a well defined function of the specificity, as shown in Figure 5.8. This is caused by numerical calculation problems resulting from the reduced statistics. The tH statistics in the 6 jet inclusive combined region are, with the requirement of ≥ 4 b-jets at a b-tag WP of 70%, insufficient for the classification. This is confirmed by the SoSB ratio plot in Figure 5.8. Here, it is observed that the SoSB ratio is < 1 for all tH classifier output thresholds, indicating that the signal can not be distinguished from background noise, see Figure 5.8.

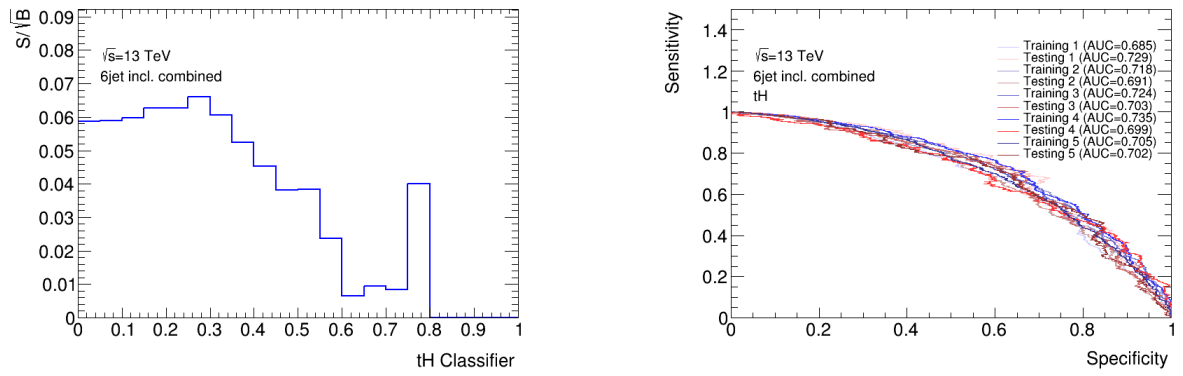

Figure 5.8.: SoSB ratio and ROC curve for the tH output class of the 5-class NN, trained and tested on the 6 jet inclusive combined region.

A representative categorical cross entropy loss and accuracy function are shown in Figure 5.9. Here, analogous to the loss functions of the binary classification NNs, the early stopping mechanism was triggered and the training loss is often underestimated by the validation loss. The multi-class NN is not able to reduce the categorical cross entropy loss and as a consequence does not improve on accuracy. The significant fluctuations of about 10% in the validation accuracy, which is about 25% of the maximum reached accuracy,

indicate a classification of events marginally better than randomly classifying the events. This is supported by the fact that one can not observe a convergence nor a trend in the validation accuracy curve.

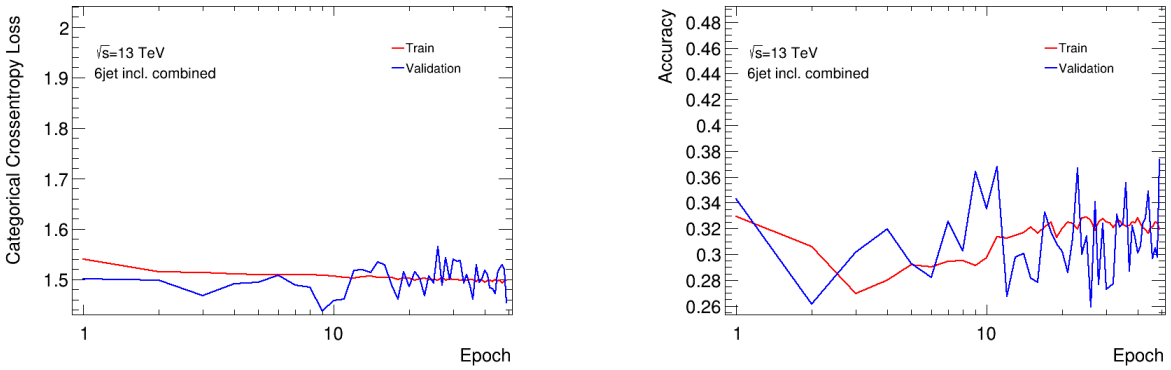

Figure 5.9.: Representative categorical cross entropy loss function (left) and accuracy (right) for the 5-class NN.

Although the variables that show the best separation power for each permutation of classes have been selected as input variables, see Section 5.3.2, the tight 70% WP cuts on the 6 jet inclusive combined region and a restricted choice of variables lead to the low performances. In the ttH ROC curve an average AUC of 0.7461 is observed, whereas the other AUCs of the ROC curves from the other four classes are, on average, smaller than 0.7400. Thus, the ttH classification performance is, on average, better than the classification performance in the other classes, this is due to the bias of variables available in the ntuples. These were not produced with the intend of classifying into multiple classes, but solely for classifying ttH signal in a binary classification.

Thus the focus of this thesis shifted towards using a larger dataset including more Monte Carlo statistics for each process and reducing the number of output classes, instead of training multi-class NNs on the 5 jet high and low regions of the small dataset. Although performance increase in classifying tH events can be expected in the 5 jet region, the performance results will still be dominated by low statistics.

5.4. Training of the Neural Networks on the Larger Dataset

To avoid bad classification performance due to low statistics, as can be seen for example in the ROC curve of the tH classifier in Figure 5.8, a larger dataset, containing more

Monte Carlo events with looser b-tagging cuts is used. Here, in the 6 jet inclusive region ≥ 6 jets with ≥ 4 b-jets at the 85% WP, and in the 5 jet region 5 jets with ≥ 3 b-jets at the 85% WP are required. Analogous to the previous sections, NNs are trained and tested on both regions. Due to computational limitations, the events in the 5 jet region were cut to a maximum number of 1.6 million events and re-weighted accordingly to preserve event yields. Processing the events from the 6 jet region was less computationally intensive, such that the whole number of available events was kept. The number of events used in the training for each class can be seen in Table 5.3.

To improve on accuracy and classifier output thresholds, output classes have been merged: In addition to a 5-class NN, 3-class and 4-class NNs have been trained. In the 3-class NNs the $tt + c$, $tt + b$ and $tt +$ light jets backgrounds have been merged into one $tt +$ jets background category and in the 4-class NNs, the $tt + c$ and $tt +$ light jets output classes have been merged to one $tt +$ light jets and c category. The multi-class NNs trained on the larger dataset use the same hyperparameters as the ones trained on the smaller dataset, as can be seen in Table B.3 in the Appendix, except for the number of folds. As the 5-fold cross validation check showed, the NNs trained on the smaller dataset performed similarly in all folds. Thus, the number of folds for the NNs trained and tested on the larger dataset, has been reduced to 4 to enlarge the training and testing subsets and provide more events.

The input variables used for the multi-class NNs trained and tested on the 6 jet inclusive and 5 jet region, listed in Tables B.5 and B.6, are selected via the same procedure as for the multi-class NN trained on the smaller dataset, see Section 5.3.2. The binary classification NN uses similar input variables to the BDTs. For the 5 jet region, no ttH reconstruction BDT output variables can be used as these are not filled properly, i.e. their histograms are empty, in contrast to the small dataset.

The variable correlation matrix for the input variables from the 6 jet region is shown in Figure 5.10. Again, one can observe a number of heavily correlated variables.

The input variables to the 6 jet inclusive NNs show large correlations between the variables that use b-tagging information, namely $\Delta\eta_{bb}^{\text{max}}$ and $\Delta\eta_{bb}^{\text{Avg.}}$, indicating that the majority of b-tagged jet pairs has a similar difference in pseudorapidity $\Delta\eta$ close to the maximum value. Furthermore, they are heavily correlated to the distribution of the average ΔR of all b-tagged jets, $\Delta R_{bb}^{\text{avg.}}$, indicating a similar detection region in the detector, as differences in the azimuthal angle $\Delta\phi$ are small compared to differences in pseudorapidity $\Delta\eta$, see Definition 4.3.

In the 5 jet region, more input variables are correlated. In addition to the variables heavily correlated in the 6 jet inclusive region, the maximum $\Delta\eta$ of any two jets, $\Delta\eta_{jj}^{\text{max}}$, is strongly correlated with the equivalent variable for b-tagged jets, $\Delta\eta_{bb}^{\text{max}}$, indicating that the maximum $\Delta\eta$ of observed jets mainly comes from b-tagged jets. These variables are also correlated with the average ΔR of any two b-tagged jets, $\Delta R_{bb}^{\text{avg.}}$, and the average $\Delta\eta$ of any two jets, $\Delta\eta_{jj,\text{Avg}}$.

Supporting the statement that the b-tagged jets are observed in a tight detector region, minimal invariant mass of all b-tagged jet pairs, $m_{bb}^{\text{min. mass}}$, is strongly correlated with the invariant mass of the two b-tagged jets with minimal ΔR, $m_{bb}^{\text{min. }\Delta R}$.

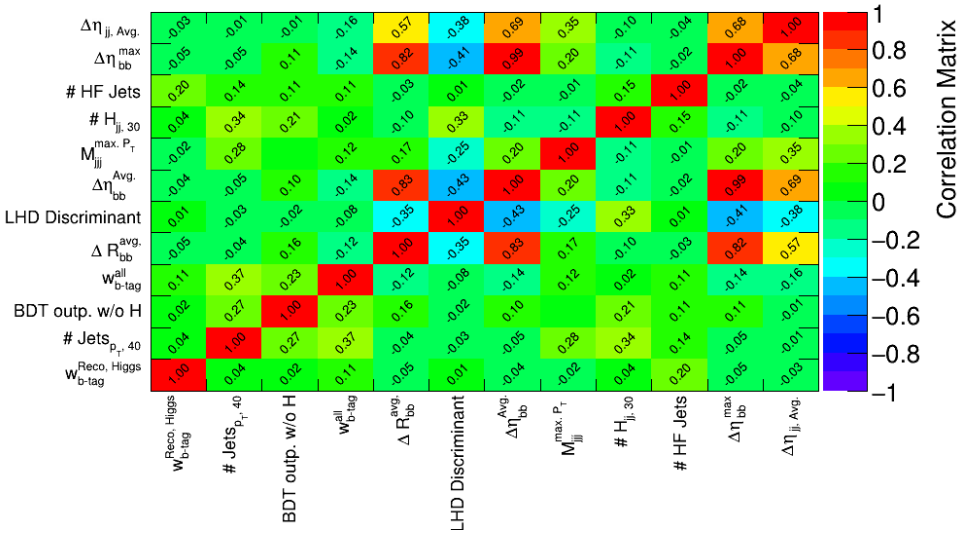

Figure 5.10.: Correlation matrix plot for the multi-class NN input variables in the 6 jet inclusive region of the larger dataset, including the tH sample.

5.4.1. Event Selection of the Larger Dataset

In the 6 jet region of the new dataset, the number of Monte Carlo events was increased by factors of up to ~ 200 for $tt + c$ events in comparison to the number of events in the 6 jet inclusive combined region available in the small dataset. In comparison to the 5 jet combined region, the number of Monte Carlo events increased by factors up to ~ 400 reached in $tt + c$ events, cf. Table 5.3. The event yield for each class is shown in Table B.4.

Since the NN uses a different, newer and larger dataset, the `mv2` b-tagging variables are replaced with the newer `DL1r` b-tagging variables. The `jet_mv2_order_N_tagWeightBin`

variables, see Appendix B.1 are replaced with their corresponding `jet_tagWeightBin_DL1r_Continuous[N]` variables, where `N` indicates the N$^{\text{th}}$ largest jet b-tagging discriminant. The newer `DL1r` variables include information, generated in a more sophisticated way using NNs.

The cuts on the transverse momenta p_{T} of the jets and on the absolute value of the pseudorapidity $|\eta|$ are analogous to those of the small dataset described in Section 5.1.

Region	#leptons	#jets	# b-tags at the 85% WP
6 jet$^{\geq 6j}_{\geq 4b}$	=1	≥ 6	≥ 4
5 jet$^{\geq 6j}_{\geq 4b}$		$= 5$	≥ 3

Table 5.4.: The non-overlapping 6 jet inclusive and 5 jet event regions of the large dataset. The regions are defined on the number of jets and b-tagged jets at 85% b-tag WP.

5.4.2. Performance Results of the Neural Networks on a Larger Dataset

In this section, the performance results for the 3-class, 4-class, and 5-class NNs trained and tested on the 6 jet inclusive and 5 jet region, see Table 5.4, are presented. In the following, they are referred to as 5 jet and 6 jet NNs. Lastly, the performance results of a binary classification NN, trained and tested on the 6 jet inclusive region, are presented.

Performance Results of the 3-class Neural Networks

In the 3 class NNs, the tt + jets background, which consists of tt+b, tt+c, and tt+light jets, is classified in addition to the two signal modes, ttH and tH.

As can be expected, the 3-class NN trained and tested on the 5 jet region classifies a larger fraction of tH events more accurately, i.e. with a higher tH classifier output value, than the one trained on the 6 jet inclusive region, cf. Figure 5.11. In classifying tH events, the Train/Test plots of the 5 jet NN show overtraining at classifier thresholds above 0.5. In contrast to that, the 6 jet NNs show smaller Train/Test ratios at the same classifier level.

The tt + jets classification distributions show more tt + jets events correctly classified, than other events falsely classified in the NNs in both regions.

