Master's Thesis

Czech
Technical
University
in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

Search for a charged Higgs boson decaying into a neutral Higgs boson and a W-boson Using Machine Learning with ATLAS Data

Bc. Azad Afandizada

Supervisor: doc. Dr. André Sopczak
May 2025

Acknowledgements

I would like to thank my supervisor doc. Dr. André Sopczak who gave me the opportunity to work on this project and who guided me during my work. Special thanks to my family who supported me during my academic studies.

Declaration

I declare that I have prepared the final thesis independently and have listed all information sources used in accordance with the Methodological Guidelines on the observance of ethical principles in the preparation of university final theses and the Framework Rules for the use of artificial intelligence at CTU for study and pedagogical purposes in Bc and NM studies. I declare that I have used artificial intelligence tools during the preparation and writing of the final thesis. I have verified the generated content. I confirm that I am aware that I am fully responsible for the content of the final thesis.

In Prague, 20. May 2025

Abstract

A charged Higgs boson remains undiscovered to this day, despite the significant impact it would have for improving our understanding of particle physics. The detection of such a particle could revolutionize areas such as the study of electroweak symmetry breaking, extensions to the Standard Model and potentially open doors to entirely new physics. This thesis focuses on the detection of the proposed charged Higgs boson using machine learning techniques applied to simulated proton-proton collisions. The main objective is to develop a machine learning model to identify and distinguish the desired events of the charged Higgs boson from background events.

Keywords: CERN, ATLAS, ROOT, tbH^+, ttH, ttZ, ttW

Supervisor: doc. Dr. André Sopczak

Abstrakt

Nabitý Higgsův boson zůstává dosud neobjevený, a to navzdory jeho potenciálně zásadnímu vlivu na naše porozumění částicové fyzice. Jeho detekce by mohla představovat průlom například ve studiu narušení elektroslabé symetrie, rozšíření Standardního Modelu nebo by mohla naznačit cestu k nové fyzice. Tato práce se zaměřuje na hledání hypotetického nabitého Higgsova bosonu pomocí technik strojového učení, aplikovaných na simulovaná data srážek proton-proton. Hlavním cílem je vytvořit model, který dokáže spolehlivě odlišit události odpovídající signálu nabitého Higgsova bosonu od událostí pozadí.

Klíčová slova: CERN, ATLAS, ROOT, tbH^+, ttH, ttZ, ttW

Contents

Figures

Tables

Chapter 1

Introduction

The Higgs boson[10] was discovered in 2012 by the ATLAS[11] and CMS[12] collaborations at CERN and has a mass of approximately 125 GeV. Often referred to as the "God particle," the Higgs boson plays a central role in particle physics as it provides mass to other fundamental particles through the Higgs field. Without it the existence of matter and the universe as we know it would not be possible.

Particle physics[13] — is the branch of science that seeks to understand the fundamental constituents of matter and the forces through which they interact. It aims to answer the most profound questions about the universe: What are the building blocks of nature? How do they acquire mass? Why is there more matter than antimatter? Despite the success of the Standard Model in describing known particles and interactions many mysteries remain unsolved such as the nature of dark matter, the imbalance of matter and antimatter and the properties of the Higgs sector itself.

Since the discovery of the Higgs boson particle physicists have focused on exploring its properties and searching for possible extensions of the Standard Model. One such extension involves the existence of *charged Higgs bosons*, predicted by theories such as the Two-Higgs-Doublet Model (2HDM) and supersymmetry (SUSY)[14]. The observation of a charged Higgs boson would provide evidence for physics beyond the Standard Model and would extend the mechanism of electroweak symmetry breaking and the generation of mass.

The goal of this thesis is to develop a machine learning model and apply it to simulated events representing proton-proton collisions at the Large Hadron Collider (LHC). The main focus is to identify and distinguish events containing the production of a charged Higgs boson from the background processes. By using data analysis techniques and machine learning algorithm this study aims to improve the efficiency and accuracy of identifying charged Higgs events.

This analysis uses a new release of the ATLAS software with high statistics and improved calibrations compared to previous studies. First, we will discuss the theory behind this study including the Standard Model. The role of the Higgs boson and possible extensions predicting charged Higgs particles. This will be followed by a detailed description of the datasets and features used in the analysis. Then, we will present the development and performance of the machine learning model using diagrams and graphs to illustrate the difference between signal and background events. Next expected limits are derived to quantify the significants. Finally, conclusions are given.

Chapter 2

Physics

2.1 The Large Hadron Collider

First of all it is important to mention the Large Hadron Collider (LHC)[15] which is the world's largest particle collider. It was constructed and is operated by the European Organization for Nuclear Research (CERN). Located on the France-Switzerland border near Geneva its construction spanned a decade, from 1998 to 2008.

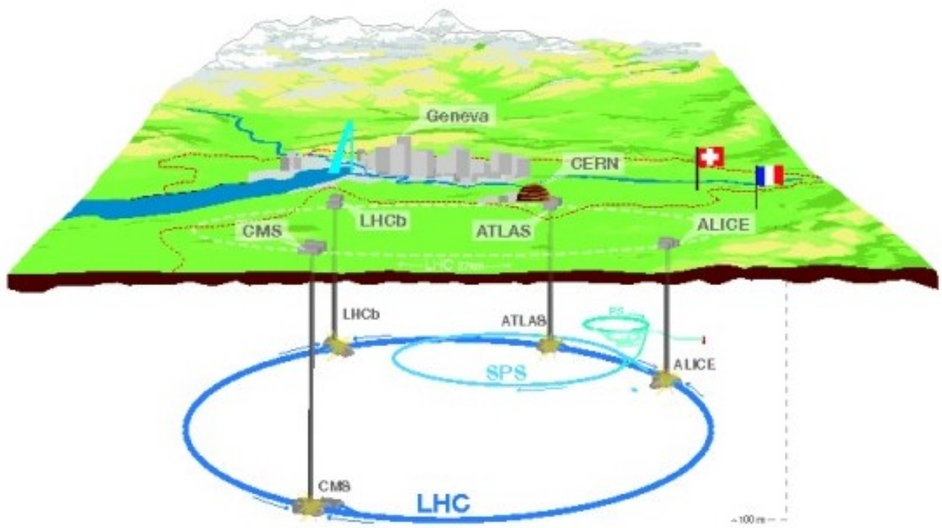

Figure 2.1: Large Hadron Collider[1].

The Large Hadron Collider plays a crucial role in modern physics by enabling the collision of subatomic particles to test theories which are beyond the Standard Model of particle physics. It continues to be used in the search for new particles including to confirm or exclude the existence of the charged Higgs boson.

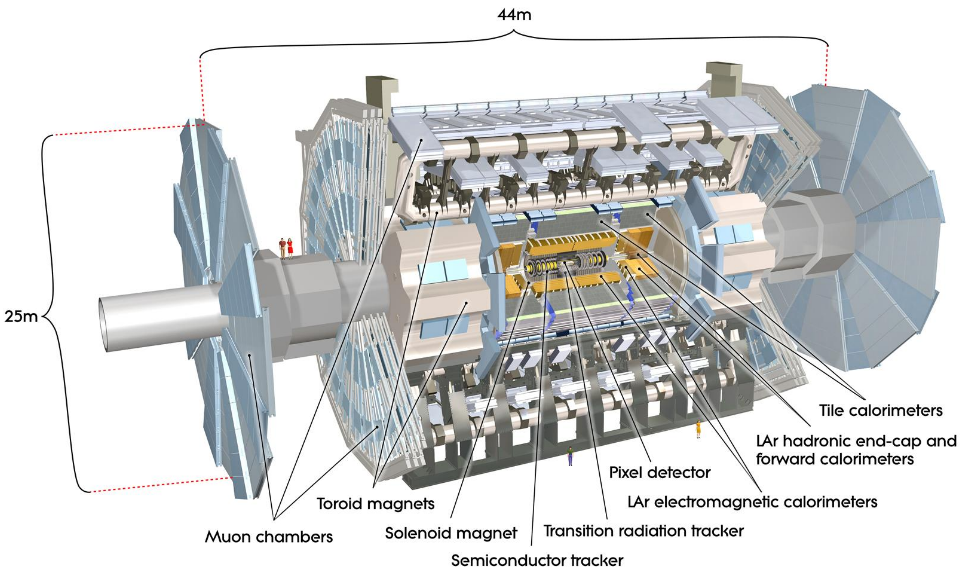

Figure 2.2: Cut-away view of the ATLAS detector[2].

The ATLAS detector[16] is a particle detector at the Large Hadron Collider. It is cylindrical and approximately 44 meters long and 25 meters in diameter. The detector consists of several subsystems arranged in layers around the interaction point. The inner layers contain the tracking system including the pixel detector, semiconductor tracker and transition radiation tracker which provide precise measurements of particle trajectories. Surrounding the tracker is the solenoid magnet which bends the paths of charged particles. Next are the electromagnetic and hadronic calorimeters (the liquid argon electromagnetic calorimeter, tile calorimeters and the liquid argon hadronic end-cap and forward calorimeters) responsible for measuring the energy of electrons, photons and hadrons. The outer layers contain the muon chambers and large toroid magnets used to detect and measure muons. This multi-layered structure allows ATLAS to identify and measure a wide variety of particles with precision.

Luminosity[17] measures the performance of a collider. How many particles are being collided per second and how densely they are packed. It tells how much data an experiment has collected and measured in inverse femtobarns:

$$1 \text{ fb}^{-1} = 10^{15} \text{ cm}^{-2}$$

The number of events expected in the detector is given by:

$$N = \mathcal{L} \times \sigma, \tag{2.1}$$

where:

- $\mathcal{L}$ is the luminosity (in our case it is 140 fb^{-1}),
- σ is the cross section of the physical process and measured in picobarns (pb).

2.2 Standard Model

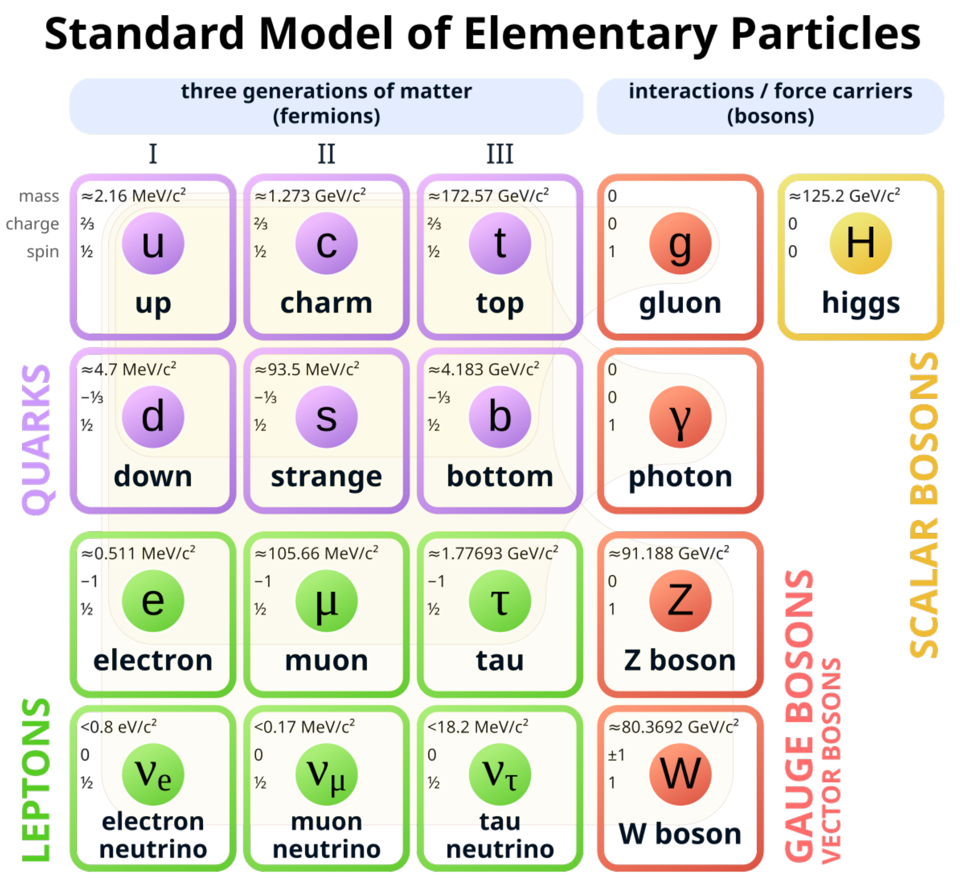

Figure 2.3: The Standard Model of particle physics[3].