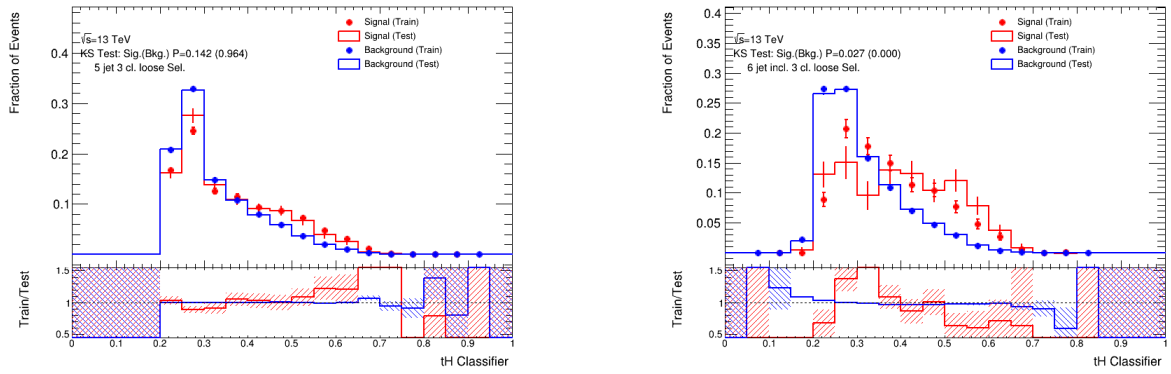

Figure 5.11.: Representative tH Train/Test plots for the 5 jet 3-class NN (left) and the 6 jet 3-class NN (right).

The 6 jet NN yields better performance in classifying ttH events than the 5 jet NN. Here, larger fractions of ttH events are classified as such, corresponding to classifier output values ≥ 0.5, as shown in Figure 5.12. In the ttH classifier output of the 6 jet NN, the ttH signal distribution peaks at a higher ttH classifier output threshold than the background distributions. This shift in distributions is validated by the confusion plot, as shown in Figure 5.12. The KS test P-values can not be considered here, as they indicate large differences in training and testing, which are not observed in the Train/Test ratio plot. No overtraining can be observed. The increased performance in the 6 jet region as opposed to the 5 jet region is further supported by the AUCs of the ROC curves, as can be seen in Table B.9.

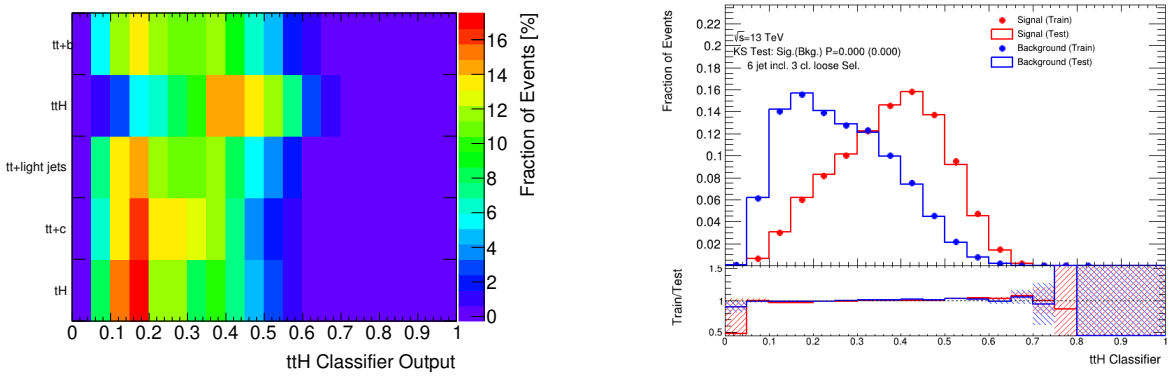

Figure 5.12.: Representative confusion (left) and Train/Test (right) plot of the 6 jet 3-class NN, trained on the larger dataset.

While the average AUC of the tH ROC curve of the 5 jet NN, 0.5903, is smaller than the the one of the 6 jet NN, 0.6918, it fluctuates less. This is analogous to the tH ROC curve of the 5-class NN operating on the smaller dataset, as seen in Figure 5.8. This

is caused by the missing tH statistics in the 6 jet inclusive region. In agreement with the Train/Test plots, displayed in Figure A.3 in the Appendix, the ttH, and $tt+$ jets classification performances are increased in NNs trained on the 6 jet region rather than on the 5 jet region.

Cuts on the classifier thresholds of each class can be applied to filter out the desired number of events in the corresponding class. These cuts can be determined using the SoSB plots.

By taking into account the SoSB ratio plots, in Figure 5.13, the classification of tH events in the 5 jet and 6 jet 3-class NNs is not advised. For both regions, the SoSB ratio for tH events and background is smaller than 1, suggesting that the tH signal can not be differentiated from background noise.

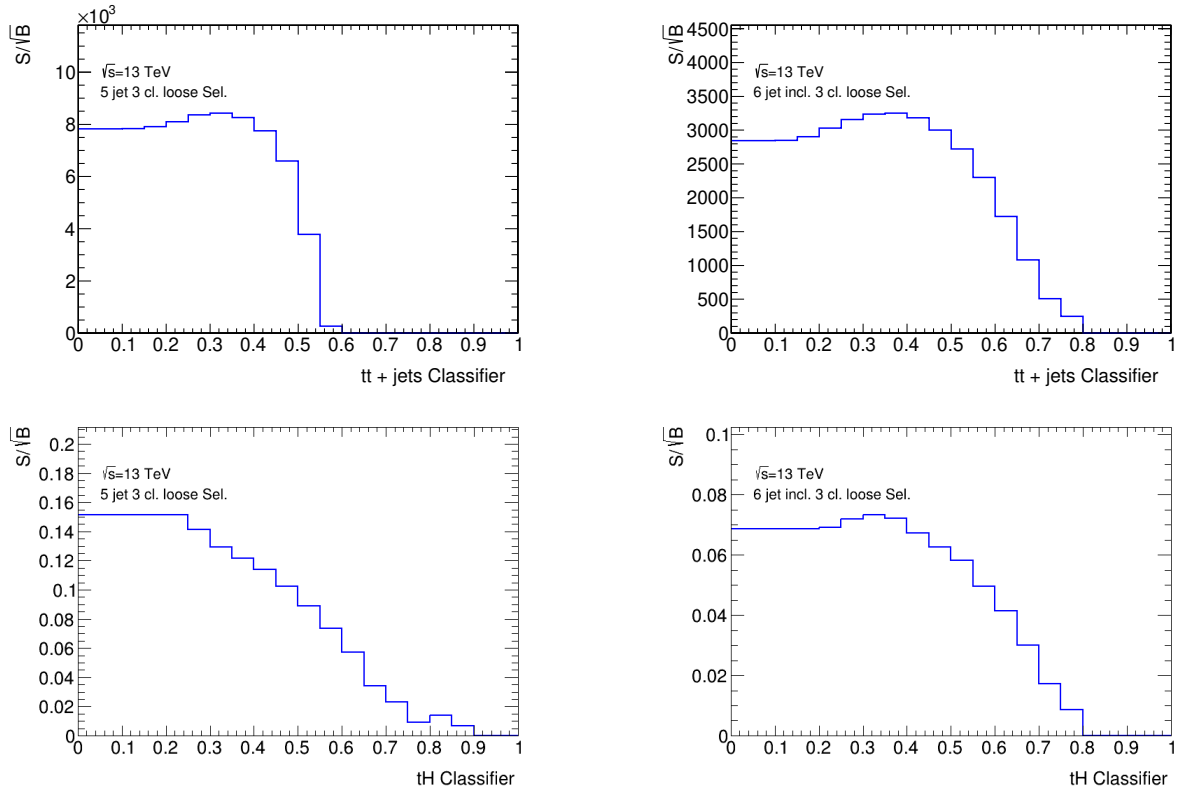

Figure 5.13.: SoSB ratio plots for the $tt+$ jets (top) and tH classification of the 5 jet (left) and 6 jet (right) 3-class NN.

The SoSB ratio for both NNs is best in the $tt+$ jets classification, peaking at a classifier output of around 0.35 with an SoSB ratio of ~ 8.5 in the 5 jet region and $\sim 3,200$ in the 6 jet inclusive region, cf. Figure 5.13.

Neither the 5 jet 3-class NN nor the 6 jet 3-class NN can minimise the categorical cross entropy loss in training significantly, as can be seen in the representative loss curve and accuracy curve, see Figure 5.14.

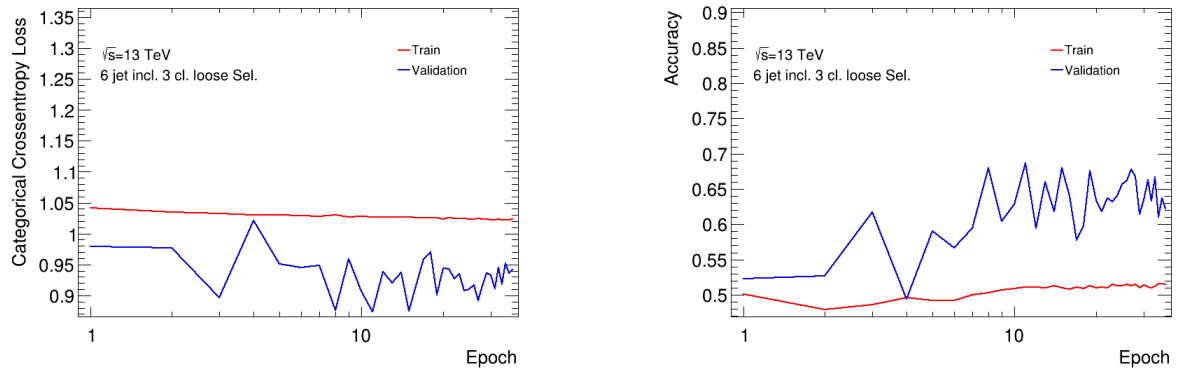

Figure 5.14.: Representative categorical cross entropy loss (left) and accuracy (right) curve for the 3-class NNs, trained and validated on the 6 jet inclusive region.

Analogous to the NNs trained on the small dataset, the early stopping mechanism was triggered in each learning process with the validation loss underestimating the training loss, as shown in Figure 5.14.

The accuracy in the 3-class NNs fluctuates by $\sim 20\%$, reaching a final accuracy of about 64%.

Overall, the ttH signal classification is improved with respect to the 5-class NN trained and tested on the small dataset using the 6 jet inclusive combined region.

The 6 jet 3-class NN performed better in the classification of ttH, and tt + jets events than the one used on the 5 jet region.

Performance Results of the 4-class Neural Networks

The 4-class NNs, analogous to the 3-class NNs, separate the two signal modes, ttH and tH, from the background events, but classify the background further into $tt + b$, and tt + light jets and c. Merging the tt + light jets and $tt + c$ classes is motivated by their confusion in the 5-class NNs, see Figure A.7.

The tt + light jets and c confusion plot for the 5 jet 4-class NN shows large fractions of $tt + c$ and tt + light jets events being classified as tt + light jets and c. For the 6 jet inclusive region, the NN exhibit a worse performance in classifying tt + light jets and c

events as such, as shown in Figure 5.15. Here, the confusion of the tt + light jets and c class with the corresponding background classes is large, all having similar classification distributions.

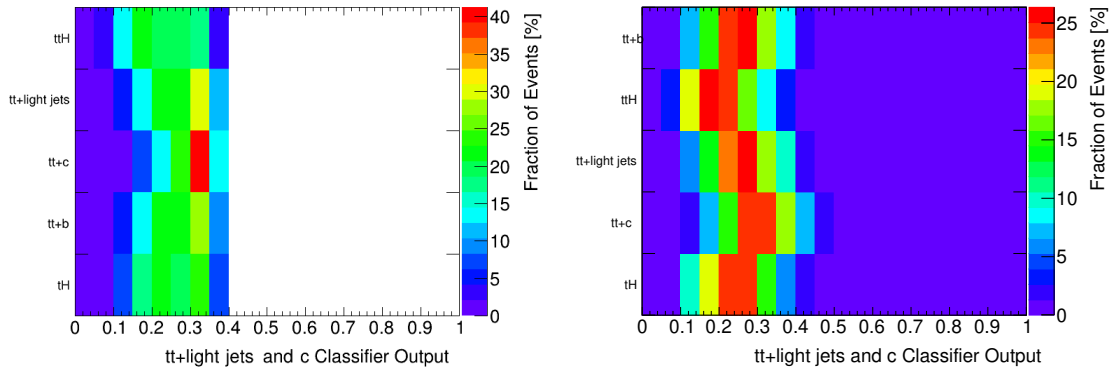

Figure 5.15.: tt + light jets and c confusion plots for the 5 jet 4-class (left) and 6 jet 4-class (right) NN.

The confusion plots for the NN operating on the 5 jet region show no classification of events with classifier thresholds above 0.5. At the same time, the 6 jet NN shows classifications being made up to classifier thresholds of 1, as can be seen in Figures 5.15 and A.4.

The ttH confusion plots show a similar performance for the 5 jet NN and 6 jet NN and the tH confusion plots show comparable performances in both regions too, as seen in Figure A.4,. The distributions for the ttH signal are similar in both regions. By applying a classifier cut at a level of 0.2, a significant fraction of background events is cut out. The same applies to the tH class at a classifier cut of 0.4, see Figure A.4.

With tH SoSB ratios peaking at values significantly smaller than 1, the tH classification capabilities of the NNs in both regions are not sufficient, due to low tH event yields.

The SoSB ratios of the $tt + b$, and tt + light jets and c classes are significantly higher compared to that of the ttH classification, but they converge to 0 for classifier outputs ≥ 0.45 in the 5 jet region, and ≥ 0.55 in the 6 jet inclusive region, see Figure 5.16 for a representative example. The classification of events into these classes is not advised, as classifications can not be made with high purity by the NNs.

The ROC curves of the NNs trained on both regions confirm the NNs' low performance in classifying tH, $tt + b$, and tt + light jets and c events. An AUC hardly above 0.5 can be observed for the ROC curves of the $tt + b$, and tt + light jets and c classifiers in both regions.

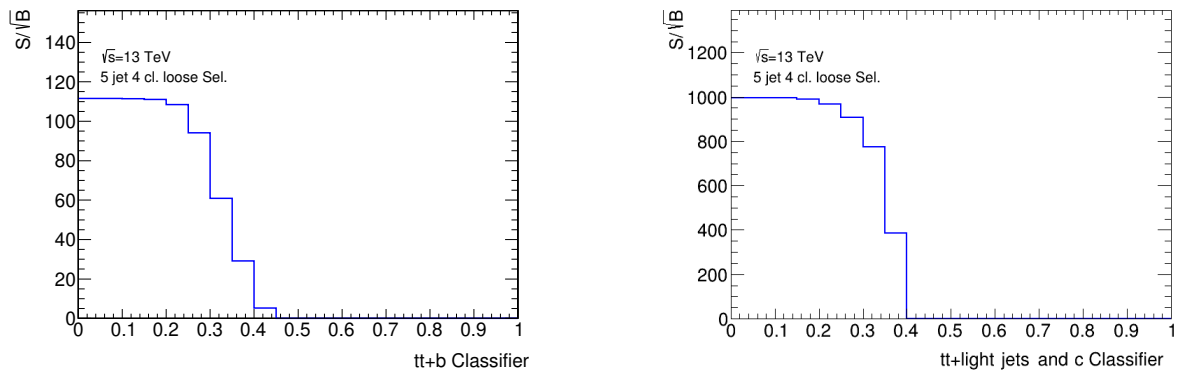

Figure 5.16.: $tt + b$ (left) and $tt +$ light jets and c (right) SoSB ratio plots for the 5 jet 4-class NN

Analogous to the previously described multi-class NN performances, the tH ROC curve is not well defined, due to the missing statistics of tH signal in the 6 jet inclusive region. While the tH statistics are improved in the 5 jet region, compared to the 6 jet inclusive region, an AUC just exceeding 0.6 is observed.

Due to the limited selection of variables optimised for ttH classification, the associated classifier outperforms the other classifications. With average AUC values of 0.7193 in the 5 jet, and 0.7287 in the 6 jet inclusive region, the ttH AUCs are larger than the AUCs of the ROC curves for other output classes.

Both NNs trained on the two different regions are neither able to minimise the categorical cross entropy loss nor to improve on accuracy, see Figure 5.17. The final accuracy is around 35%, which, by considering the number of output classes, is not a significant improvement compared to a random classification of events. Early stopping was triggered in both NNs.

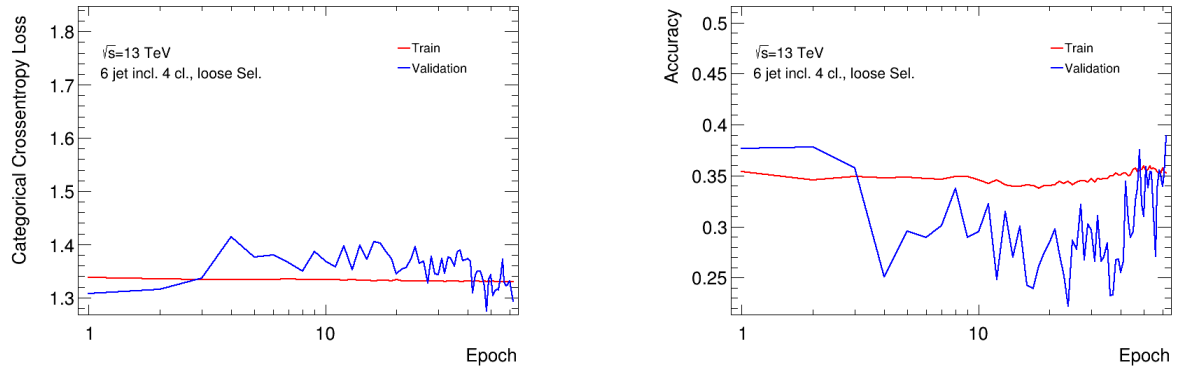

Figure 5.17.: Representative categorical cross entropy loss (left) and accuracy (right) curve for the 4-class NNs, trained and validated on the 6 jet inclusive region.

Performance Results of the 5-class Neural Networks

In the following, the classification performance of the 5 jet and 6 jet 5-class NNs, trained on the larger dataset is presented. A comparison to the 5-class NN trained and tested on the small dataset is made in Section 6.3.

Analogous to the multi-class NNs presented in previous sections, the 5-class NNs show decreased classification capabilities of tH events, due to low event yields and insufficient classification, cf. Sections 5.4.2 and 5.4.2. This is supported by the corresponding SoSB ratio plots, in Figures A.7 and A.8 in the Appendix.

The confusion plot for the 5 jet 5-class NN shows a similar ttH classification performance to the 4-class NN operating on the same region, see Figure A.6. Analogous, the confusion plot for the tt+c classifier of the 5 jet 5-class NN, is similar to that of the tt+light jets and c classifier of the 5 jet 4-class NN, cf. Figures A.4 and A.7.

The 6 jet 5-class NN shows less confusion of the $tt+c$ events with other classes, as shown in Figure 5.18, when compared to the tt+light jets and c confusion plot of the 6 jet 4-class NN.

In both NNs operating on the 5 jet and 6 jet inclusive regions, the confusion plots indicate significant fractions of tH, ttH and $tt+c$ events being classified as $tt+b$ events with approximately the same classifier thresholds as the signal events themselves, see Figure 5.18.

Overall, it is observed that the 6 jet 5-class NN performs best in classifying ttH and $tt+c$ events, which is further supported by the SoSB ratio plots, indicating identifiable peaks in

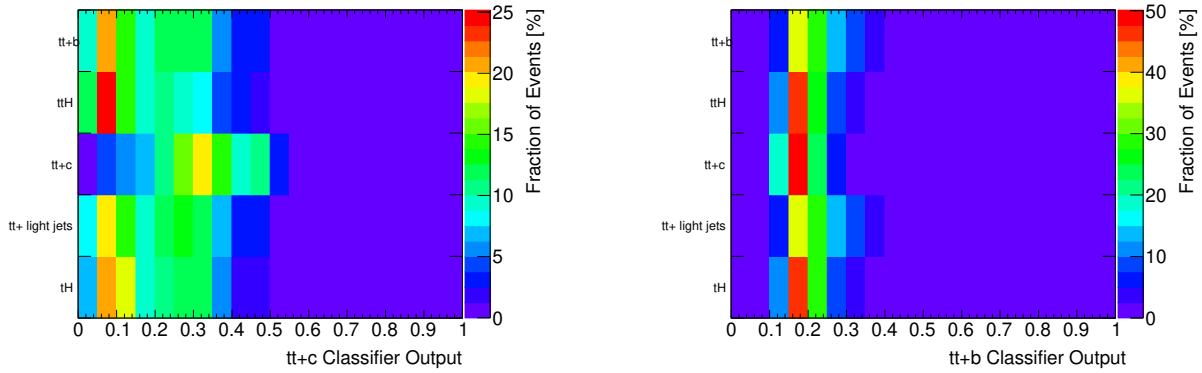

Figure 5.18.: $tt + c$ (left) and $tt + b$ (right) confusion plots of the 6 jet 5-class NN.

the SoSB ratios, cf. Figures A.7 and A.8. The other SoSB ratios do not show identifiable peaks in SoSB ratio or, on average, are smaller than 1, meaning the respective signal can not be distinguished from background noise.

The AUCs of the ROC curves of both 5-class NNs compound this performance evaluation. The AUCs of the $tt+b$ and $tt+$light jets ROC curves are, on average, below 0.6, indicating a marginally better classification performance than random classification, see Table B.11.

The SoSB ratio plots of the 5 jet 5-class NN exhibit similar distributions to the ones of the 6 jet NN. At the same time the AUCs of the ROC curves of the 5 jet NN indicate a worse classification performance in all output classes except for the tH output class, as can be seen in Table B.11 in the Appendix.

Neither the 5 jet 5-class NN nor the 6 jet 5-class NN is able to minimise the loss function significantly, see A.8. Thus, it is indicated that the local minimum that was found, results in no performance increase compared to a random classification of events. The early stopping mechanism was triggered in each learning process.

The observed classification performance is expected given the similar classification performance of the 4-class NNs.

Performance Results of the binary classification Neural Network

A binary classification NN, similar to the ones trained on the small dataset, is used for the classification of events from the larger dataset into ttH or background events.

The NN shares the same hyperparameters with the simpler binary classification NNs. In addition to that, it uses the same kinematic variables and variables from the likelihood calculations as inputs. The correlation matrix is shown in Figure A.9 in the Appendix.

The performance of the 6 jet binary classification NN can be compared to the BDT's performance in classifying events of the region with the loosest selection, requiring ≥ 6 jets including ≥ 4 b-jets at the 85% WP. Analogous to the BDTs used in previous analyses, the tH signal is neglected.

The confusion plot in Figure 5.19, shows more than 11% of ttH events being classified as such with binary classifier output values ~ 0.7. Significant fractions of background events to the ttH class are classified as such with binary classifier outputs ≤ 0.5.

The Train/Test ratio shows no signs of overtraining, as it is approximately 1 in the binary classifier threshold interval [0.1,0.9]. The loss function indicates no undertraining, cf. Figure 5.20.

The ROC curve in Figure 5.20, is well defined, suggesting no numerical errors when computing it. The ttH event yields are large enough for the NN to be able to separate signal from background fluctuations, as indicated by an SoSB ratio > 1, which peaks at a NN output threshold of ~ 0.6. This shows a confident classification of signal events. It is advised to apply a threshold of 0.6 to the NN's output to gain the best SoSB ratio. Then, the ROC curve can be used to fine tune the trade-off between the true positive rate and true negative rate. With an average AUC of 0.755, the NN's performance on the region with the loosest selection considered in this thesis, is comparable to that of the BDTs used in previous analyses validated on a region requiring a tighter selection.

The binary classification NN is not able to minimise the loss function for training by a significant amount. Thus, the accuracy only improved marginally compared to the initial accuracy, see Figure 5.20. The validation loss is underestimated by the training loss and the learning process was stopped early by the early stopping mechanism.

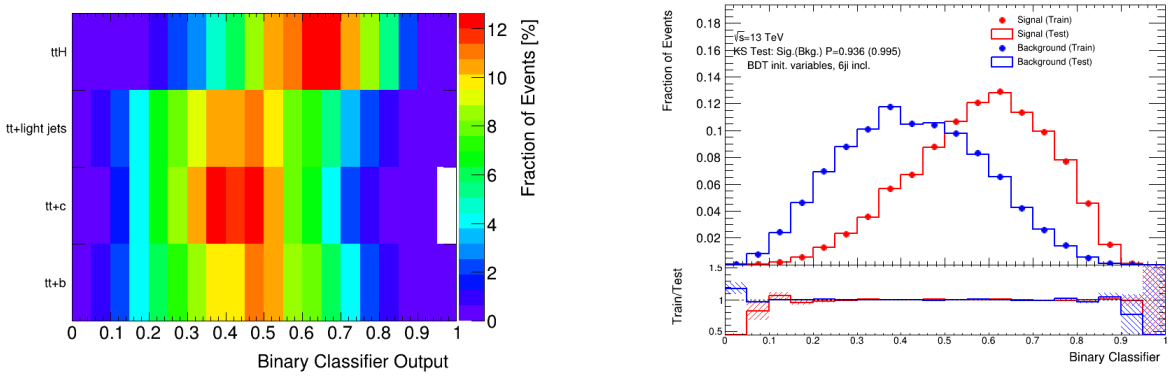

Figure 5.19.: Representative confusion (left) and Train/Test (right) plot of the 6 jet binary classification NN, trained on the larger dataset.

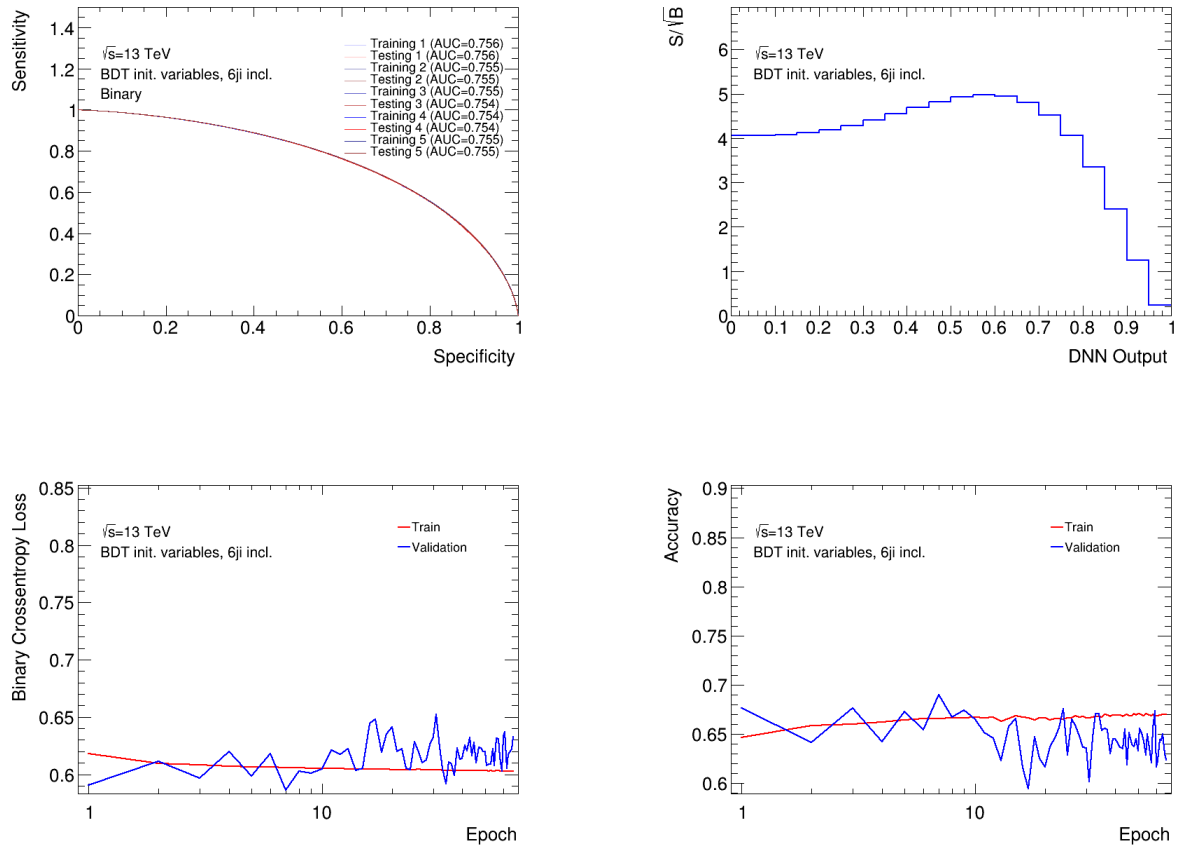

Figure 5.20.: ROC curve (top left), SoSB ratio (top right), representative binary cross entropy loss (bottom left) and representative accuracy (bottom right) plots of the binary classification NN, trained and validated on the large dataset.