The Standard Model[18] classifies all known elementary particles and describes three of the four fundamental forces: electromagnetic, weak and strong interactions (gravity is not included in the Standard Model). The particles are divided into two main categories: fermions and bosons.

Fermions[19] are matter particles and come in two types: *quarks* and *leptons*. They are organized into three generations of increasing mass. Each generation contains two quarks and two leptons. Quarks carry color charge and interact via the strong force while leptons do not.

Bosons[20] are force carriers: the photon (γ) mediates the electromagnetic force, the $W^{\pm}$ and Z bosons mediate the weak force and the gluon (g) mediates the strong force. The recently discovered Higgs boson (H) is responsible for giving mass to the fundamental particles via the Higgs mechanism.

Each particle's electric charge, weak isospin, and mass are also indicated in the Figure 2.3. The Standard Model has been successful in describing experimental results and predicting phenomena confirmed by experiments although it does not include gravity or dark matter.

2.3 The Higgs Mechanism

The Higgs mechanism[21] is a fundamental process in the Standard Model of particle physics that explains how elementary particles get mass. According to gauge theory the fundamental forces should be mediated by massless gauge bosons. However, in nature the $W^{\pm}$ and Z bosons that mediate the weak interaction are massive. The Higgs mechanism solves this problem.

In this mechanism a scalar field known as the *Higgs field* fills all of space. Elementary particles acquire mass through their interactions with the Higgs field. As the universe cooled after the Big Bang the Higgs field underwent *spontaneous symmetry breaking* settling into a non-zero vacuum expectation value (VEV). This process broke the electroweak symmetry and allowed the $W^{\pm}$ and Z bosons to get mass while the photon remained massless.

2.4 Signals and Background Events

The charged Higgs boson (H^+) is a hypothetical particle predicted by extensions of the Standard Model (SM) such as the Minimal Supersymmetric Standard Model (MSSM) and the Two-Higgs-Doublet Model (2HDM). Unlike the neutral Higgs boson discovered in 2012, the charged Higgs boson has an electric charge and the potential existence of new physics. The production of charged Higgs bosons in a proton-proton collider can be divided into two mass regimes, based on the Higgs boson's mass relative to the top quark mass (m_t):

The production of charged Higgs bosons in a proton-proton collider can be divided into two mass regimes, based on the Higgs boson's mass relative to the top quark mass (m_t):

- **Low Mass Region** ($m_{H^+} < m_t$): In this scenario, the charged Higgs boson is produced through the decay of a top quark ($t \to H^+b$) within top-antitop ($t\bar{t}$) pairs.

- **High Mass Region** ($m_{H^+} > m_t$): Here the main production processes involve the fusion of bottom and top quarks which leads to reactions such as $pp \to H^+tb$.

The decay modes of the charged Higgs boson depend on its mass and the specific theoretical model. In this analysis we focus on the decay chain $H^+ \to WH$, followed by $H \to \tau\tau$ resulting in the final state $tbH^+ \to tbWH \to tbW\tau\tau$. The analysis is conducted in the $2l$SS1τ channel which studies events characterized by two same-sign leptons and one hadronically decaying tau lepton (τ).

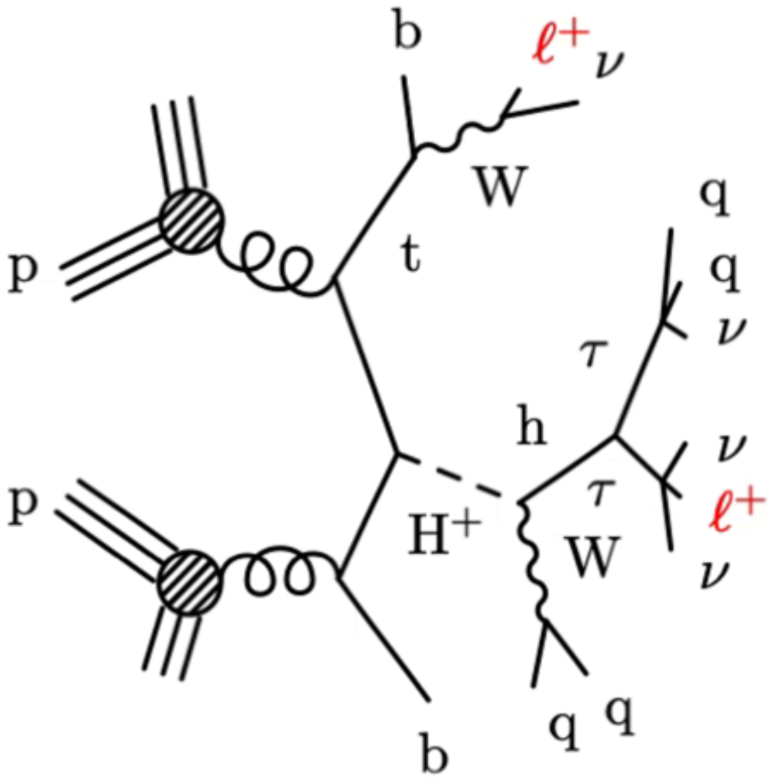

Figure 2.4: Feynman diagram showing occurrence of tbH^{+}[4].

Particles in Feynman diagram:

A Feynman diagram is a graphical description of the behavior and interaction of subatomic particles. To understand the diagram we have to know the particles themselves.

p[22]- Protons are the colliding particles in this diagram. They are composed of quarks (two up quarks and one down quark) and gluons which interact via the strong force.

b[23] – The bottom quark is a fundamental particle from the third generation of quarks and has a charge of -1/3.

t[24] – The top quark is the heaviest of all known quarks, belonging to the third generation. It has a charge of +2/3.

H^{+} - Charged Higgs boson which is a hypothetical particle. It carries a positive electric charge.

W[25] - The W Boson is a mediator of the weak force, carrying an electric charge. It is responsible for processes involving weak interactions.

l^+ [26] - A positively charged light lepton (electron or muon).

ν [27] – Neutrino a neutral and nearly massless particle a member of the lepton family. Neutrinos interact only via the weak force, making them extremely difficult to detect.

h - Neutral Higgs boson which gives mass to other elementary particles. Decays into two τ

τ - The tau lepton is a heavy third-generation lepton with a negative electric charge. It has a relatively short lifetime and decays into other particles.

q - Light quarks fundamental particles that form protons and neutrons.

As we do not know the mass of charged Higgs boson, we will use several predicted mass ranges. We will consider the following masses for charged Higgs boson: 250 GeV, 800 GeV and 3000 GeV.

Backgrounds:

During proton-proton collisions, various other events can occur alongside the production of the charged Higgs boson which we categorize as background events. These background events must be analyzed and separated from the signal as they can be mistaken for the production and decay of charged Higgs boson. This happens because the initial interaction such as two gluons interacting and the final observed particles can appear identical for both signal and background processes. For us to determine whether charged Higgs boson production has occurred it is important to study and understand background events. We will consider only 3 most important background reactions.

ttH: Is a process where a Higgs boson is produced in association with a top quark-antiquark pair during proton-proton collisions (Figure 2.5).

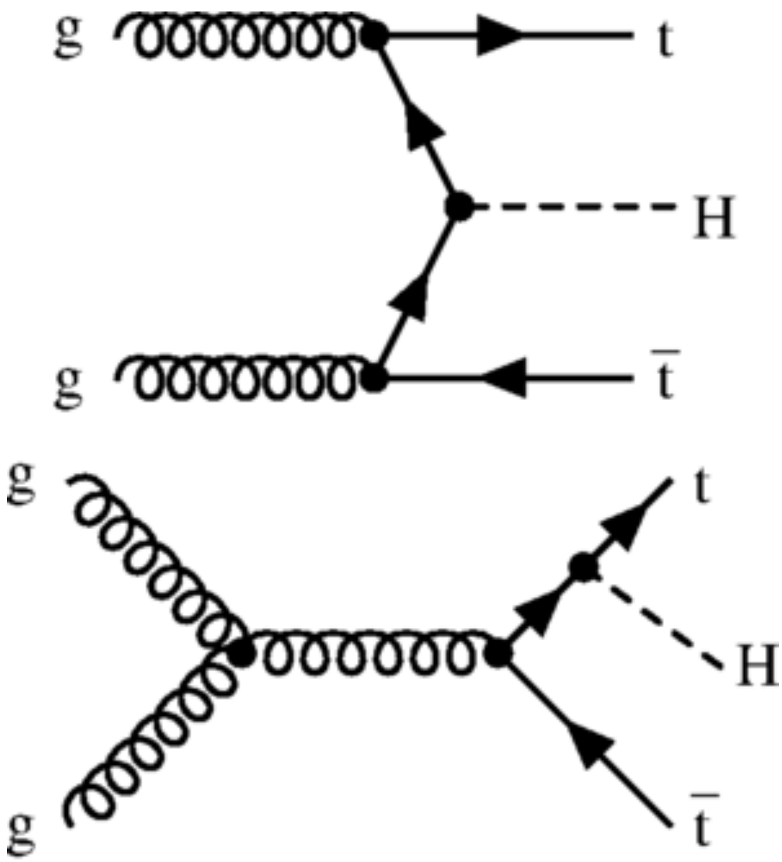

Figure 2.5: Feynman diagram for ttH production[5].

ttW: Is a process where a top quark-antiquark pair is produced in association with a W boson during proton-proton collisions (Figure 2.6).

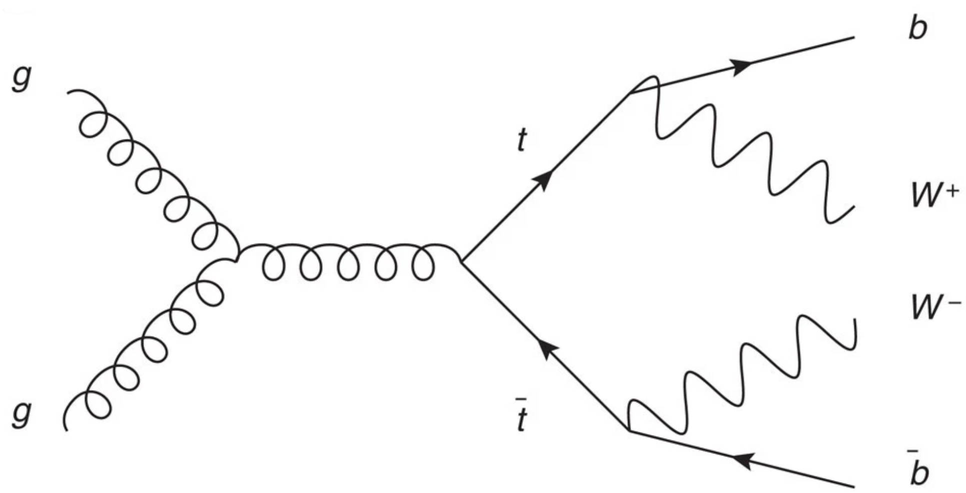

Figure 2.6: Feynman diagram for ttW production [6].

ttZ: Is a process where a top quark-antiquark is produced in association with a Z boson during proton-proton collisions (Figure 2.7).

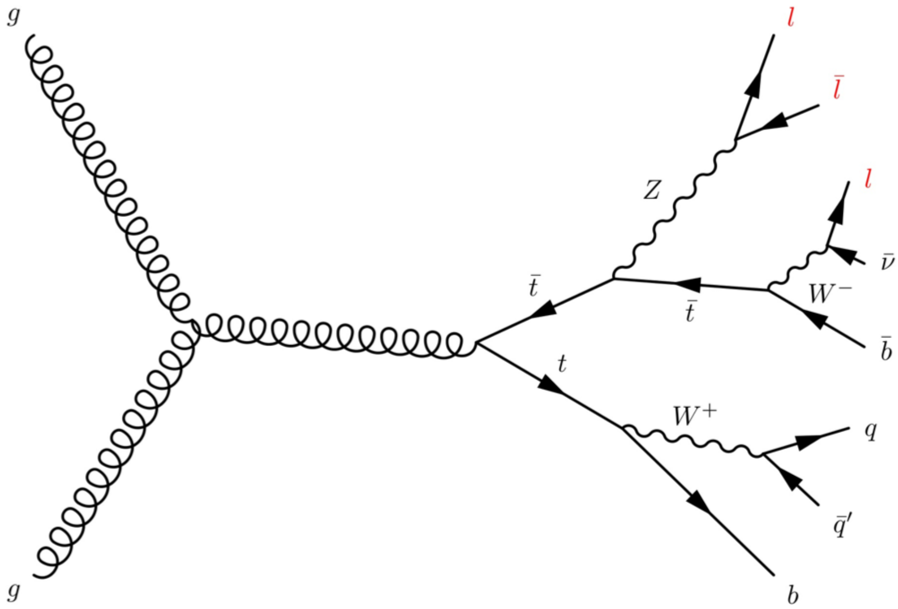

Figure 2.7: Feynman diagram for ttZ production [7].