6. Discussion of the Results

In this Chapter, the classification performance of the NNs is compared to that of the BDTs.

In Section 6.1, the binary classification NNs trained and validated on the 6 jet inclusive combined region are compared to the BDTs from previous analyses trained on the same region. In Section 6.2, the BDT's performance on the 6 jet inclusive region of the larger dataset requiring ≥ 4 b-jets at 85% WP is compared to a binary classification NN trained with similar input variables on a comparable dataset. In Section 6.3, the performance of the multi-class NNs used in this thesis is evaluated.

6.1. Performance Comparison of the Boosted Decision Trees and Neural Networks Trained on the Small Dataset

In Section 5.2, it was shown that the BDT trained and validated on the 6 jet inclusive combined region requiring ≥ 4 b-jets at 70% WP, was the best performing BDT in that region with an AUC of 0.758. When validated on the 6 jet inclusive region, requiring ≥ 4 b-jets at 60% WP, delivered the overall largest AUC of 0.761. These classification performance results are compared to the performance results of the binary classification NNs trained on the same dataset and regions, described in Section 5.3.3.

The average AUCs of the binary classification NNs are shown in Table B.7 in the Appendix. It is observed that none of the training nor testing processes during the 5 folds have a lower AUC than that of the BDT. The performance increases of the NN in each region, with respect to the BDTs, are shown in Table 6.1.

The overall best performance of a binary classification NN was observed when trained and tested on the 5 jet high region.

NN	Performance increase [%]		
	avg. training	avg. testing	avg. overall
simple NN, 6 jet incl.	0.53	0.45	0.49
complex NN, 6 jet incl.	0.95	0.71	0.83
simple NN, 6 jet high	1.08	0.97	1.03

Table 6.1.: Average performance increase, based on the AUCs of the ROC curves, of the NNs with respect to the BDTs used in previous analyses.

In conclusion, an increase in performance when using a binary classification NN instead of a BDT in the probed regions can be observed. Especially in the region with a tight b-tag requirement, the NN performs about 1% better. The more simple NNs increase performance with respect to the BDTs used in previous analyses, although the performance increase in terms of AUC of the ROC curve is smaller than that of the complex NN. No hyperparameter optimisation was attempted in this thesis. By optimising the hyperparameters one might gain an even bigger increase in performance.

6.2. Performance Comparison of the Boosted Decision Trees and Neural Network Trained on the Large Dataset

A binary classification NN was trained on the 6 jet inclusive region of the larger dataset, see Section 5.4.2, using the same kinematic and likelihood variables as the BDTs used in previous ATLAS analyses. The b-tagging variables are switched to the `DL1r` ones since the `mv2c10` variables are not available in that dataset.

In the following, the tight region is referred to the region requiring ≥ 6 jets including $\geq 4b$-jets at 60% WP. The firm region requires ≥ 6 jets including $\geq 4b$-jets at 70% WP. The loose region is defined as the region requiring ≥ 6 jets including $\geq 4b$-jets at 85% WP.

When comparing the correlation matrices of the NNs from the old and the new larger dataset, it is observed that the correlations are similar to each other. Differences of less than $\pm 10\%$ of the total linear correlation are observed. Here, the different selections need to be taken into account to explain the differences. The correlation matrix for the exact BDT input variables is calculated for the firm region, see Figure 5.1. The firm region is selection-wise closest to the updated input variables. Here, the correlation matrix in

Figure A.9 of the Appendix is calculated for the loose region.

As a result the different datasets and input variables are comparable with each other. Thus, the binary classification performances of the BDTs and the binary classification performance of the NN can be compared. The BDT trained on the loose region and evaluated on the events of the firm region lead to ROC curves with an AUC of 0.755. When validated on the tight region an AUC of 0.757 was reported. The binary classification NN, trained and tested on the loose region produces a ROC curve with an average AUC of 0.755 during training and testing. When only comparing the AUCs, this results in an equal performance of the NN and BDT trained on the loose region and tested on the firm region, and a marginal performance decrease, when compared to the BDT trained on the loose region, but validated on the tight region. The NN trained on the loose region should be tested on the regions with the tighter cuts to evaluate differences in performance.

These results do not necessarily indicate an overall worse classification performance. The BDTs trained on the firm region and validated on the tight region increased in AUC. A similar behaviour might be observed for the NN trained on the loose region.

In addition to that, the comparison is only partially valid, as two different datasets were used. Although the correlation matrices were comparable, the newer `DL1r` can still lead to different performance results. To check possible performance differences caused by the different input samples, the BDTs have to be used on the larger dataset and performances need to be compared.

Lastly, no hyperparameter optimisation was done. The hyperparameters of the NN were chosen adequately for the classification task, but were not optimised. As the results in Section 6.1 show, the simpler models underfit the data slightly. A performance increase is likely when increasing the complexity of the NNs.

6.3. Performance Evaluation of the Multi-Class Neural Networks

It is observed that the 6 jet 5-class NN of the larger dataset performed better than the one used on the 6 jet combined region of the small dataset overall. The performance increase in terms of AUCs for each output class can be seen in Table 6.2.

It has to be noted that the NNs show classifications for large fractions of events from classes other than ttH with classifier thresholds significantly below 0.5, indicating that

	tH	ttH	$tt+$ light jets	$tt+c$	$tt+b$
Performance increase [%]	-2.13	0.27	9.96	0.20	8.42

Table 6.2.: Average performance increase, based on the AUCs of the ROC curves, of the NNs trained on the larger dataset with respect to the NNs trained on the smaller dataset.

the NNs do not confidently classify the events from these classes as such.

A significant difference in confusion plots from the 5-class and 4-class NNs can be observed. The intent of reducing the number of classes by merging the $tt+$light jets and $tt+c$ classes, as these processes are thought of leaving similar variable distributions, resulted in large confusions of the 4-class NN in all input classes, as shown in Figure 5.15.

The 3-class NNs are the only NNs that show a significant over- or undertraining. It is observed in the tH event classification. While the SoSB ratio for the tH classifier is always significantly smaller than 1 it shows a small peak at a high classifier threshold of ~ 0.85. These phenomena can not be explained and need to be investigated further. The classification performance can be tested further by employing a tH vs. background binary classification NN on the larger dataset. It is to be evaluated if this binary classification NN validates these observations.

In general, a worse overall multi-class classification in the 5 jet region compared to the 6 jet inclusive region was observed. Only the tH classification was improved in the 5 jet region due to increased statistics which are low in the 6 jet region.

This can be explained by the restriction of available input variables to the NNs trained and validated on the 5 jet region. Here, the ttH output variables of the reconstruction BDT can not be used.

Overall, the lack of classification performance in classes different to ttH of the multi-class NNs can be explained by the bias towards the input variables. These were originally chosen for the ttH vs. background classification and not meant to be used for the multi-class classification. The classifications of the multi-class NN are hardly better than random classification. Thus, the initial idea of classifying the background further by splitting $tt+b$ into $tt+1b$, $tt+2b$, and $tt+\geq 3b$ was discarded.

Hyperparameter tuning is another source of inefficiency. Only two different sets of hyperparameters are used in this analysis. It is to be checked if the number of trainable parameters needs to be increased drastically for better multi-class classification results.

7. Conclusion and Outlook

In the first part of this thesis the ttH vs. $tt+$jets binary classification performance of 6 NNs using the same dataset as the BDTs from previous analyses was presented. Here, the classification performances of the simple and complex NN are improved compared to the BDT's performance. Using a larger dataset with enlarged statistics, but looser selection, requiring ≥ 6 jets with ≥ 4 b-jets, did not improve the classification results. Thus, it is indicated that the hyperparameters and the input variable selection have to be optimised.

It was also shown that the more complex NNs, having more trainable parameters than the others, performed best.

For the best ttH vs. background binary classification results, it is advised to train a NN using the hyperparameters of the complex NNs on the 6 jet inclusive combined region and evaluating its performance on the other regions. After an in-depth analysis of the performance, hyperparameter optimisation is recommended.

The multi-class NNs, using a selection of input variables optimised for the ttH vs. background binary classification, did not perform sufficiently well to be used in upcoming analyses regardless of the dataset used. The NNs were not able to minimise the loss function and improve in the applied evaluation metrics significantly.

Initially a better multi-class classification performance was anticipated, but a further splitting of the background, such as classifying events via the number of observed b-jets, is not possible given the current performance.

For improving on the multi-class classification performance, it is recommended to use variables less biased towards the classification of subsets of the output classes. Thus, classification performance in all output classes could be optimised.

By implementing neural networks in ttH classification a better classification performance, compared to the previously used boosted decision trees, can be achieved. The improved classification algorithm can be used in future analyses to support the event selection and background reduction, which then allows for further Higgs boson analyses.

Bibliography

[1] S. Abachi, et al. (D0 Collaboration), *Observation of the Top Quark*, Phys. Rev. Lett. **74**, 2632 (1995).

[2] F. Abe, et al. (CDF Collaboration), *Observation of Top Quark Production in $\overline{p}p$ Collisions with the Collider Detector at Fermilab*, Phys. Rev. Lett. **74**, 2626 (1995).

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

[4] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B **716(1)**, 30 (2012).

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

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

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

[8] ATLAS Collaboration, *Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector*, Phys. Lett. B **784**, 173 (2018).

[9] CMS Collaboration, *Observation of $t\bar{t}H$ Production*, Phys. Rev. Lett. **120**, 231801 (2018).

[10] C. N. Yang, R. L. Mills, *Conservation of Isotopic Spin and Isotopic Gauge Invariance*, Phys. Rev. **96**, 191 (1954).

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

[12] S. L. Glashow, *Partial-symmetries of weak interactions*, Nucl. Phys. **22(4)**, 579 (1961).

[13] A. Salam, *Weak and Electromagnetic Interactions*, Conf. Proc. C **680519**, 367 (1968).

[14] R. N. Mohapatra, R. E. Marshak, *Local $B-L$ Symmetry of Electroweak Interactions, Majorana Neutrinos, and Neutron Oscillations*, Phys. Rev. Lett. **44**, 1316 (1980).

[15] P. Zyla, et al. (Particle Data Group), *Review of Particle Physics*, Prog. Theor. Exp. Phys. **2020(8)** (2020).

[16] D. Clowe, et al., *A direct empirical proof of the existence of dark matter*, Astrophys. J. Lett. **648**, L109 (2006).

[17] A. V. Gladyshev, D. I. Kazakov, *Supersymmetry and LHC*, Phys. Atom. Nucl. **70**, 1553 (2007).

[18] P. Langacker, *Grand unified theories and proton decay*, Phys. Rep. **72(4)**, 185 (1981).

[19] E. Witten, *Mass hierarchies in supersymmetric theories*, Phys. Lett. B **105(4)**, 267 (1981).

[20] S. P. Martin, *A Supersymmetry primer*, Adv. Ser. Direct. High Energy Phys. **18**, 1 (1998).

[21] ATLAS Collaboration, *Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector*, Phys. Lett. B **786**, 59 (2018).

[22] CMS Collaboration, *Observation of Higgs boson decay to bottom quarks*, Phys. Rev. Lett. **121(12)**, 121801 (2018).

[23] ATLAS Collaboration, *Observation and measurement of Higgs boson decays to WW^* with the ATLAS detector*, Phys. Rev. D **92(1)**, 012006 (2015).

[24] W. Beenakker, et al., *Higgs Radiation Off Top Quarks at the Tevatron and the LHC*, Phys. Rev. Lett. **87**, 201805 (2001).

[25] S. Dawson, et al., *Associated top quark Higgs boson production at the LHC*, Phys. Rev. D **67**, 071503 (2003).

[26] Z. Kunszt, *Associated production of heavy Higgs boson with top quarks*, Nucl. Phys. B **247(2)**, 339 (1984).

[27] J. N. Ng, P. Zakarauskas, *QCD-parton calculation of conjoined production of Higgs bosons and heavy flavors in $p\bar{p}$ collisions*, Phys. Rev. D **29**, 876 (1984).

[28] C. Englert, et al., *Precision Measurements of Higgs Couplings: Implications for New Physics Scales*, J. Phys. G **41**, 113001 (2014).

[29] A. D. Martin, et al., *Parton distributions for the LHC*, Eur. Phys. J. C **63**, 189 (2009).

[30] LHC Higgs Cross Section Working Group, D. de Florian, et al. (Eds.), *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, CERN Yellow Reports: Monographs, Vol. 2/2017, 2017.

[31] A. M. Turing, *I.-COMPUTING MACHINERY AND INTELLIGENCE*, Mind **LIX(236)**, 433 (1950).

[32] ATLAS, *Search for the standard model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Rev. D **97**, 072016 (2018).

[33] R. Pascanu, T. Mikolov, Y. Bengio, *Understanding the exploding gradient problem*, CoRR **abs/1211.5063** (2012).

[34] Z. Zhang, M. R. Sabuncu, *Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels*, CoRR **abs/1805.07836** (2018).

[35] F. Chollet, et al., *Keras*, `https://keras.io` (2015).

[36] D. P. Kingma, J. Ba, *Adam: A Method for Stochastic Optimization*, ICLR **abs/1412.6980** (2017).

[37] D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Learning representations by back-propagating errors*, Nature **323(6088)**, 533 (1986).

[38] A. N. Kolmogorow, *Sulla Determinazione Empirica di Una Legge di Distribuzione*, Giornale dell'Istituto Italiano degli Attuari **4**, 83 (1933).

[39] N. W. Smirnow, *Sur les Écarts de la Courbe de Distribution Empirique*, Recueil Mathématique (Matematiceskii Sbornik) **6**, 3 (1939).

[40] L. Evans, P. Bryant, *LHC Machine*, J. Instrum. **3(08)**, S08001 (2008).

[41] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, J. Instrum. **3(08)**, S08003 (2008).

[42] CMS Collaboration, *The CMS Experiment at the CERN LHC*, J. Instrum. **3**, S08004 (2008).

[43] ALICE Collaboration, *The ALICE experiment at the CERN LHC*, J. Instrum. **3**, S08002 (2008).

[44] LHCb Collaboration, *The LHCb Detector at the LHC*, J. Instrum. **3**, S08005 (2008).

[45] A. Bejar, et al. (Eds.)*High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*, CERN Yellow Reports: Monographs, Vol. 10/2020, 2020.

[46] ATLAS Collaboration, *ATLAS inner detector: Technical Design Report, 1*, Technical design report. ATLAS, CERN, Geneva, 1997, CERN-LHCC-97-016.

[47] B. Mindur, *ATLAS Transition Radiation Tracker (TRT): Straw tubes for tracking and particle identification at the Large Hadron Collider*, Nucl. Instrum. Methods Phys. Res. A **845**, 257 (2017).

[48] ATLAS Collaboration, *ATLAS pixel detector electronics and sensors*, J. Instrum. **3(07)**, P07007 (2008).

[49] ATLAS Collaboration, *Measurements of spatial resolution of ATLAS pixel detectors*, Nucl. Instrum. Meth. A **465**, 112 (2000).

[50] J. R. Pater, *The ATLAS SemiConductor Tracker operation and performance*, J. Instrum. **7(04)**, C04001 (2012).

[51] ATLAS Collaboration, *The ATLAS Inner Detector commissioning and calibration*, Eur. Phys. J. C **70**, 787 (2010).

[52] ATLAS Collaboration, *ATLAS calorimeter performance Technical Design Report*, CERN-LHCC-96-40 (1996).

[53] D.-E. Boumediene, *ATLAS Calorimeter: Run 2 performance and Phase-II upgrades*, PoS **EPS-HEP2017**, 485 (2017).

[54] ATLAS Collaboration, *ATLAS muon spectrometer: Technical Design Report*, Technical design report. ATLAS, CERN, Geneva, 1997.

[55] ATLAS Collaboration, *Technical Design Report for the Phase-II Upgrade of the ATLAS Muon Spectrometer*, CERN-LHCC-2017-017, ATLAS-TDR-026, CERN, Geneva, 2017.

[56] ATLAS Collaboration, *The ATLAS Pixel Detector Upgrade at the HL-LHC*, ATL-ITK-PROC-2020-002, CERN, Geneva, 2020.

[57] ATLAS Collaboration, *Technical Design Report for the ATLAS Inner Tracker Pixel Detector*, CERN-LHCC-2017-021, ATLAS-TDR-030, CERN, Geneva, 2017.

[58] ATLAS Collaboration, *Operation of the ATLAS trigger system in Run 2*, J. Instrum. **15(10)**, P10004 (2020).

[59] N. Berger, et al., *The ATLAS high level trigger steering*, J. Phys. Conf. Ser. **119**, 022013 (2008).

[60] S. Agostinelli, et al., *Geant4-a simulation toolkit*, Nucl. Instrum. Methods Phys. Res. A **506(3)**, 250 (2003).

[61] W. Lukas, *Fast Simulation for ATLAS: Atlfast-II and ISF*, J. Phys. Conf. Ser. **396**, 022031 (2012).

[62] P. Nason, B. Webber, *Next-to-Leading-Order Event Generators*, Ann. Rev. Nucl. Part. Sci. **62**, 187 (2012).

[63] G. Corcella, et al., *HERWIG 6: An Event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)*, J. High Energy Phys. **01**, 010 (2001).

[64] T. Sjöstrand, S. Mrenna, P. Skands, *PYTHIA 6.4 physics and manual*, J. High Energy Phys. **2006(05)**, 026 (2006).

[65] T. Gleisberg, et al., *SHERPA 1., a proof-of-concept version*, J. High Energy Phys. **2004(02)**, 056 (2004).

[66] ATLAS Collaboration, *Measurement of the Higgs boson decaying to b-quarks produced in association with a top-quark pair in pp collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, ATLAS-CONF-2020-058 (2020).

[67] ATLAS Collaboration, *Electron reconstruction and identification efficiency measurements with the ATLAS detector using the 2011 LHC proton–proton collision data*, Eur. Phys. J. C **74(7)**, 2941 (2014).

[68] ATLAS Collaboration, *Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton-proton collision data*, ATLAS-CONF-2016-024 (2016).

[69] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton-proton collision data at* $\sqrt{s} = 13$ *TeV*, Eur. Phys. J. C **76(5)**, 292 (2016).

[70] ATLAS Collaboration, *Reconstruction, Energy Calibration, and Identification of Hadronically Decaying Tau Leptons in the ATLAS Experiment for Run-2 of the LHC*, ATL-PHYS-PUB-2015-045 (2015).

[71] ATLAS Collaboration, *Jet energy scale and resolution measured in proton-proton collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, CERN-EP-2020-083 (2020).

[72] ATLAS Collaboration, *ATLAS b-jet identification performance and efficiency measurement with* $t\bar{t}$ *events in pp collisions at* $\sqrt{s} = 13$ *TeV*, Eur. Phys. J. C **79(11)**, 970 (2019).

[73] Y. Coadou, Personal Communication.

[74] V. Barger, J. Ohnemus, R. J. N. Phillips, *Event shape criteria for single-lepton top-quark signals*, Phys. Rev. D **48**, R3953 (1993).

[75] G. C. Fox, S. Wolfram, *Observables for the Analysis of Event Shapes in* e^+e^- *Annihilation and Other Processes*, Phys. Rev. Lett. **41**, 1581 (1978).

A. Further Plots

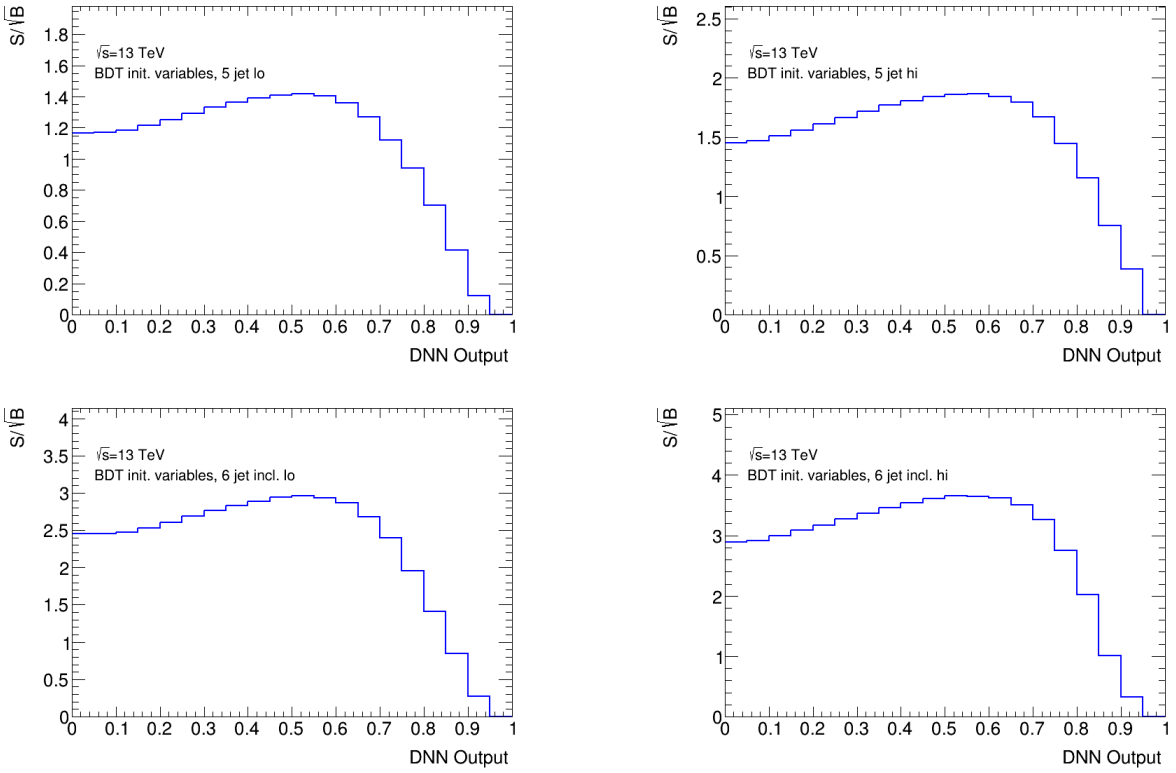

Figure A.1.: Signal over square root of background (SoSB) plots of the binary classification NNs, trained and validated on the 5 jet low (top left) and high (top right), 6 jet inclusive low (bottom left) and high (bottom right) regions, trained and validated on the 6 jet inclusive high and low region combined.

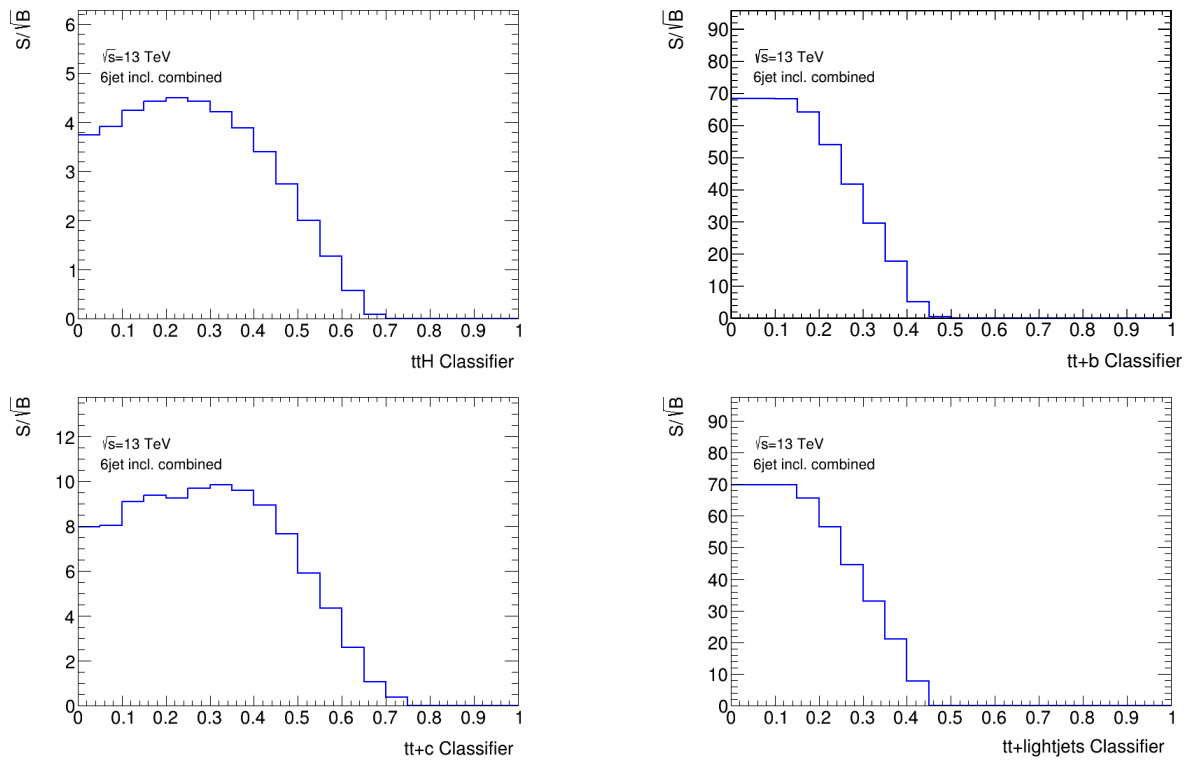

Figure A.2.: SoSB plots for 4 of the 5 output classes of the 6 jet multi-class NN using the smaller dataset. ttH (top left), $tt + b$ (top right), $tt + c$ (bottom left) and $tt +$ light jets (bottom right).