2.5 Monte Carlo

Monte Carlo (MC)[28] simulations play a fundamental role in particle physics analyses. They are used to model both the signal and background processes by simulating particle collisions, their resulting interactions and the detector response.

In this study MC simulations were employed to generate proton-proton collision events at a center-of-mass energy of $\sqrt{s} = 13$ TeV, corresponding to the conditions at the Large Hadron Collider. The simulations include theoretical calculations of hard scattering processes, parton showering, hadronization, and a detailed emulation of the ATLAS detector response.

Signal events corresponding to charged Higgs boson production (tbH$^+$) and background processes such as ttW, ttH and ttZ were produced using state-of-the-art MC generators. These simulations are crucial for training and validating the machine learning model developed in this thesis as they provide labeled datasets that represent the expected physics signatures and detector effects.

Chapter 3

Machine Learning

Machine learning and artificial intelligence are becoming widespread and are playing a central role in many areas of today's life. ML techniques are useful for simplifying complex classification tasks. In our case machine learning can be applied to distinguish the tbH^+ signal from background events on statistical basis.

3.1 Software Used

Python:
Python [29] is a widely used programming language especially in fields such as data analysis and machine learning. In this project, Python was used to design and develop machine learning models and to build visualizations. Additionally tools like TRExFitter which were used to create graphs and analyze data are written in Python.

TensorFlow:
In this analysis the machine learning model was implemented using **TensorFlow** [30] an open-source deep learning framework developed by the Google Brain team. TensorFlow provides a highly flexible and efficient ecosystem for building, training and deploying machine learning models.

API (Application Programming Interface):
It is a set of definitions and protocols that allow users to interact with software libraries or frameworks in a structured way. In this work the Keras API [31] was used to define, train, and evaluate neural network models efficiently.

TRExFitter:
TRExFitter [32] is a powerful statistical analysis tool widely used in high energy physics. Built on top of ROOT and RooStats, it provides a range of functionalities, including:

- Fitting distributions
- Calculating significance and exclusion limits
- Visualizing outputs, such as producing plots by pseudo-merging results

Its flexibility and precision make it a main resource for analyzing and interpreting complex datasets in particle physics

FastFrame framework:
The FastFrame [33] is a newly developed software framework that was used for the creation and preprocessing of the Ntuples. It provided a standardized and flexible way to select the desired $2\ell SS1\tau$ channel, including both the signal events (tbH^+) and background processes (ttW, ttH, and ttZ). This allowed for the efficient extraction of events within the targeted phase space. The processed Ntuples generated by FastFrame were used as the input dataset for the machine learning analysis.

ROOT:
ROOT [34] is a comprehensive data analysis framework developed by CERN used for processing large datasets in high-energy physics. It offers tools for:

- Data processing
- Statistical analysis
- Visualization of results

The datasets used in ROOT are typically structured as Ntuples which consist of branches corresponding to various variables recorded by the ATLAS detector. These branches contain information on particle properties such as types, momenta, and energies providing a detailed view of particle interactions observed during collisions.

3.2 Basic Terminology

Activation Function - Decides whether a node should be activated or not. Meaning that it will decide whether the node's input to the network is important or not in the process of prediction [35].

Loss Function - A loss function is a mathematical formula that calculates the difference between the predicted output of a machine learning model and the actual target value. It serves as the objective that the learning algorithm minimizes during training [36].

Epochs - It is one complete pass through the training dataset. During each epoch the model sees and learns from every training sample once [37].

Batch Size - It is the number of samples processed (input–label pairs) in one iteration of training a model [38].

Weights - Are trainable parameters that determine how much influence each input feature has on the model's output [39]. Each node computes

$$z = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b \tag{3.1}$$

where,

- x_i — input feature,
- w_i — corresponding weight,
- b — bias term (additional trainable parameter),
- z — weighted sum, passed to the activation function.

Learning Rate - Is a hyperparameter that controls the step size during the optimization process. It determines how much the model's weights are updated in response to the estimated error at each iteration [40].

3.3 Neural Network

Neural network [41] is a type of machine learning model which is said to be similar to the structure and function of the human brain. It consists of layers of interconnected nodes that can learn to recognize patterns in data. Each node performs a weighted sum of its inputs then applies a non-linear activation function and passes the result to the next layer. When a neural network contains multiple hidden layers between the input and output layers it is called a **Multilayer Perceptron**.

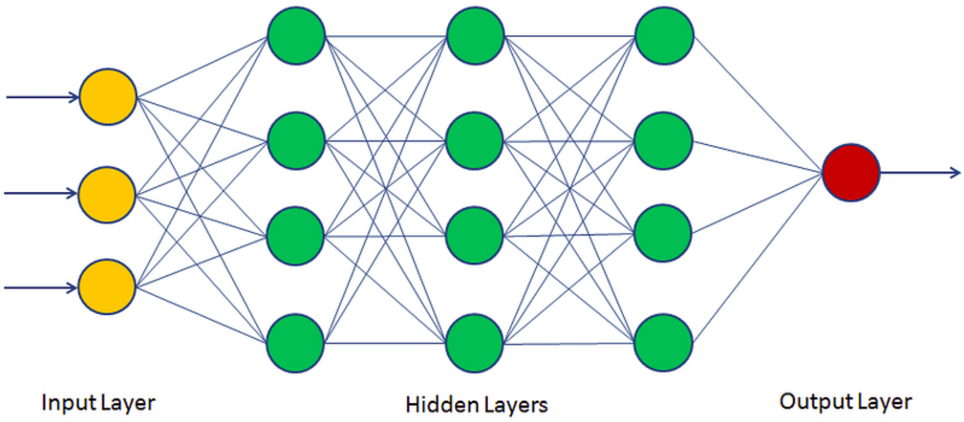

Figure 3.1: Schematic diagram of a Multilayer Perceptron [8].

In this work a sequential neural network was constructed using the Keras API within TensorFlow. The model includes several fully connected layers, batch normalization, activation functions, was trained using the `Adam` optimizer and the binary cross-entropy loss function. The training process was monitored with early stopping to prevent overfitting and model evaluation was performed using ROC curves and AUC scores.

In this analysis a multilayer perceptron was developed to distinguish between signal and background events. The task was formulated as a binary classification problem where each event was labeled either as signal (1) or background (0). The neural network was trained using the loss function and the final output layer which consisted of a single neuron with a sigmoid activation function to produce a probability score between 0 and 1. The probability represented that a given event is signal.

Binary Cross-Entropy Loss Function:
The binary cross-entropy loss function [42] was used in our classification task. It measures the difference between the predicted probabilities and the actual class labels and is defined as

$$\mathcal{L}(y, \hat{y}) = -\left[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right], \tag{3.2}$$

where $y \in \{0, 1\}$ is the true label and $\hat{y} \in [0, 1]$ is the predicted probability.

ReLU Activation Function:

The Rectified Linear Unit (ReLU) [43] was used as the activation function in the hidden layers of the network. It introduces non-linearity while remaining computationally efficient. It is defined as

$$\mathrm{ReLU}(x) = \max(0, x), \tag{3.3}$$

where

- x is the input to the activation function. Usually the output of a linear transformation in a neural network layer (a weighted sum of inputs plus bias).

This means it outputs zero for all negative inputs and passes positive inputs unchanged. It helps prevent the vanishing gradient problem and accelerates convergence during training.

Sigmoid Activation Function:
The sigmoid function [44] is an S-shaped function which is used to convert output scores into probabilities. It is given as

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{3.4}$$

Adam Optimizer

Adam (Adaptive Moment Estimation) [45] is one of the most widely used optimization algorithms in training models. It combines the advantages of two other popular methods: AdaGrad and RMSProp.

The Adam optimizer maintains exponentially decaying averages of past gradients (first moment estimates) and past squared gradients (second moment estimates). These averages are used to adaptively adjust the learning rate for each parameter. The update rule for the parameters is given by

$$\begin{aligned}
m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t \\
v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \\
\hat{m}_t &= \frac{m_t}{1 - \beta_1^t} \\
\hat{v}_t &= \frac{v_t}{1 - \beta_2^t} \\
\theta_t &= \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon},
\end{aligned}$$

where

- g_t is the gradient at time step t,
- m_t and v_t are the estimates of the first and second moments,
- β_1 and β_2 are decay rates for the moment estimates,
- ϵ is a small constant to prevent division by zero,
- α is the learning rate.

Adam is particularly effective for problems with large datasets and high-dimensional parameter spaces. It requires little memory and typically performs well with minimal parameter setting which makes it a default choice for many neural network applications.

3.4 Data Processing

In this analysis a (70%–30%) split was used to divide each dataset into training and validation subsets. The input features were standardized for stable and efficient learning. After the machine learning model was trained and applied each event was normalized to obtain the expected number of events in the detector based on luminosity and cross-section values.

For visualization purposes the signal events were scaled to a cross-section of 0.1 pb ensuring that their contribution remains within the same order of magnitude as the background.

3.5 Evaluation Metric

A commonly used evaluation metric for measuring the separation strength between signal and background is the **significance** defined as

$$Z_0 = \frac{S}{\sqrt{B}}$$

,

where:

- S is the expected number of signal events
- B is the expected number of background events

To calculate S and B, the following expressions are used

$$S = w \cdot TP, \quad B = w \cdot FP$$

,

where

- w is a normalization factor that converts the raw number of events into an expected yield based on cross-section and luminosity,
- TP (true positives) is the number of correctly classified signal events,
- FP (false positives) is the number of background events incorrectly classified as signal.

The significance Z_0 provides an approximate estimate of how likely it is that a signal can be distinguished from background purely based on statistical fluctuations. A higher Z_0 indicates stronger evidence for a true signal.

Chapter 4

Analysis

4.1 ATLAS Simulation

The signal and the background reactions were simulated by the ATLAS collaboration. The datasets were provided in Ntuple format with features describing each event. Table 4.1 lists the reactions which were taken into account for this analysis. The weighted event number corresponds to the number of events expected to be observed in the recorded data. Thus, it converts the number of simulated events to the number of expected events to be observed.

Channel	Ntuples	Preselected	Weighted
tbH^+_{250}	390883	8319	44.6
tbH^+_{800}	491327	14355	98.3
tbH^+_{3000}	287942	4766	32.5
ttZ	332456	1136	2.4
ttW	4790809	23013	68.0
ttH	10446740	67682	68.1

Table 4.1: Reactions, number of events in Ntuples, number of preselected events, weighted preselected events.

The preselection criteria was two same sign leptons, one hadronic τ, a minimum of four jets and at least one b-tagged jet[1]. The table shows that in comparison with previous analyses [46, 47], this analysis has much larger provided Ntuple dataset, for example in the previous analysis there were 834,970 events, and in the current analysis, we have 10,446,740 events.

[1]The ATLAS FastFrame tool was used to apply the preselection and create the signal and background Ntuples

4.2 Features

The input to the machine learning model were selected features stored in ROOT files. Each signal sample contain over 300 features and only 15 most important were selected. These were chosen based on the previous analyses and their influence on the classification task [46, 47]. Those features are

- `el_phi`: Azimuthal angle (ϕ) of the electron in the transverse plane.
- `el_eta`: Pseudorapidity (η) of the electron which describes its direction relative to the beam axis.
- `el_pt_NOSYS`: Transverse momentum (p_T) of the electron.
- `mu_phi`: Azimuthal angle of the muon.
- `mu_eta`: Pseudorapidity of the muon.
- `mu_pt_NOSYS`: Transverse momentum of the muon.
- `tau_phi`: Azimuthal angle of the hadronic tau.
- `tau_eta`: Pseudorapidity of the tau.
- `tau_pt_NOSYS`: Transverse momentum of the tau.
- `tau_vz`: Z-position (longitudinal coordinate) of the tau decay vertex.
- `jet_pt_NOSYS`: Transverse momentum of the leading jet.
- `jet_eta`: Pseudorapidity of the leading jet.
- `jet_phi`: Azimuthal angle of the leading jet.
- `jet_GN2v01_FixedCutBEff_70_select`: A b-tagging variable indicating whether the jet passes a fixed b-tagging efficiency threshold of approximately 70%.
- `MET_wpSet0_phi_NOSYS`: Azimuthal angle of the missing transverse energy (MET).
- `MET_wpSet0_met_NOSYS`: Magnitude of the missing transverse energy (MET) representing undetected particles such as neutrinos.