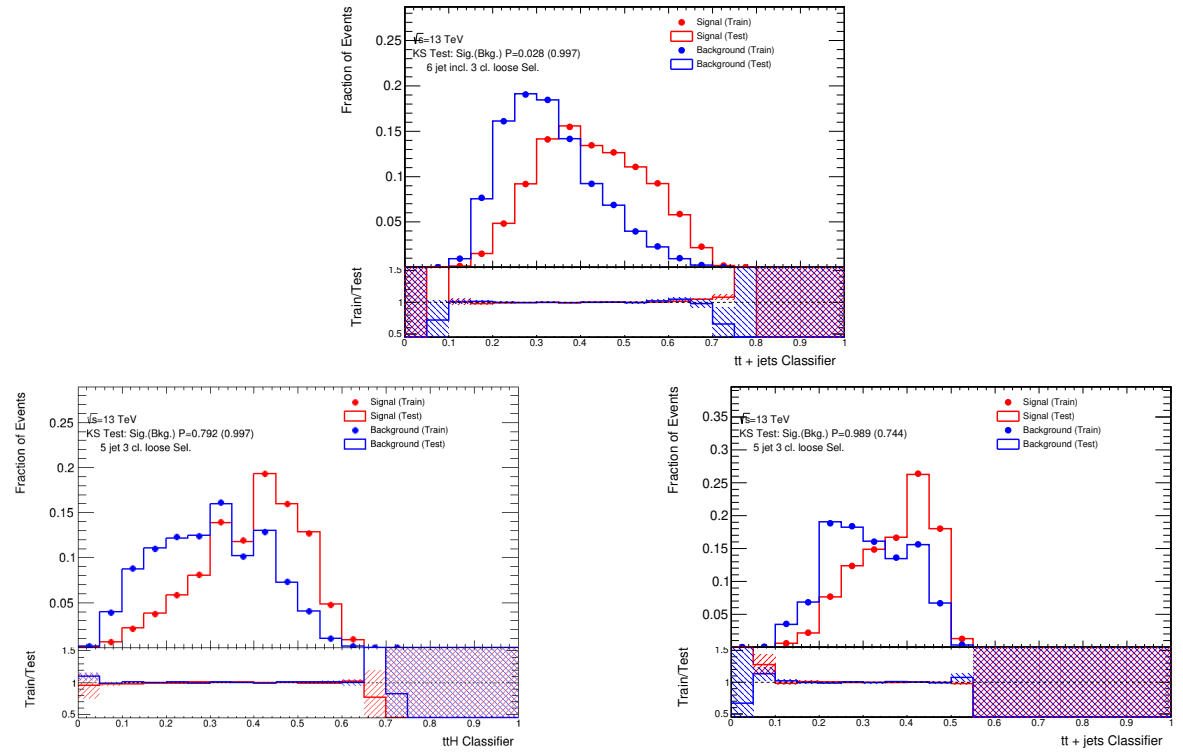

Figure A.3.: $tt +$ jets (top) Train/Test plot for the 6 jet, 3-class NN. ttH (bottom left), and $tt +$ jets (bottom right) classification plots for the 5 jet, 3-class NN trained.

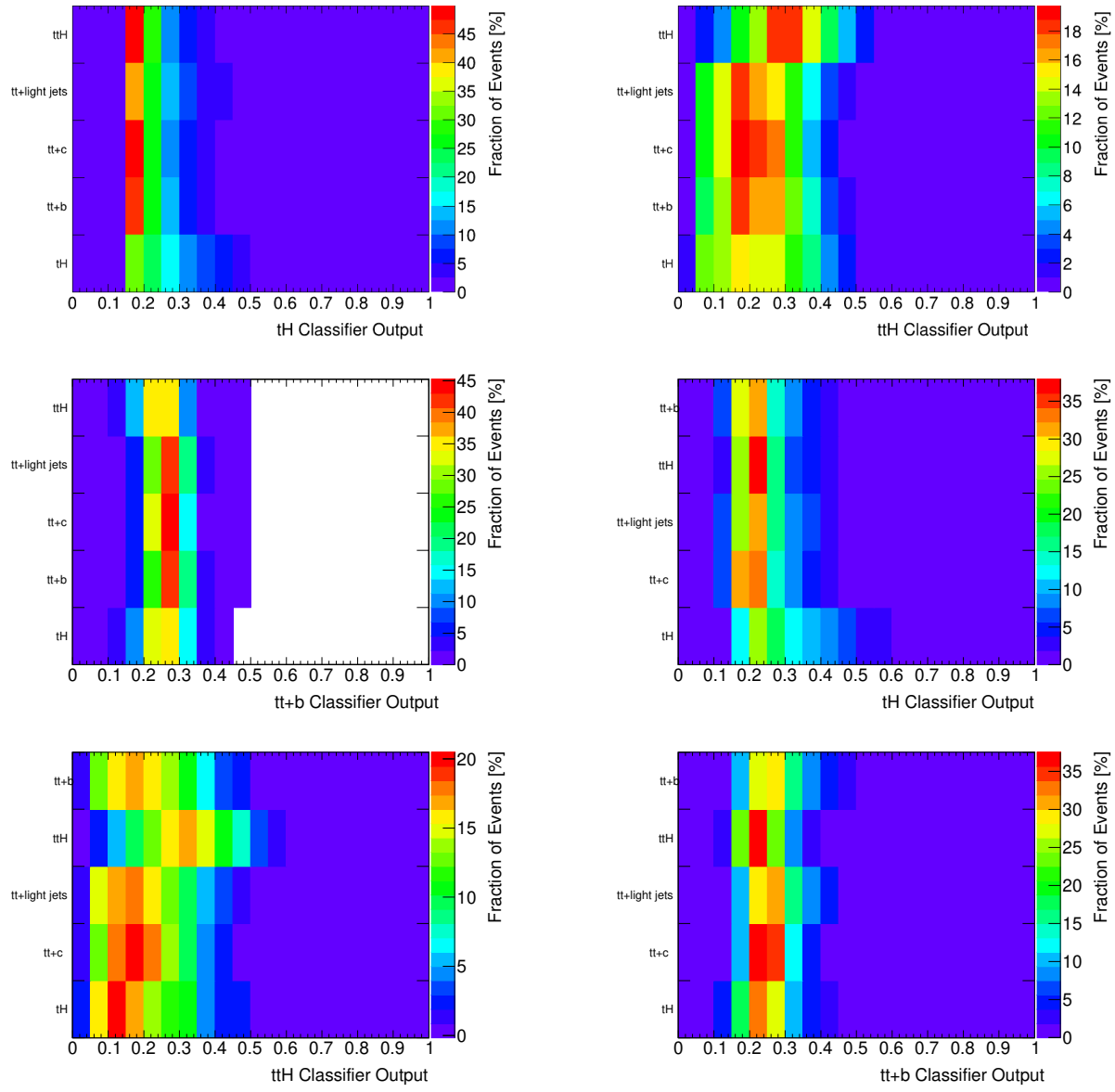

Figure A.4.: tH (top left), ttH (top right), $tt + b$ (middle left) confusion plots for the 4-class NN, trained and validated on the 5 jet region. tH (middle right), ttH (bottom left), $tt + b$ (bottom right) and confusion plots for the 4-class NN, trained and validated on the 6 jet inclusive region.

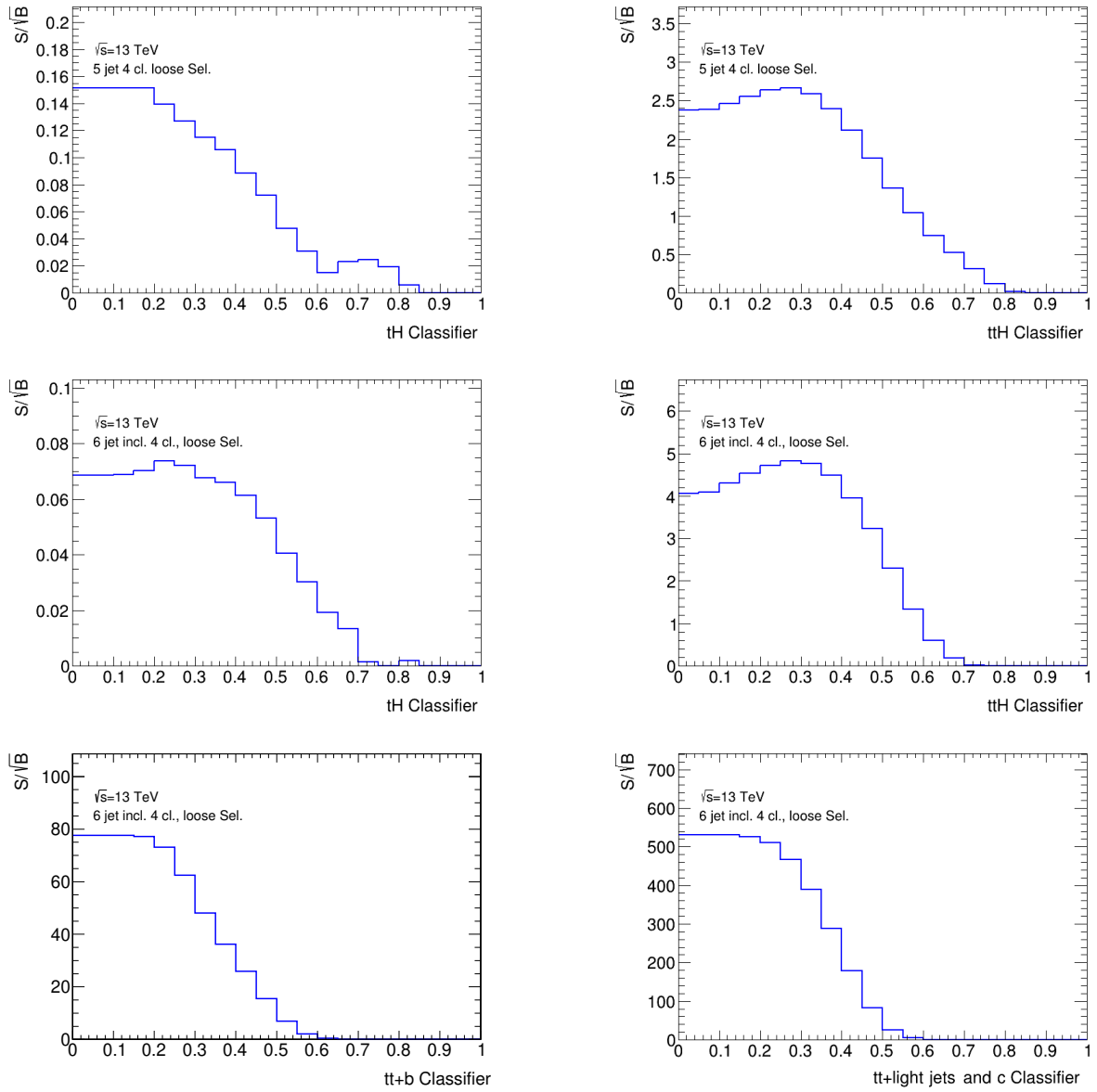

Figure A.5.: tH (top left), ttH (top right) SoSB ratio plots for the 4-class NN, trained and validated on the 5 jet region. tH (middle left), ttH (middle right), $tt + b$ (bottom left) and $tt +$ light jets (bottom right) SoSB ratio plots for the 4-class NN, trained and validated on the 6 jet region.

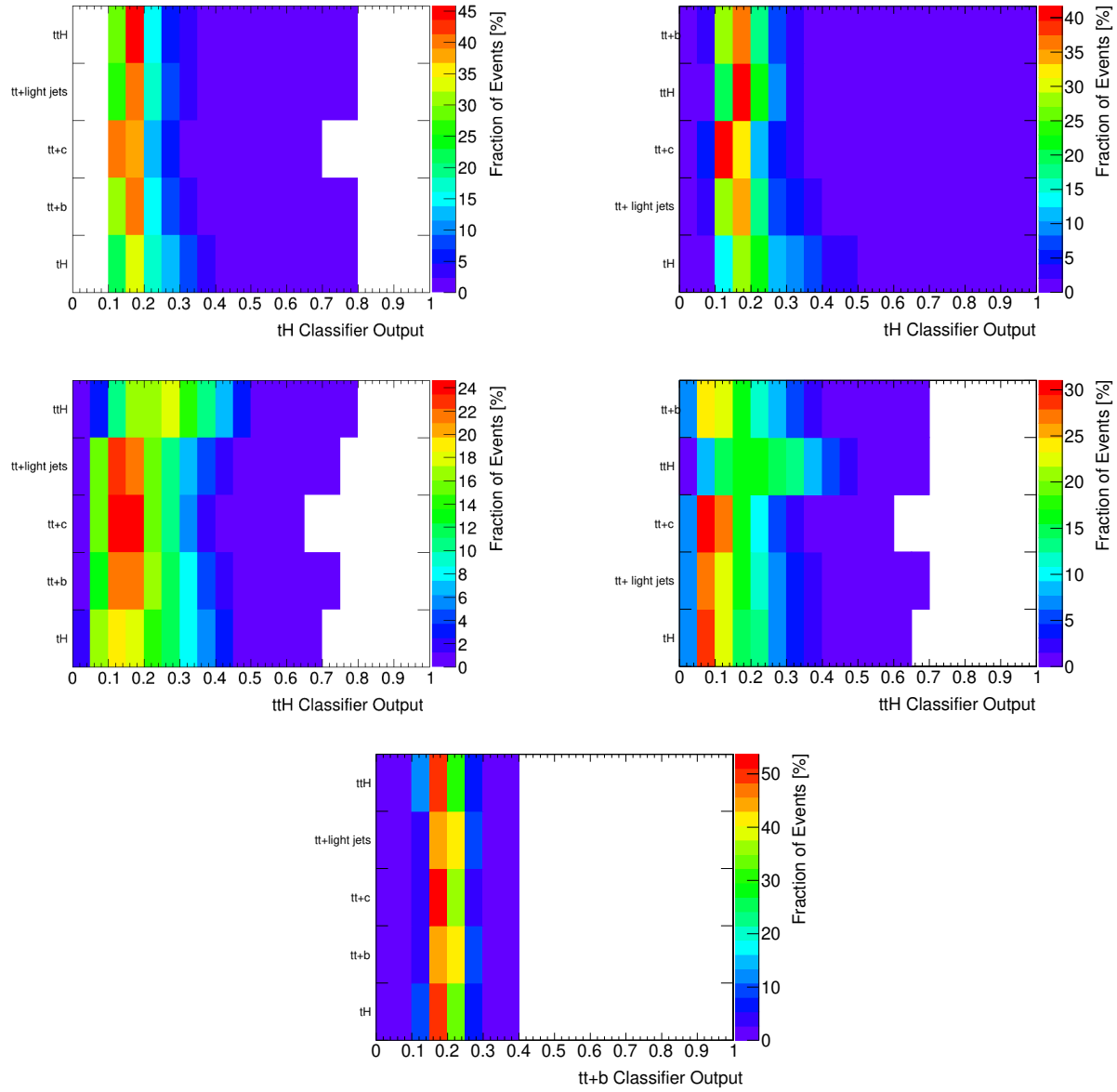

Figure A.6.: tH (top), ttH (middle) and confusion plots for the 5-class NN, trained and evaluated on the 5 jet (left) and 6 jet inclusive (right) regions. $tt + b$ (bottom) confusion plot for the 5 jet 5-class NN.

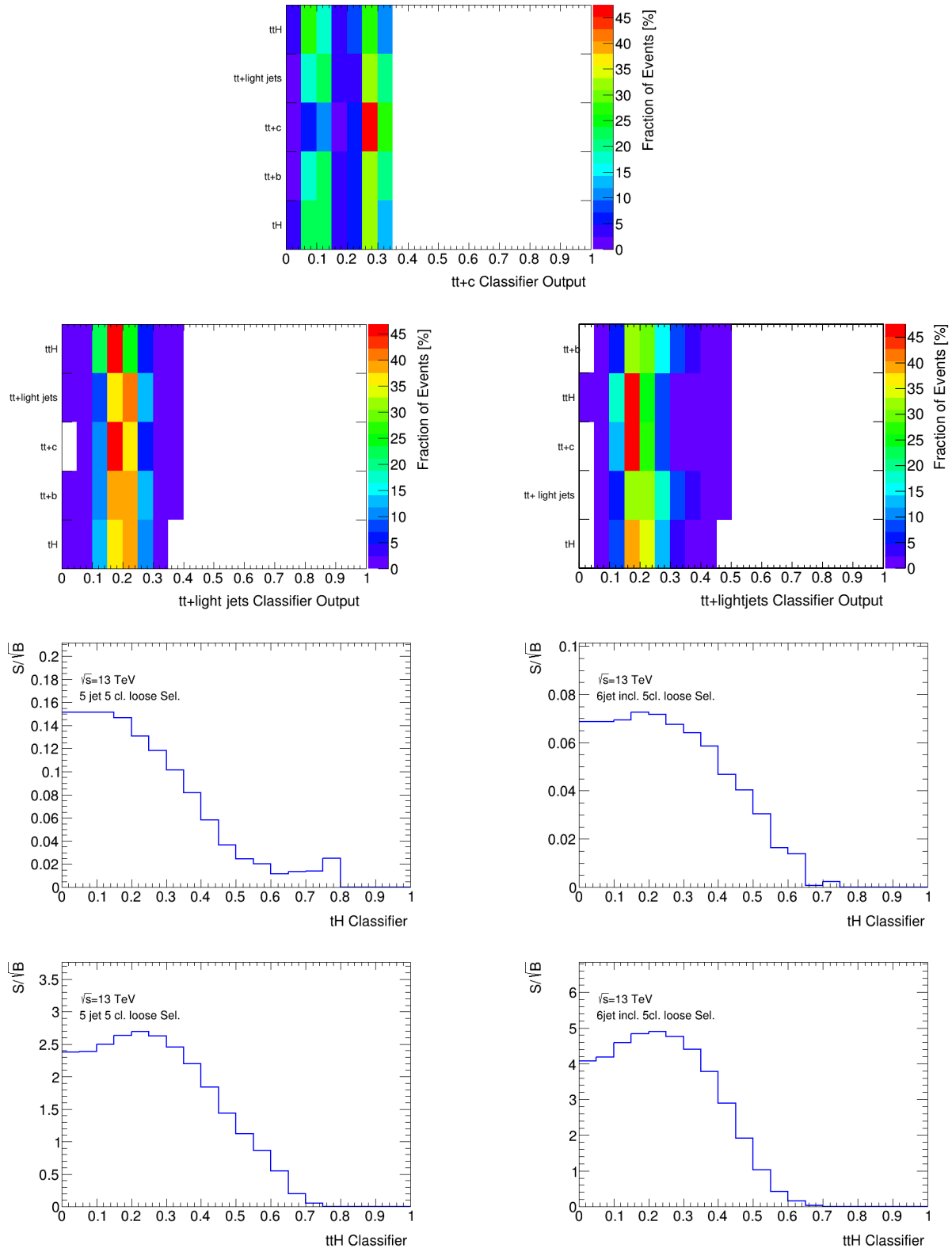

Figure A.7.: $tt + c$ (top) confusion plot for the 5 jet 5-class NN. $tt +$ light jets (second to top) confusion plots for the 5-class NN, trained and evaluated on the 5 jet (left) and 6 jet inclusive (right) regions. tH (second to bottom) and ttH (bottom) SoSB ratio plots for the for the 5-class NN, trained and evaluated on the 5 jet (left) and 6 jet inclusive (right) regions.

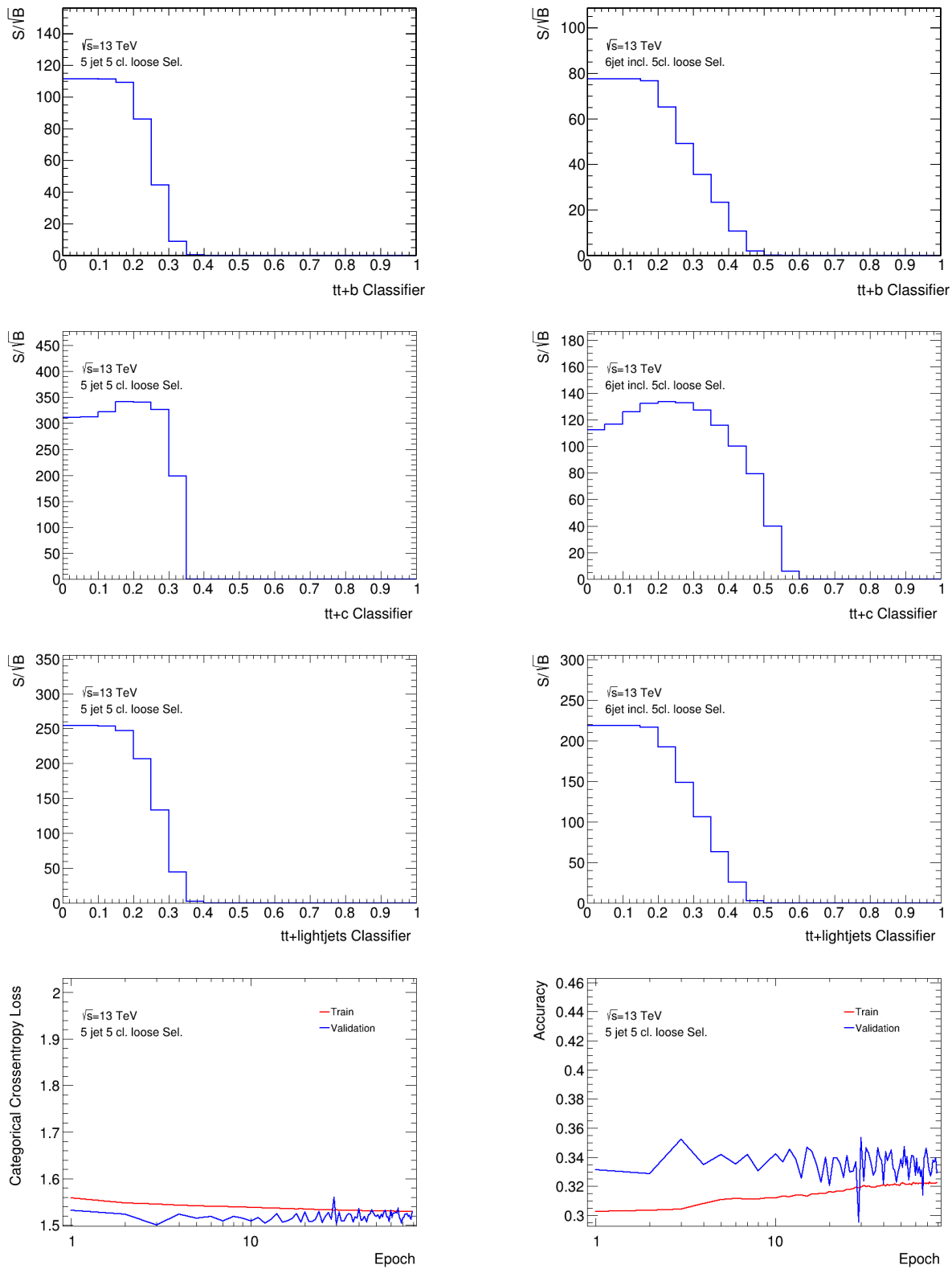

Figure A.8.: $tt + b$ (top), $tt + c$ (second to top) and $tt +$ light jets (second to bottom) SoSB ratio plots for the for the 5-class NN, trained and evaluated on the 5 jet (left) and 6 jet inclusive (right) regions. Representative categorical cross entropy loss (bottom left) and accuracy (bottom right) SoSB plots for the for the 5-class NNs.

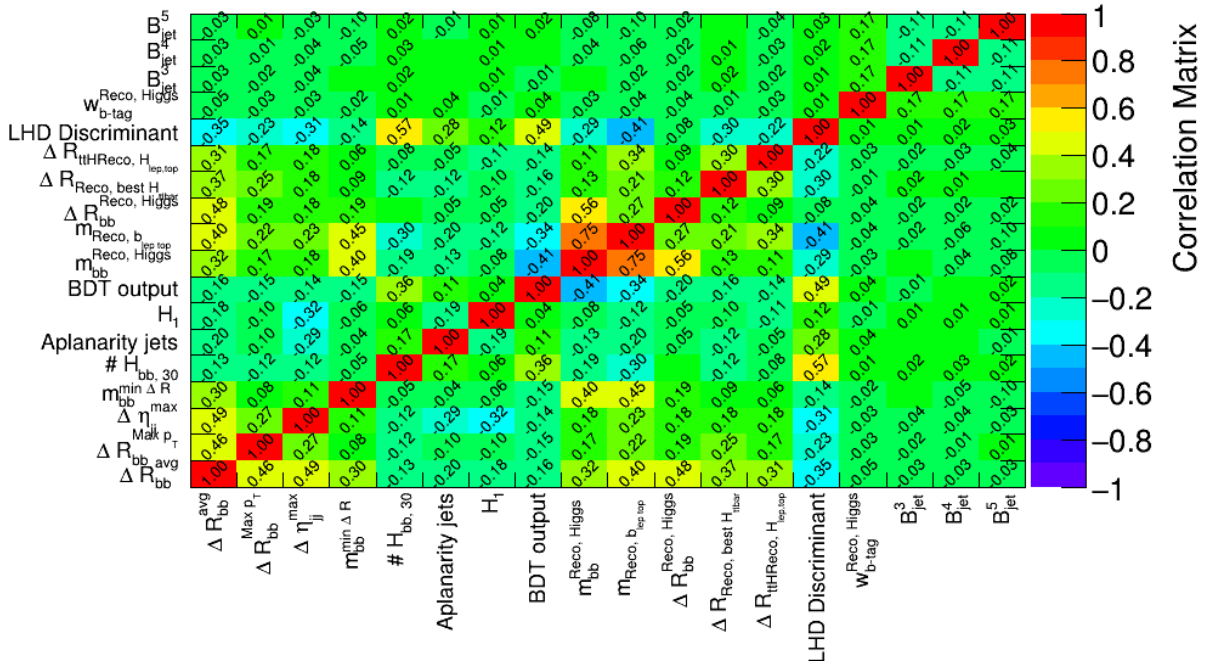

Figure A.9.: Correlation matrix for the input variables of the binary-classification NN,trained and validated on the larger dataset in the 6 jet inclusive region.

B. Further Tables

Variable	Variable name in code, † indicates Higgs information	Definition
	Kinematic variables	
$\Delta R_{bb}^{\text{avg}}$	`dRbb_avg_Sort4`	Average ΔR of all b-tagged jet pairs
$\Delta R_{bb}^{\text{max } p_{\text{T}}}$	`dRbb_MaxPt_Sort4`	ΔR between two b-tagged jets with the largest p_{T} vector sum
$\Delta\eta_{jj}^{\text{max}}$	`dEtajj_MaxdEta`	Maximum $\Delta\eta$ between any two jets
$m_{bb}^{\text{min } \Delta R}$	`Mbb_MindR_Sort4`	Mass of two b-tagged jets with the smallest ΔR
$\#H_{bb,30}$	`nHiggsbb30_Sort4`	Number of b-tagged jet pairs with an invariant mass within $\pm 30\,\text{GeV}$ of m_H
Aplanarity jets	`Aplanarity_jets`	$1.5 \cdot \lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor [74]
H_1	`H1_all`	Second Fox-Wolfram moment computed using all jets and the lepton [75]
	Variables from the reconstruction BDT	
BDT output	`TTHReco_withH_best_TTHReco_withH`†	Reconstruction BDT output
$m_{bb}^{\text{Reco, Higgs}}$	`TTHReco_best_Higgs_mass`	invariant mass of two b-tagged jets $\sim m_H$ (Higgs candidate mass)
$m_{\text{Reco},b_{\text{lep top}}}$	`TTHReco_best_Higgsbleptop_mass`	Mass of Higgs candidate and b-jet from leptonic top candidate
$\Delta R_{bb}^{\text{Reco, Higgs}}$	`TTHReco_best_bbHiggs_dR`	ΔR between b-jets from the Higgs candidate
$\Delta R_{\text{Reco, best H}_{\text{ttbar}}}$	`TTHReco_withH_best_Higgsttbar_dR`†	ΔR between Higgs candidate and $t\bar{t}$ candidate system
$\Delta R_{\text{Reco, H}_{\text{lep top}}}$	`TTHReco_best_Higgsleptop_dR`	ΔR between Higgs candidate and leptonic top candidate

Variables from likelihood calculations		
LHD Discriminant	`LHD_Discriminant`	Likelihood discriminant
Variables from b-tagging		
$w_{b\text{-tag}}^{\text{Reco, Higgs}}$	`TTHReco_best_bbHiggs_tagWeightBin_sum`	Sum of b-tagging discriminants of jets from the best Higgs candidate from the reconstruction BDT
B_{jet}^{3}	`jet_mv2_order_3_tagWeightBin`	3$^{\text{rd}}$ largest jet b-tagging discriminant
B_{jet}^{4}	`jet_mv2_order_4_tagWeightBin`	4$^{\text{th}}$ largest jet b-tagging discriminant
B_{jet}^{5}	`jet_mv2_order_5_tagWeightBin`	5$^{\text{th}}$ largest jet b-tagging discriminant

Table B.1.: List of BDT input variables and their definitions, see [32, 66]. Variables indicated with "`Sort4`" depend on b-tagged jets. Jets are sorted by their pseudo-continuous b-tag score and, in case of an equal pseudo-continuous b-tag score, by p_T. Variables indicated with † are from a reconstruction BDT using Higgs boson information, those not having a † stem from the reconstruction BDT without Higgs boson information. In the six jet hi region, defined in 5.1, B_{jet}^{3} and B_{jet}^{4} carry no discriminating power, also affecting the sum of b-tagging discriminants. Equation 4.3 is the definition for ΔR.