The figures below show how some of the selected features contribute to the separation between signal and background events.

Figure 4.1 shows the distribution of the transverse momentum of the leading lepton combining all tbH^+ mass points and the backgrounds (ttH, ttW, and ttZ) evaluated in the preselected region. The vertical axis is plotted on a logarithmic scale to emphasize both low and high population regions.

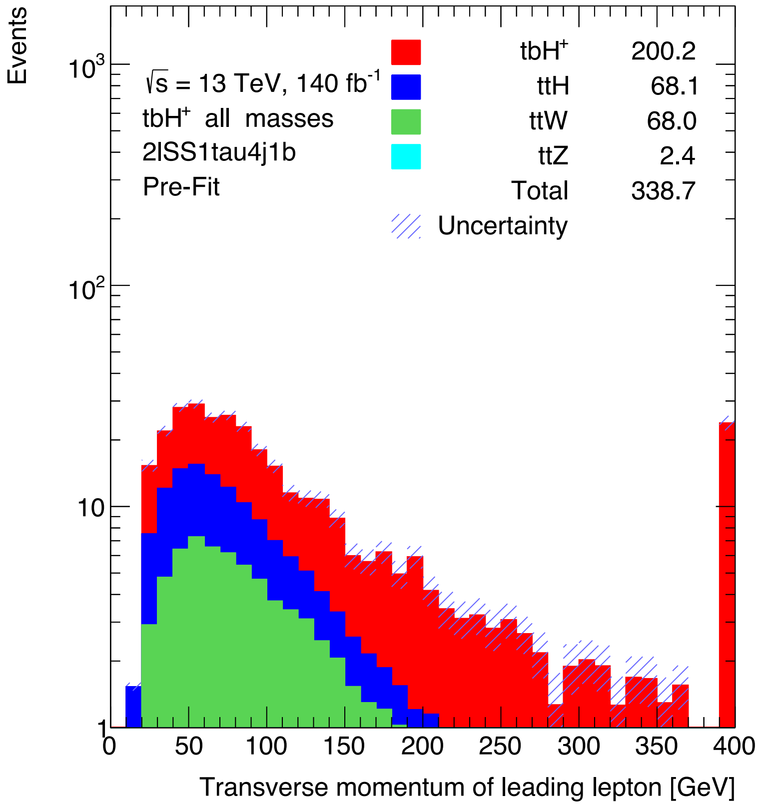

Figure 4.1: Transverse momentum of leading lepton. The three signal samples with masses 250, 800 and 3000 GeV are combined.

From the distribution above it can be seen that tbH^+ signal events tend to appear more on the right hand side. This suggests that signal events are more likely to contain high leading leptons compared to background events, particularly in the tail of the distribution ($p_T > 100$ GeV). The peak at the end is the overflow bin. All events with higher values are collected in the last bin.

Figure 4.2 shows the distribution of the transverse momentum of the leading hadronic τ candidate for the tbH$^+$ signal and background processes.

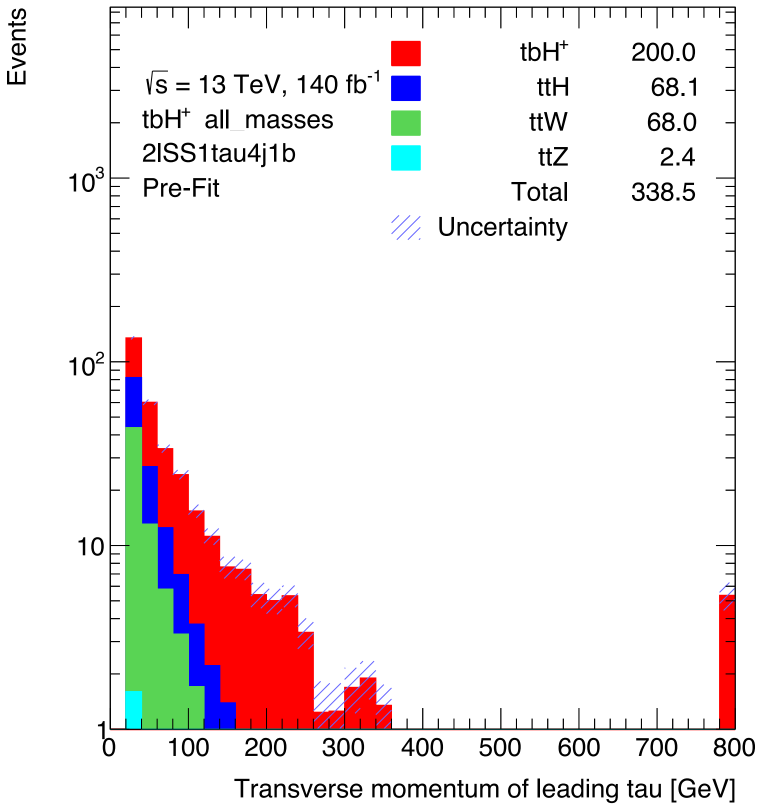

Figure 4.2: Transverse momentum of leading tau. The three signal samples with masses 250, 800 and 3000 GeV are combined.

As with the leading lepton the signal tends to have more energetic τ compared to the background particularly in the higher p_T region.

Figures 4.3 and 4.4 show the distributions of the number of b-tagged jets (nbJets) and the total number of jets (nJets) respectively for signal and background processes.

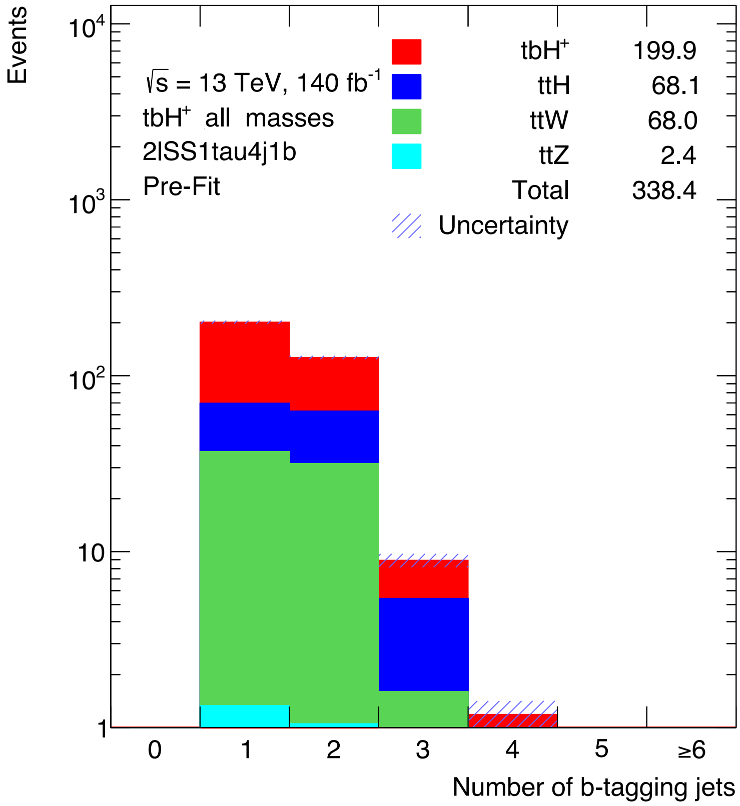

Figure 4.3: Number of b-tagging jets. The three signal samples with masses 250, 800 and 3000 GeV are combined.

The `nbJets` variable provides a decent separation between signal and background events. The tbH^{+} signal events are more likely to contain multiple b-jets due to the presence of top quark decays in the final state.

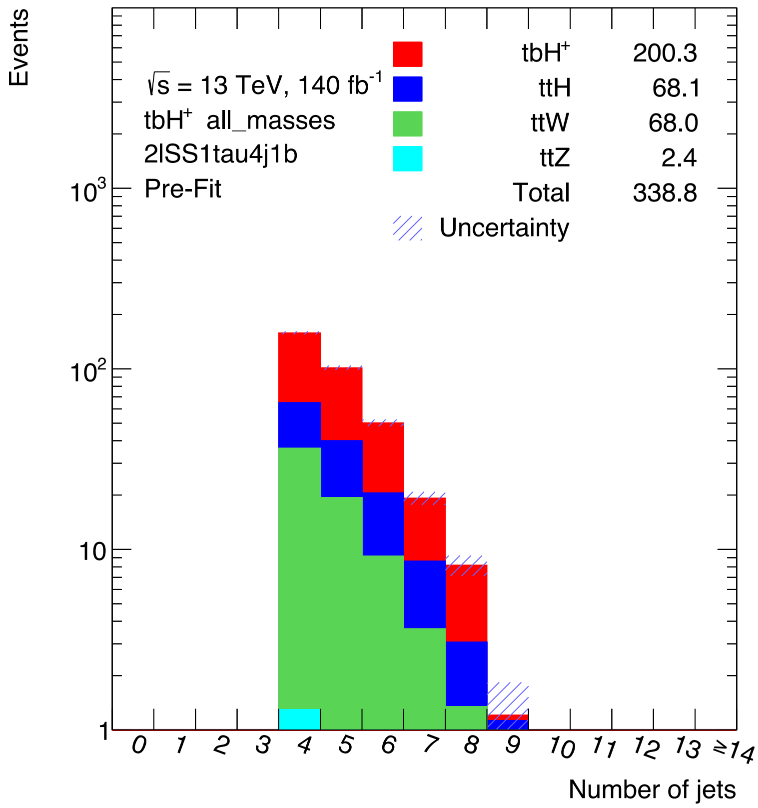

Figure 4.4: Number of jets. The three signal samples with masses 250, 800 and 3000 GeV are combined.

Similarly the `nJets` distribution in Figure 4.4 shows that tbH$^+$ events typically feature a larger overall number of jets compared to background processes. This is again due to the decay of the signal process which involves particles that decay hadronically. All those variables serve as key inputs to the machine learning classifier.

4.3 Training on separate masses

4.3.1 Network output

In the first experiment the training and testing will be done on each individual masses of tbH$^+$. Figures 4.5, 4.6 and 4.7 present the output distributions of the trained neural network for the three signal masses: 250 GeV, 800 GeV and 3000 GeV respectively. The horizontal axis represents the classifier's output score, so the prediction of an event being classified as signal ($p(y = \mathrm{tbH} \mid x)$) ranging from 0 to 1. The vertical axis shows the number of events on a logarithmic scale.

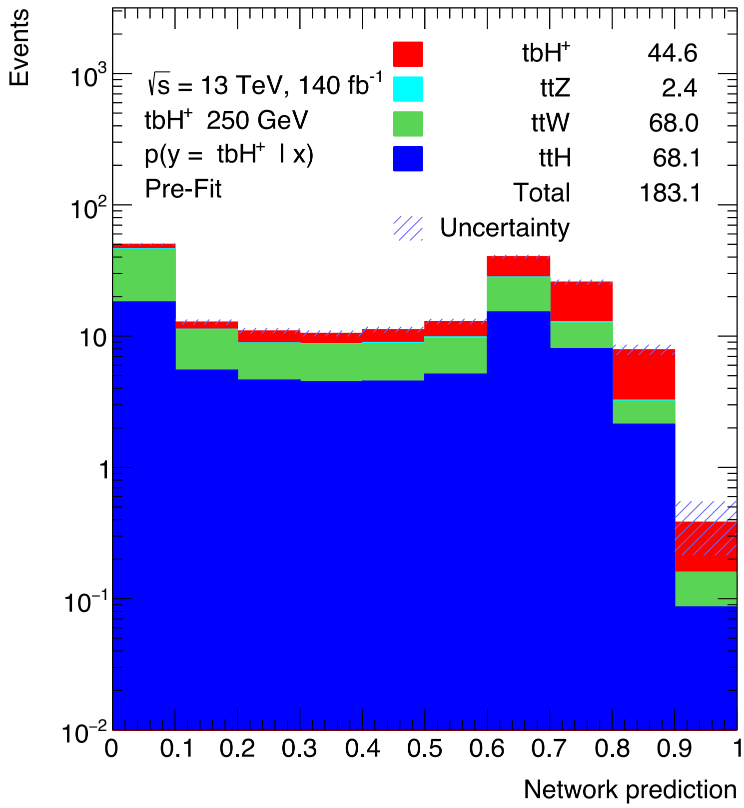

Figure 4.5: Network output for 250 GeV mass. The network training was performed on for 250 GeV mass sample separately. $\sigma(tbH^+) = 0.1$ pb assumed.