Variable	Variable name in code, † indicates Higgs information	Definition
	Kinematic variables	
$\Delta R_{bb}^{\text{avg}}$	`dRbb_avg_Sort4`	Average ΔR of all b-tagged jet pairs
$\Delta\eta_{bb}^{\text{Avg}}$	`dEtabb_Avg_Sort4`	Average $\Delta\eta$ of all b-tagged jet pairs
$\Delta\eta_{bb}^{\text{max}}$	`dEtabb_MaxdEta_Sort4`	Maximum $\Delta\eta$ of all b-tagged jet pairs
$m_{jjj}^{\text{max } p_{\text{T}}}$	`Mjjj_MaxPt`	Invariant mass of the three jets with highest p_{T}
$m_{bb}^{\text{min. mass}}$	`Mbb_MinM_Sort4`	Minimum mass of all b-tagged jet pairs
	Variables from the reconstruction BDT	
BDT outp. w/o H	`TTHReco_best_TTHReco`	Reconstruction BDT output without Higgs information
$m_{\text{Higgs}}^{\text{Reco}}$	`TTHReco_withH_best_Higgs_mass`†	Higgs mass from the reconstruction BDT that is closest to m_H
	Variables from likelihood calculations	
LHD Discriminant	`LHD_Discriminant`	Likelihood discriminant
	Variables from b-tagging	
B_{jet}^{3}	`jet_tagWeightBin_order_3_tagWeightBin`	3rd largest jet b-tagging discriminant
$w_{b\text{-tag}}^{\text{all}}$	`btag_sum_all`	Sum of all b-tagged jets
# b-tags	`nBTags_MV2c10_60`	Number of b-tagged jets at 60% WP

Table B.2.: List of multi-class NN input variables and their definitions. Variables indicated with "`Sort4`" depend on b-tagged jets. Jets are sorted by their pseudo-continuous b-tag score and, in case of an equal pseudo-continuous b-tag score, by p_T. Variables indicated with † are from a reconstruction BDT using Higgs boson information, those not having a † stem from the reconstruction BDT without Higgs boson information. Equation 4.3 is the definition for ΔR.

Hyperparameter	Settings	
	Simple NN	Complex NN
Number of hidden layers	3	5
Number of nodes in each layer	15	20
Activation functions in each layer	3 × ReLU	5 × ReLU
Output activation function	Sigmoid	Sigmoid
Metrics	accuracy, MSE, binary/ categorical cross entropy	accuracy, MSE, binary/ categorical cross entropy
Epochs	500	500
Loss	binary cross entropy	binary cross entropy
Learning rate	0.001	0.001
Batch size	1,024	1,024
Patience	30	30
$\Delta_{\text{min.}}$	0.0001	0.0001
Folds	5	5
Dropout index	1	2
Dropout probability	0.3	0.25
Trainable parameters	781	1,761

Table B.3.: Hyperparameters of the 5 NNs trained on the 5 jet high and low, 6 jet inclusive high and low and 6 jet inclusive high and low combined regions. The hyperparameters are defined in Section 3.7

Region	Event yield ± Poisson error				
	tH	ttH	tt + light jets	$tt + c$	$tt + b$
Small dataset					
6 jet, high$^{\geq 6j}_{\geq 4b,\text{hi}}$	3.4 ± 1.8	213.1 ± 14.6	$2,638.7 \pm 51.4$	109.8 ± 10.5	$2,669.4 \pm 51.7$
6 jet, low$^{\geq 6j}_{<4b,\text{lo}}$	3.3 ± 1.8	207.0 ± 14.4	$3,239.6 \pm 56.9$	765.4 ± 27.7	$3,131.3 \pm 56.0$
5 jet, high$^{5j}_{\geq 4b,\text{hi}}$	3.4 ± 1.8	64.4 ± 8.0	897.7 ± 30.0	55.0 ± 7.4	$1,013.6 \pm 31.8$
5 jet, low$^{5j}_{<4b,\text{lo}}$	3.3 ± 1.8	64.4 ± 8.0	$1,240.7 \pm 35.2$	396.8 ± 19.9	$1,403.9 \pm 37.5$
6 jet, high & low combined	6.7 ± 2.6	420.1 ± 20.5	$5,878.3 \pm 76.7$	875.2 ± 29.6	$5,800.7 \pm 76.2$
Large dataset					
6 jet$^{\geq 6j}_{\geq 4b,@85\%}$	22.3 ± 4.7	$1,311.7 \pm 36.2$	$50,897.0 \pm 225.6$	$30,689.8 \pm 175.2$	$22,354.3 \pm 149.5$
5 jet$^{\geq 6j}_{\geq 3b,@85\%}$	81.4 ± 9.0	$1,277.6 \pm 35.7$	$108,241.0 \pm 329.0$	$125,979.4 \pm 354.9$	$54,153.2 \pm 231.7$

Table B.4.: The event yield and Poisson error in each category of the small and large datasets.

Variable	Variable name in code	Definition
	Kinematic variables	
$\Delta R_{bb}^{\text{avg}}$	`dRbb_avg_Sort4`	Average ΔR of all b-tagged jet pairs
$\Delta\eta_{jj,\text{Avg}}$	`dEtajj_Avg`	Average $\Delta\eta$ between any two jets
$\Delta\eta_{bb}^{\text{max}}$	`dEtabb_MaxdEta_Sort4`	Maximum $\Delta\eta$ of all b-tagged jet pairs
# HF Jets	`nHFJets`	Number of heavy flavour jets
$\#H_{jj,30}$	`nHiggsjj30`	Number of arbitrary jet pairs with an invariant mass within $\pm 30\,\text{GeV}$ of m_H
$\Delta\eta_{bb}^{\text{Avg.}}$	`dEtabb_Avg_Sort4`	Average $\Delta\eta$ of all b-tagged jet pairs
$M_{jjj}^{\text{max } p_{\text{T}}}$	`Mjjj_MaxPt`	Invariant mass of 3 jet system with the highest p_{T}
# $\text{Jets}_{p_{\text{T}},40}$	`nJets_Pt40`	Number of jets with $p_{\text{T}} \geq 40\,\text{GeV}$
	Variables from the reconstruction BDT	
BDT output w/o H	`TTHReco_best_TTHReco`	Reconstruction BDT output without Higgs information
$w_{b-\text{tag}}^{\text{Reco, Higgs}}$	`TTHReco_best_bbHiggs_tagWeightBin_sum`	Sum of the two b-tagged jets coming from the reconstructed Higgs boson
	Variables from likelihood calculations	
LHD Discriminant	`LHD_Discriminant`	Likelihood discriminant
	Variables from b-tagging	
$w_{b\text{-tag}}^{\text{all}}$	`btag_sum_all`	Sum of all b-tagged jets

Table B.5.: Input variables to the multi-class NNs trained and validated on 6 jet region of the large dataset.

Variable	Variable name in code	Definition
Kinematic variables		
$\Delta R_{bb}^{\text{avg}}$	`dRbb_avg_Sort4`	Average ΔR of all b-tagged jet pairs
$\Delta R_{\text{lep. H}}$	`dR_lepH`	
$\Delta\eta_{jj}^{\text{max}}$	`dEtajj_MaxdEta`	Maximum of $\Delta\eta$ between two jets
$\Delta\eta_{jj,\text{Avg.}}$	`dEtajj_Avg`	Average $\Delta\eta$ of two jets
$\Delta\eta_{bb}^{\text{max}}$	`dEtabb_MaxdEta_Sort4`	Maximum $\Delta\eta$ of all b-tagged jet pairs
$\Delta\eta_{bb}^{\text{Avg.}}$	`dEtabb_Avg_Sort4`	Average $\Delta\eta$ of all b-tagged jet pairs
$m_{bj}^{\text{max. } p_{\text{T}}}$	`Mbj_MaxPt_Sort4`	Invariant mass of a b-tagged and arbitrary jet system with highest p_{T}
$m_{bb}^{\text{min. mass}}$	`Mbb_MinM_Sort4`	Minimum invariant mass of all b-tagged jet pairs
$m_{bb}^{\text{min.}\Delta R}$	`Mbb_MindR_Sort4`	Invariant mass of the b-tagged jet pair with the smallest ΔR
$m_{jjj}^{\text{max } p_{\text{T}}}$	`Mjjj_MaxPt`	Invariant mass of 3 jet system with the highest p_{T}
Variables from likelihood calculations		
LHD Discriminant	`LHD_Discriminant`	Likelihood discriminant
Variables from b-tagging		
$w_{b\text{-tag}}^{\text{all}}$	`btag_sum_all`	Sum of all b-tagged jets

Table B.6.: Input variables to the multi-class NNs trained and validated on 5 jet region of the large dataset. Variables from the $t\bar{t}H$ reconstruction BDT are not included, as this reconstruction requires at least 6 jets.

Region	AUC avg. training	AUC avg. testing	AUC avg. overall
5 jet low	0.7534	0.7508	0.7521
5 jet high	0.7756	0.7726	0.7741
6 jet incl. low	0.7502	0.7488	0.7495
6 jet incl. high	0.7682	0.7674	0.7678
6 jet incl. high and low comb. simple	0.7200	0.7614	0.7617
6 jet incl. high and low comb. complex	0.7652	0.7634	0.7643

Table B.7.: Average AUCs of the ROC curves during training, testing and overall average of the binary-classification NNs, trained on the small dataset.

	$t\bar{t}H$	tH	$t\bar{t}+b$	$t\bar{t}+c$	$t\bar{t}+$ light jets
AUC avg. training	0.7462	0.7134	0.5330	0.7394	0.5412
AUC avg. testing	0.7460	0.7048	0.5328	0.7356	0.5408
AUC avg. overall	0.7461	0.7091	0.5329	0.7375	0.5410

Table B.8.: Average AUCs of the ROC curves during training, testing and overall average of the multi-class NN, trained on the small dataset, for each output class.

	5 jet			6 jet inclusive		
	AUC avg. training	AUC avg. testing	AUC avg. overall	AUC avg. training	AUC avg. testing	AUC avg. overall
ttH	0.7095	0.7090	0.7093	0.7400	0.7400	0.7400
tH	0.5918	0.5888	0.5903	0.6938	0.6898	0.6918
$tt+$ jets	0.6875	0.6875	0.6875	0.7260	0.7259	0.7260

Table B.9.: Average AUCs of the ROC curves during training, testing and overall average of the 3-class NN, trained on the large dataset, for each output class.

	5 jet			6 jet inclusive		
	AUC avg. training	AUC avg. testing	AUC avg. overall	AUC avg. training	AUC avg. testing	AUC avg. overall
ttH	0.7193	0.7193	0.7193	0.7410	0.7163	0.7287
tH	0.6135	0.6100	0.6118	0.7020	0.6963	0.6992
$tt + b$	0.5313	0.5313	0.5313	0.5318	0.5318	0.5318
$tt +$ light and c	0.5695	0.5693	0.5694	0.5833	0.5833	0.5833

Table B.10.: Average AUCs of the ROC curves during training, testing and overall average of the 4-class NN, trained on the large dataset, for each output class.

	5 jet			6 jet inclusive		
	AUC avg. training	AUC avg. testing	AUC avg. overall	AUC avg. training	AUC avg. testing	AUC avg. overall
ttH	0.7235	0.7230	0.7233	0.7483	0.7480	0.7482
tH	0.6295	0.6265	0.6280	0.6953	0.6928	0.6941
$tt + b$	0.5478	0.5480	0.5479	0.5778	0.5778	0.5778
$tt + c$	0.6333	0.6333	0.6333	0.7393	0.7388	0.7391
$tt +$ light jets	0.5740	0.5738	0.5739	0.5948	0.5950	0.5949

Table B.11.: Average AUCs of the ROC curves during training, testing and overall average of the 5-class NN, trained on the large dataset, for each output class.

Acknowledgements

I would like to thank Prof. Dr. Arnulf Quadt for providing me with the great opportunity of writing my bachelor's thesis in this wonderful working group and a big "thank you" to the whole working group for the very warm welcome. I also want to thank Prof. Dr. Stan Lai for his time and support as a second referee. I would like to thank Dr. Jelena Jovicevic and Dr. Elizaveta Shabalina for their support and technical supervision of this thesis. Thank you to Steffen for providing me with such a sophisticated MVA framework, I am glad to have been able to contribute to the code.

Lastly, thank you so much Chris! Thank you for your continuous support during the 12 weeks, no matter if it was 1 p.m. or 1 a.m., whether it was debugging code, answering physics questions, helping when the whole thing broke down (again), or reading through my thesis, you were always able to help me!

Erklärung nach §13(9) der Prüfungsordnung für den Bachelor-Studiengang Physik und den Master-Studiengang Physik an der Universität Göttingen:
Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe.
Darüberhinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, im Rahmen einer nichtbestandenen Prüfung an dieser oder einer anderen Hochschule eingereicht wurde.

Göttingen, den 27. September 2021

(Konrad Helms)