As expected signal events (red) mostly appear at the high-score region (close to 1) which indicates that the classifier can effectively assign high score to true signal events. The background processes are mostly concentrated in the low score region showing good separation between signal and background.

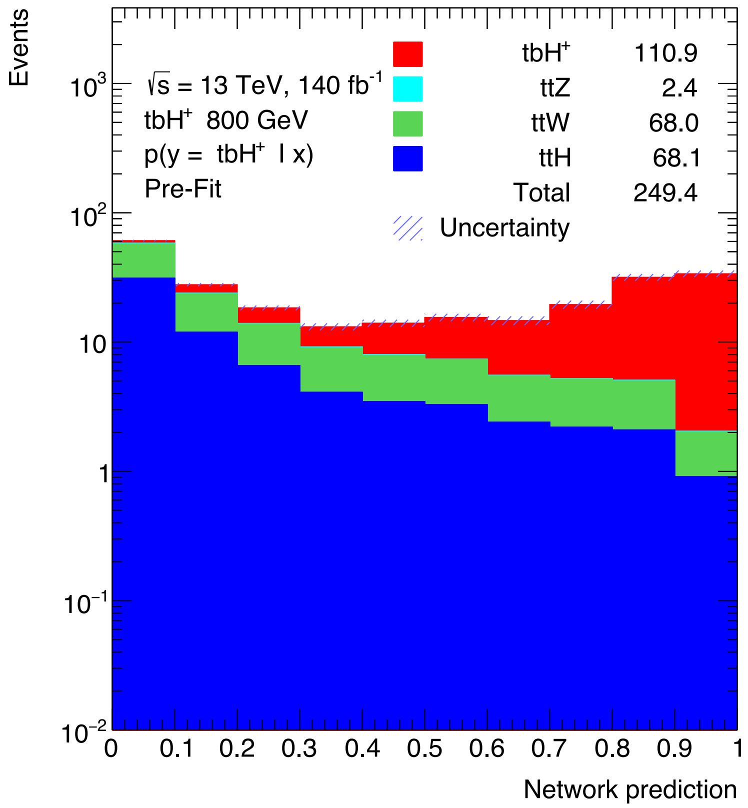

Figure 4.6: Network output for 800 GeV mass. The network training was performed on for 250 GeV mass sample separately. $\sigma(tbH^+) = 0.1$ pb assumed.

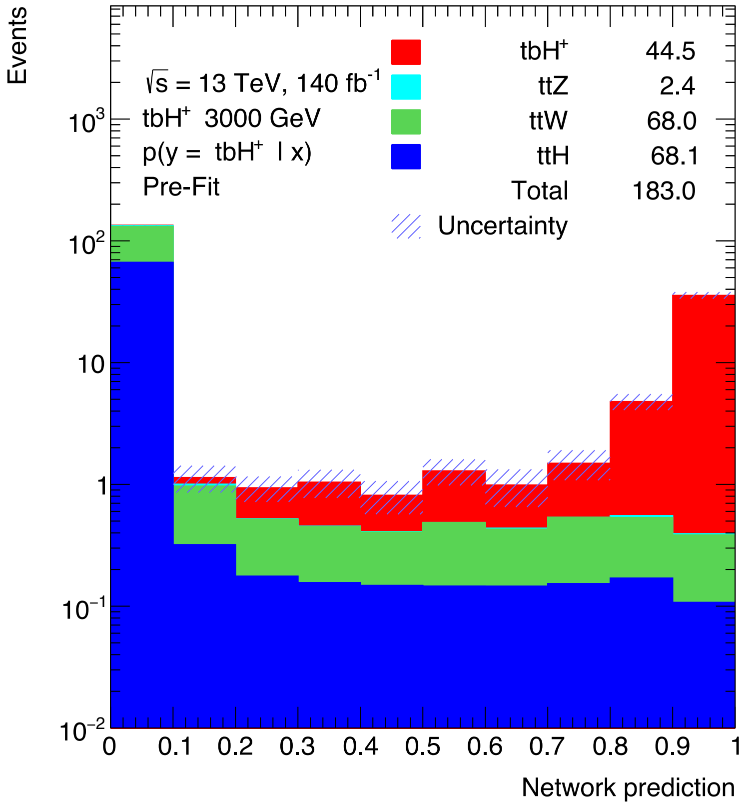

Figure 4.7: Network output for 3000 GeV mass. The network training was performed on for 250 GeV mass sample separately. $\sigma(tbH^+) = 0.1$ pb assumed.

The separation becomes more clear for higher mass signals particularly in the 3000 GeV mass region where the signal dominates the highest-score bins. This reflects the classifier's ability to find better differences between the signal and background at higher mass regions. This occurs because the particles in the final state of the signal have kinematic properties that are very similar to those of particles produced in the background processes.

Figure 4.8 shows the composition of background events that were assigned a high score by the neural network. The most contribution come from the ttW and ttH processes with ttZ being less significant.

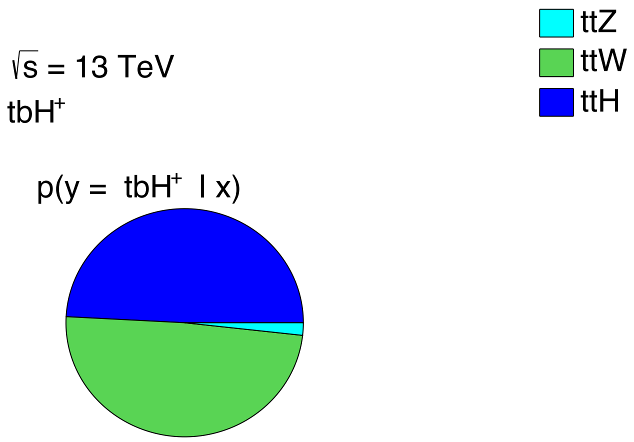

Figure 4.8: Background composition.

4.3.2 Expected limit

Figure 4.9 shows the expected 95% confidence level (CL) upper limits on $\sigma(pp \to tbH^+) \times \mathrm{BR}(H^+ \to \tau\tau)$ for the previous analysis. It includes several charged Higgs boson masses: 250, 300, 800, 1500, 2000, and 3000 GeV. These limits were derived with the same luminosity of 140 fb^{-1} at a center-of-mass energy of $\sqrt{s}$ = 13 TeV. The results were obtained using the asymptotic approximation.

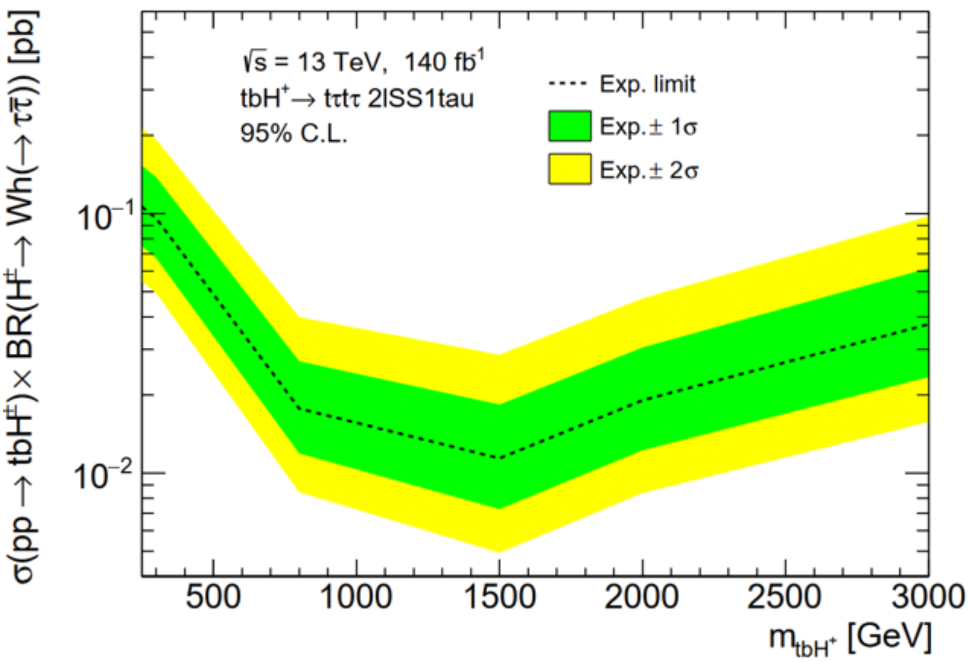

Figure 4.9: Expected 95% confidence level (CL) upper limits on the cross-section for tbH$^+$ production with statistical uncertainties for release 21[9].

The dashed line represents the median expected exclusion limit while the green and yellow bands correspond to the $\pm 1\sigma$ and $\pm 2\sigma$ intervals respectively.

The lowest expected limit is observed near 1500–2000 GeV, indicating the highest sensitivity of the previous analysis in this mass region.

Figure 4.10 shows the expected 95% CL upper limits on $\sigma(pp \to tbH^+) \cdot \mathrm{BR}(H^+ \to \tau\tau)$ for our charged Higgs boson masses: 250 GeV, 800 GeV, and 3000 GeV using statistical uncertainties.

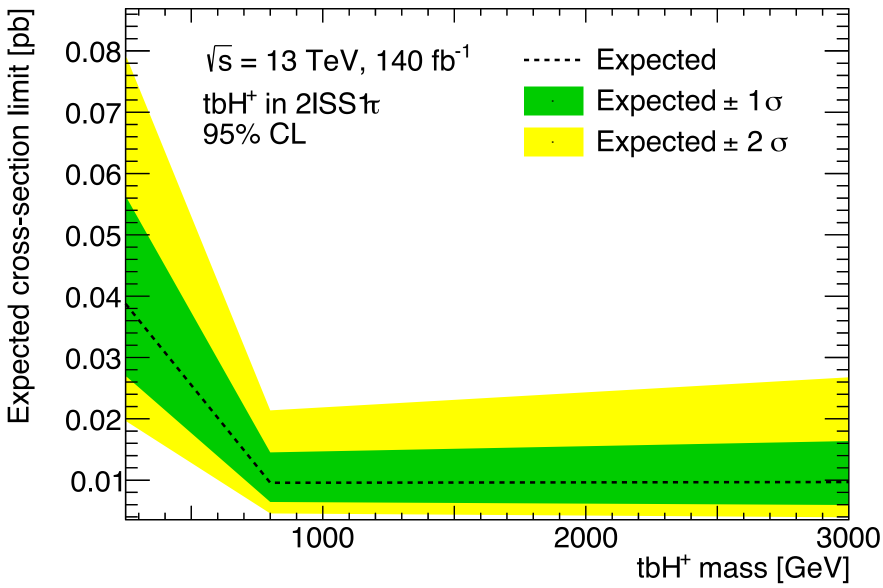

Figure 4.10: Expected 95% confidence level (CL) upper limits on the cross-section for tbH$^+$ production with statistical uncertainties for release 22 for separate mass training.

The expected limits are:

- **250 GeV:** 0.039 pb
- **800 GeV:** 0.010 pb
- **3000 GeV:** 0.010 pb

We can see the improvement in the sensitivity in high mass regions. However, the result for the 250 GeV mass point shows a higher limit which is likely due to the current incompleteness of the feature set used for classification and similarity with background events.

This highlights the importance of extending and optimizing the input variables used in the model to increase discrimination power across all signal regions. For example, a higher weight could be given to lower mass events in the training.

4.3.3 Loss function development

Figure 4.11 shows the development of the binary cross-entropy loss during the training for signal with a mass of 250 GeV. The training loss rapidly decreases in the initial epochs and stabilizes at a low value (below 0.1), indicating that the model successfully fits the training data.

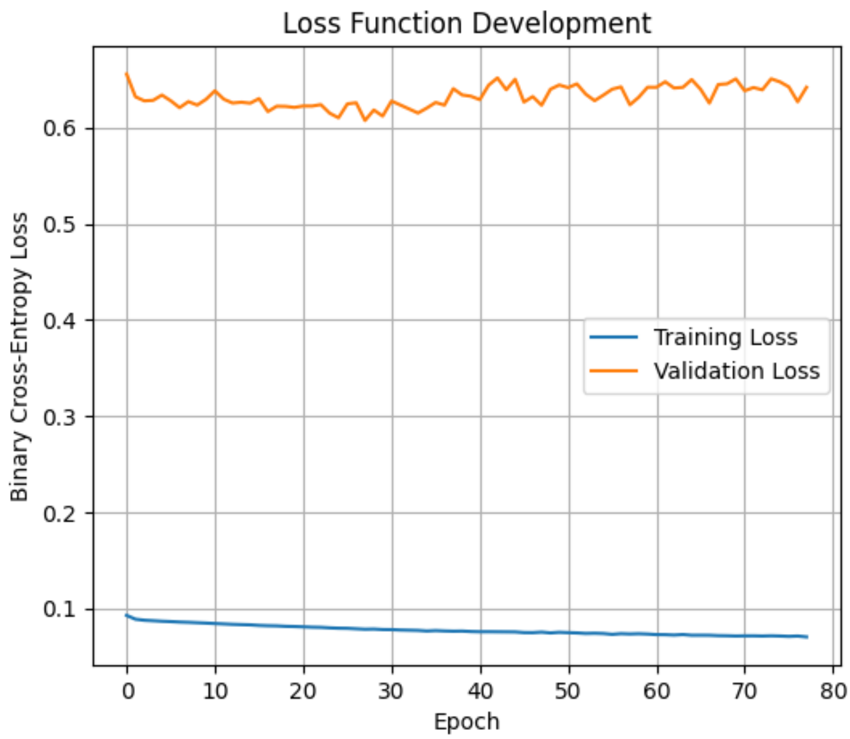

Figure 4.11: Loss function development for 250 GeV.

The validation loss remains significantly higher than the training loss, fluctuating between 0.6 and 0.7 over the training. This behavior suggests that the set of input features may not be sufficient to effectively distinguish the 250 GeV signal from the background. The other reason might be due to the fact that 250 GeV signal behaves similar to the background events. Figures 4.12 and 4.13 presents the loss function development for the mass points: 800 GeV and 3000 GeV.

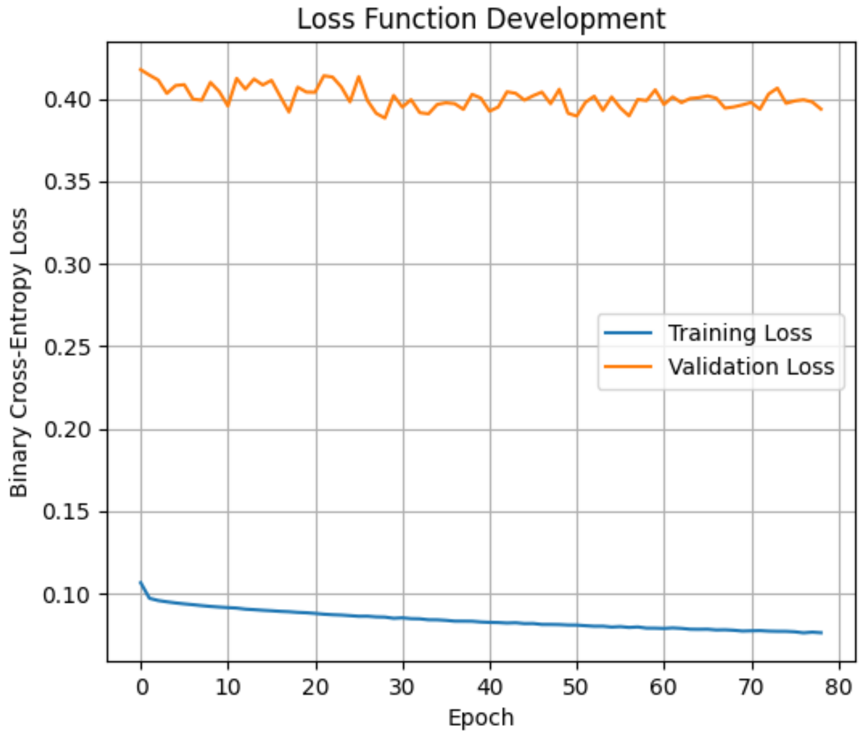

Figure 4.12: Loss function development for 800 GeV.

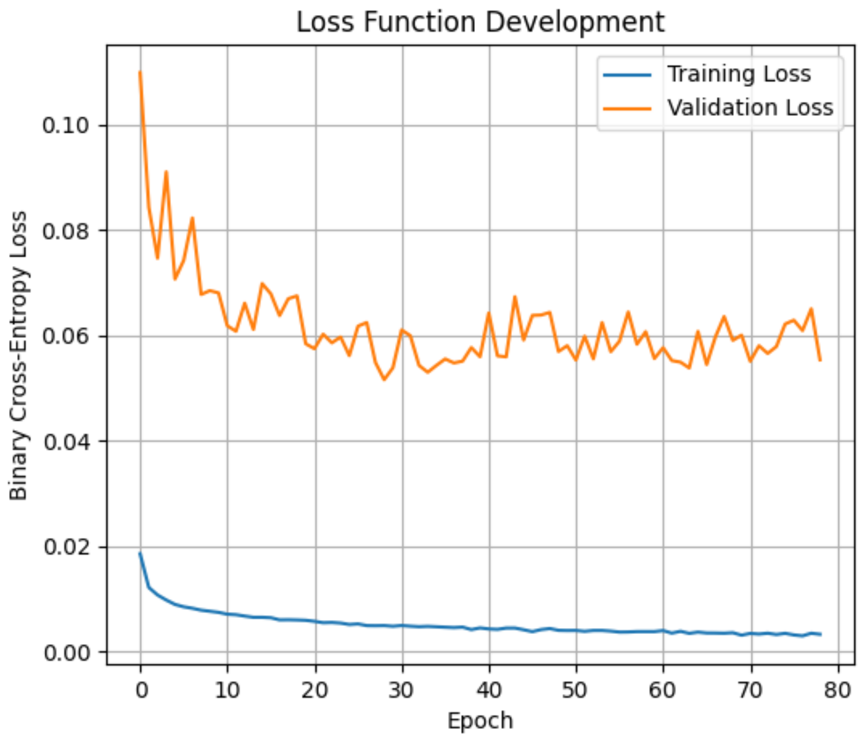

Figure 4.13: Loss function development for 3000 GeV.

At **800 GeV**, the training loss decreases quickly and stabilizes at a value below 0.1. However, the validation loss remains consistently higher fluctuating around 0.43–0.45 with no significant downward trend.

At **3000 GeV**, the model demonstrates much better generalization. The training loss drops below 0.01, and the validation loss remains stable around 0.06–0.08 with some fluctuations. This improved behavior may result from stronger signal background separation at higher mass points, where the kinematic features of signal events are more distinct.

Overall, the model shows more robust learning behavior and generalization at 800 and 3000 GeV compared to 250 GeV highlighting the influence of signal mass on classification difficulty.

4.3.4 ROC curves

The Receiver Operating Characteristic (ROC) [48] curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 – specificity) at various classification thresholds. The curve is a graphical representation of a binary classifier's performance across all possible decision thresholds. The area under the ROC curve (AUC) shows the overall ability of the model to distinguish between the two classes (signal and background). An AUC value of 1.0 corresponds to a perfect classifier, while a value of 0.5 indicates random guessing. A steeper curve that rises quickly toward the top-left corner indicates better classification performance. The ROC curves for different signal masses are shown in Figures 4.14, 4.15, and 4.16.

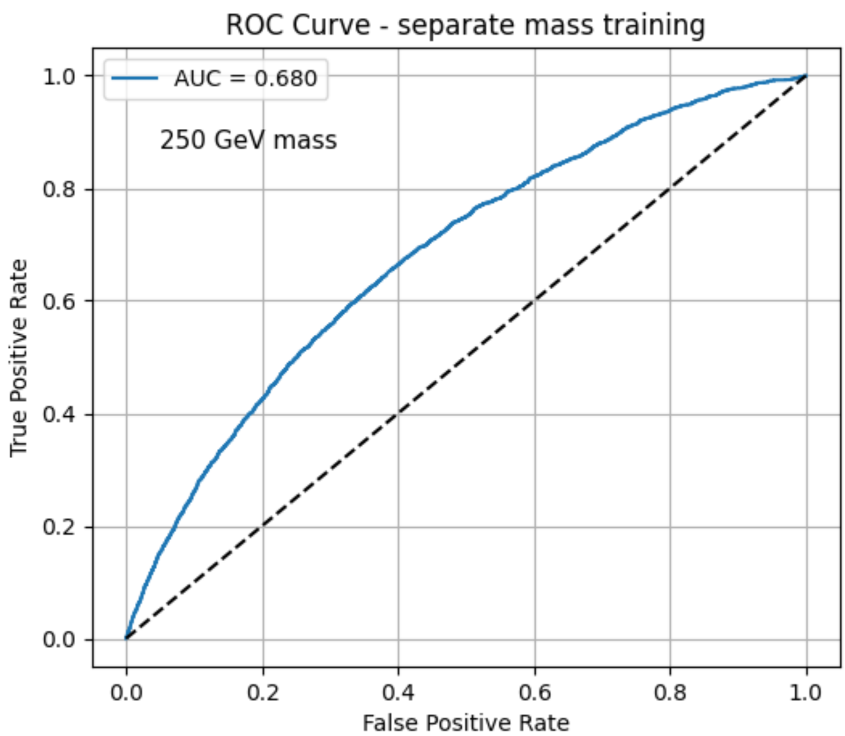

Figure 4.14: ROC curve for 250 GeV in separate mass training.

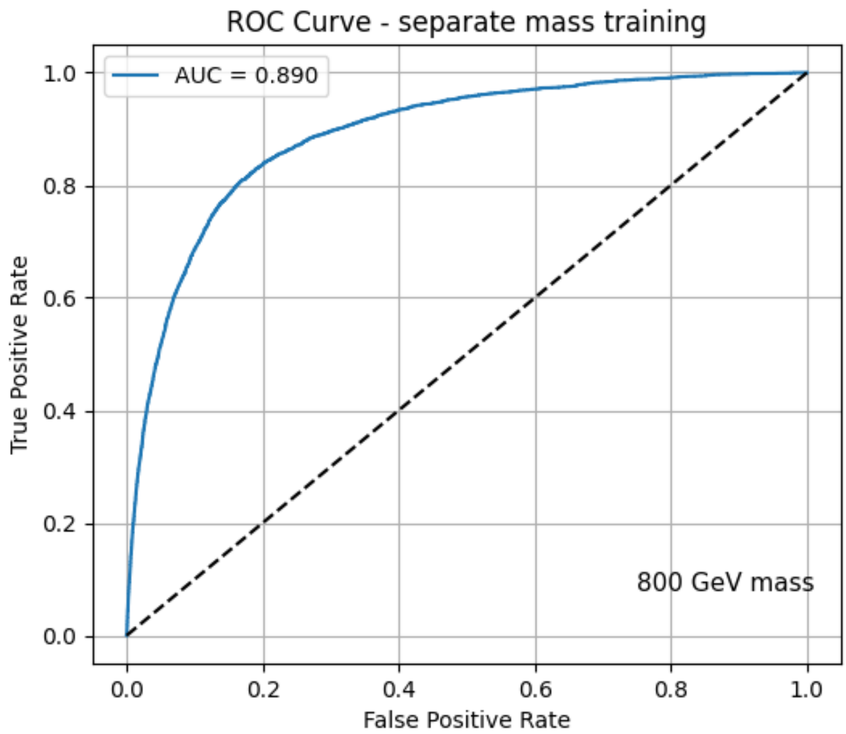

Figure 4.15: ROC curve for 800 GeV in separate mass training.

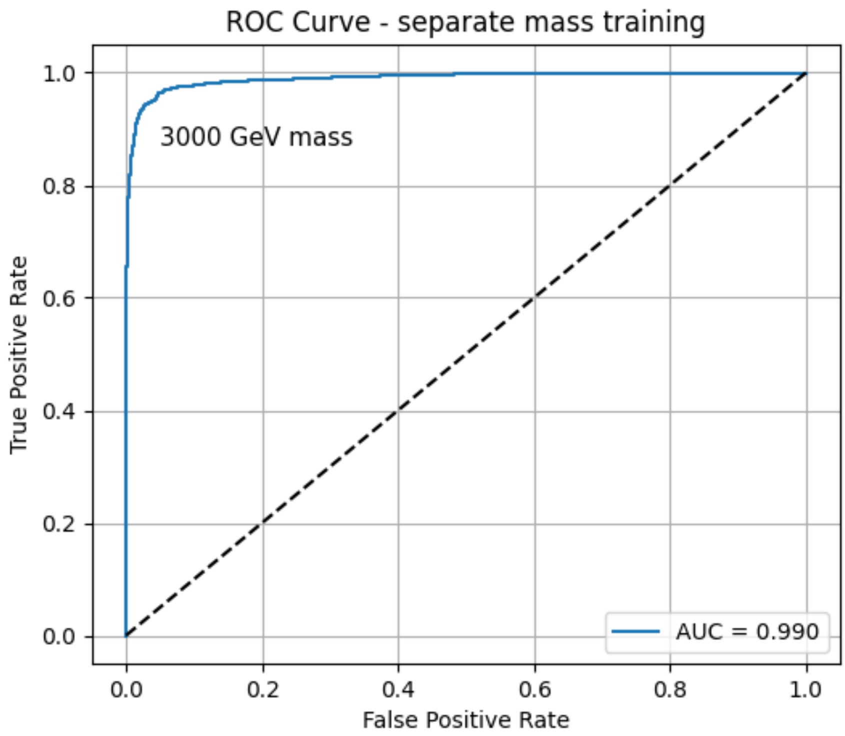

Figure 4.16: ROC curve for 3000 GeV in separate mass training.

- **250 GeV:** The ROC curve for the 250 GeV signal yields an AUC value of 0.680. This relatively low AUC indicates that the model struggles to distinguish the signal from the background effectively at low mass values.
- **800 GeV:** The ROC curve for the 800 GeV signal shows an improvement with an AUC value of 0.890. This demonstrates a strong capability of the model to separate signal and background.
- **3000 GeV:** For the 3000 GeV mass, the ROC curve achieves an AUC value of 0.990. This very high AUC indicates very good separation between signal and background. This reflects that the high mass signals are very distinct from the background and are easier for the model to classify correctly.

Overall, the model's performance improves with increasing signal mass. This highlights the fact that distinguishing high mass signals is easier due to more kinematic differences relative to background processes.

4.4 Combined mass training

In this section we focus on the second experiment where we are going to combine all signal masses into single set during the training phase. The training is performed on this set together with the background events. During the testing phase the signal events are separated based on their mass. Figures 4.17, 4.18 and 4.19 show the output distribution of the network.

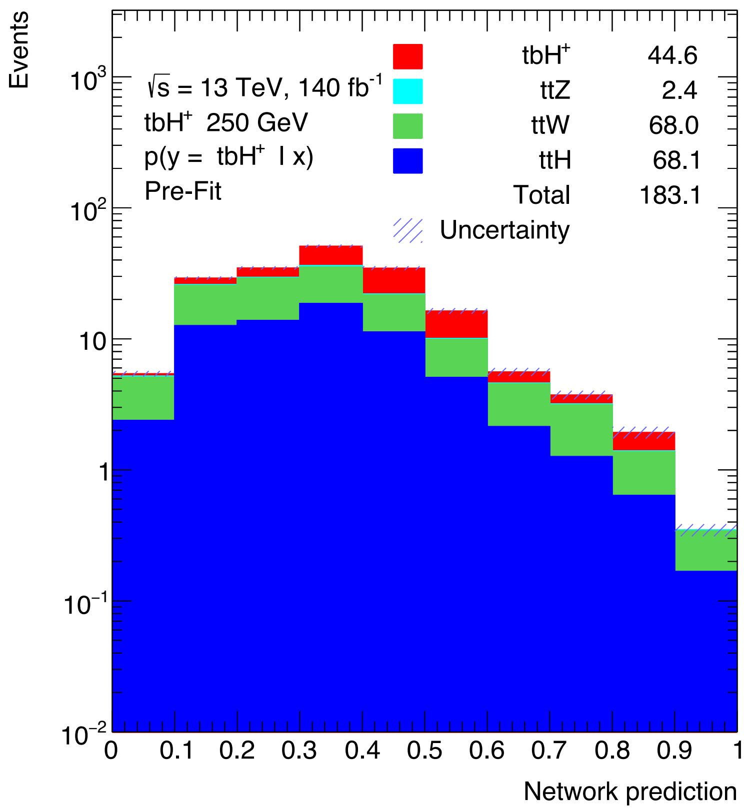

Figure 4.17: Network for 250 GeV mass. The network training was performed on combined masses. $\sigma(tbH^{+}) = 0.1$ pb assumed.

For the low mass region the network prediction in the combined mass training peaks at intermediate values. In the range 0.4–0.7 with only a small number of events reaching values above 0.8. The background remains dominant in the lower part of the spectrum (below 0.5) resulting in an overlap between signal and background. This overlap suggests that the classifier has limited ability to distinguish low mass signals from background when trained on a mixture of signal masses. Compared when the model was trained only on the 250 GeV the separation was better with the signal values shifting towards to 1.

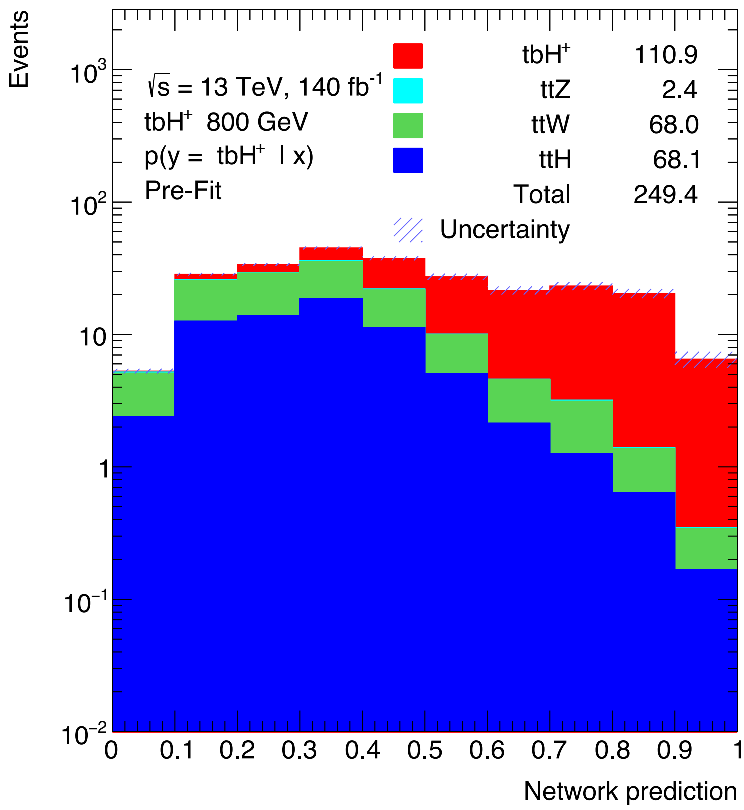

Figure 4.18: Network for 800 GeV mass. The network training was performed on combined masses. $\sigma(tbH^+) = 0.1$ pb assumed.

For the intermediate mass point the signal distribution shows better results compared to the 250 GeV. The majority of signal events lie within the 0.4–0.9 interval with a high number of events reaching the highest value bins above 0.8. This indicates a better separation from the background which is concentrated below 0.5. The signal to background classification is improved in this region even under the combined mass training. Compared to training on a single mass, the performance remains relatively robust suggesting that the features of 800 GeV signal are more distinguishable and less affected by the presence of other masses during training.

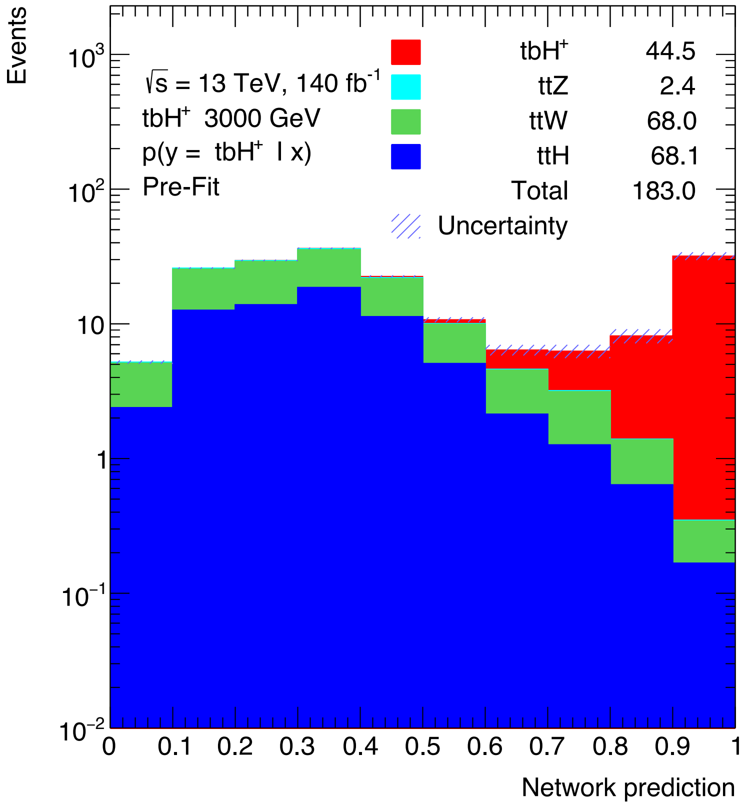

Figure 4.19: Network for 3000 GeV mass. The network training was performed on combined masses. $\sigma(tbH^+) = 0.1$ pb assumed.

In the high mass region the signal distribution shows a clear peak in the highest range (above 0.9) which indicates strong separation from the background. Most background events remain in the lower interval (below 0.5) with some overlap with the signal. This separation demonstrates that the model successfully distinguishes features of the high mass signal even in the combined training scenario. Compared to separate mass training the performance remains stable or even slightly improved showing that high mass signals are easier to distinguish due to their more unique kinematic signatures.

4.4.1 ROC curves

Figures 4.20, 4.21 and 4.22 show ROC curves on individual signal masses after the model was trained on a dataset combining all three mass points.

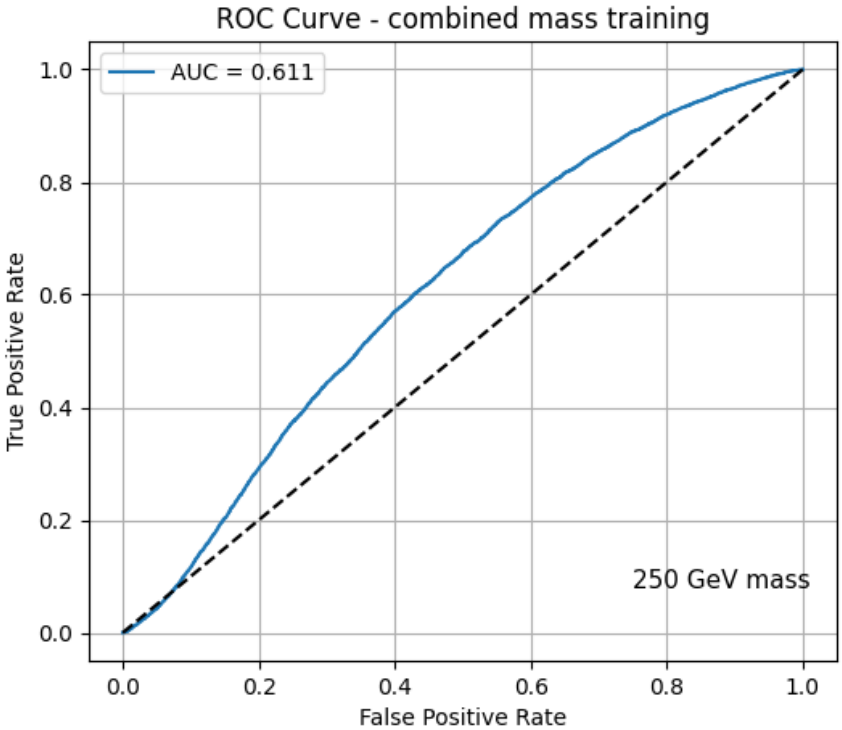

Figure 4.20: ROC curve for 250 GeV in combined mass training.

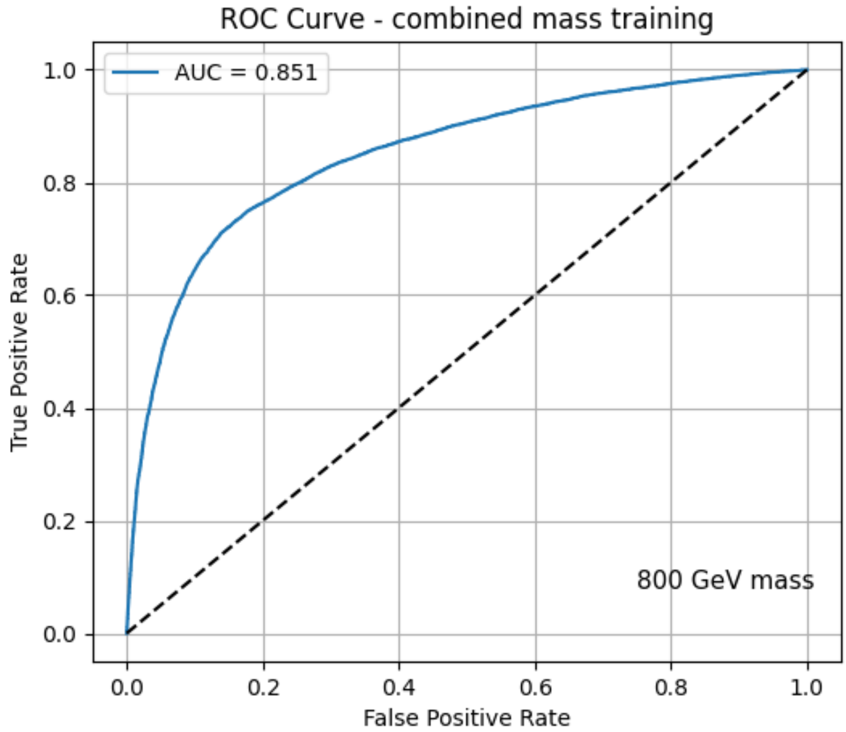

Figure 4.21: ROC curve for 800 GeV in combined mass training.

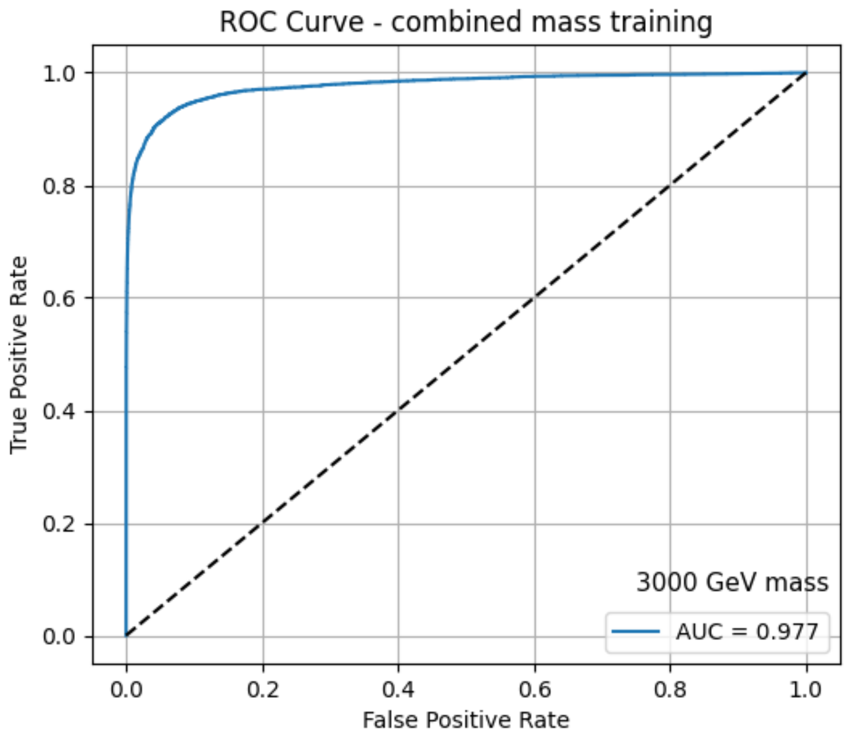

Figure 4.22: ROC curve for 3000 GeV in combined mass training.

- **250 GeV:** The AUC dropped to **0.611**, which is significantly lower than in the separate mass training. This indicates that the classifier struggles even more to identify low-mass signals when trained on a mixed dataset.
- **800 GeV:** The AUC is 0.851, slightly lower than in the separate training.
- **3000 GeV:** The AUC remains high at 0.977, although it is slightly reduced compared to the separate mass training.

These results demonstrate that combining different signal mass points during training leads to a decrease in performance especially for lower mass signals.

4.4.2 Expected limits

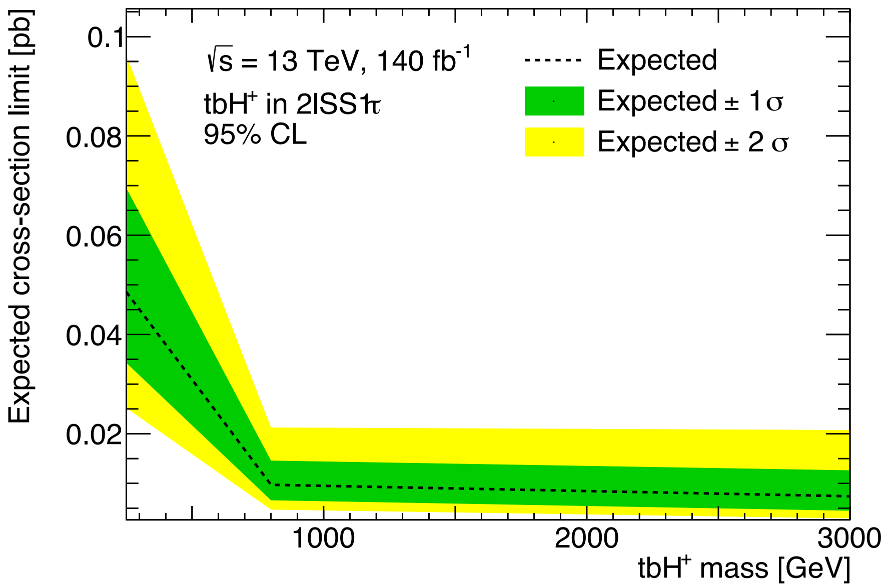

Figure 4.23: Expected 95% confidence level (CL) upper limits on the cross-section for tbH^+ production with statistical uncertainties for release 22 for combined mass training.

The expected limits are:

- **250 GeV:** 0.049 pb
- **800 GeV:** 0.010 pb
- **3000 GeV:** 0.007 pb

The expected limit for 250 GeV increased to 0.049 pb which indicates decrease in sensitivity. However, for 800 GeV it stayed the same and for 3000 GeV it decreased to 0.007 pb.

It can be seen that training the classifier on combined signal masses leads to a trade-off between generalization and performance. While the AUC values decreased for all individual mass points compared to separate mass training the expected limits remained stable or even improved for higher masses. This suggests that combined training provides a more robust model particularly for heavy signals, but at the cost of reduced sensitivity to low mass signals. Now, what if we increase the weight of low mass signal?

4.5 Combined mass training with increased weight of low region

In this section we focus on the third experiment where all signal masses are still combined into single set during the training phase. Due to the fact that for a model it is more difficult to separate low mass region from the background, the weight is increased by 10 times for 250 GeV. Figures 4.24, 4.25 and 4.26 show the output distribution of the network.

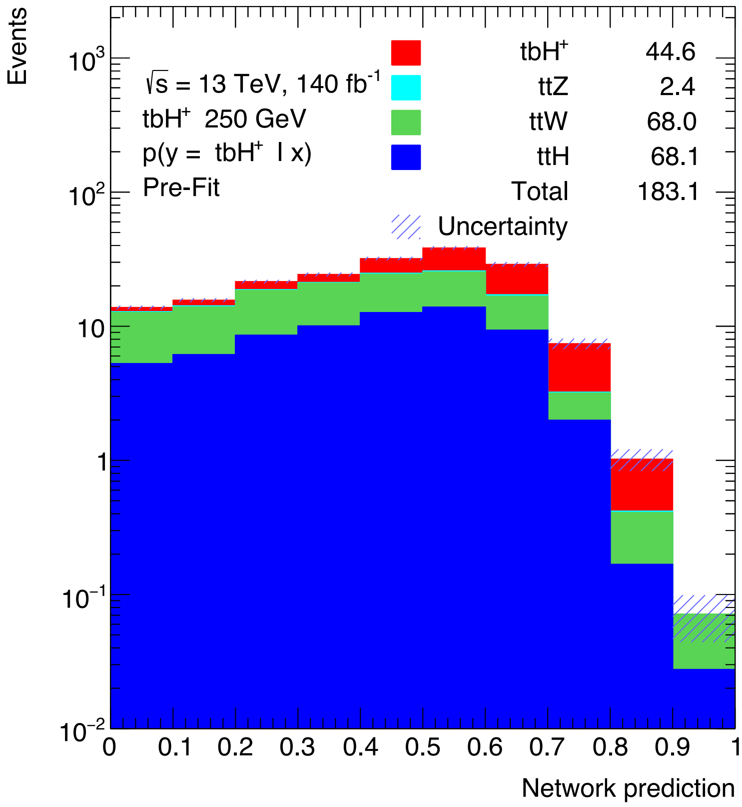

Figure 4.24: Network for 250 GeV mass. The network training was performed on combined masses with increased weight in low region. $\sigma(tbH^{+}) = 0.1$ pb assumed.

The result for the low mass region shows that the signal distribution peaks in the middle value range (0.5–0.8) with only a small fraction of events reaching the highest value bins. The background remains dominant in the lower scores (0–0.5). The overlap between signal and background shows that distinguishing low mass signals is challenging especially in the combined mass training.

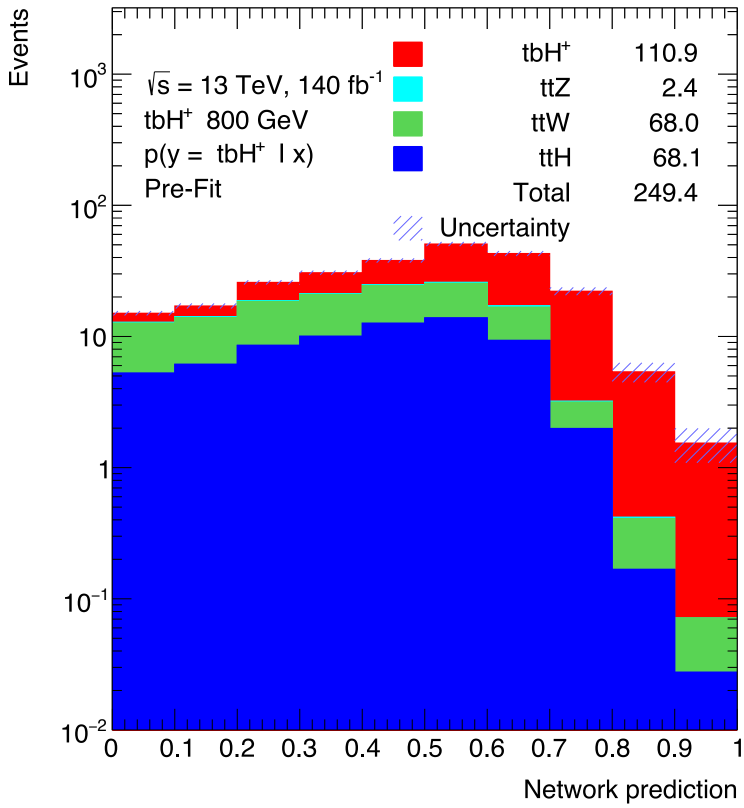

Figure 4.25: Network for 800 GeV mass. The network training was performed on combined masses with increased weight in low region. $\sigma(tbH^+) = 0.1$ pb assumed.

In the mid mass region the model performance improves significantly. We can see a lot of signal events is observed in the high value bins ($p > 0.7$) while the background remains mostly in the lower bins. The signal is more distinct, because of more notable kinematic features compared to the background. This indicates that the combined training scheme is effective in this mass region.

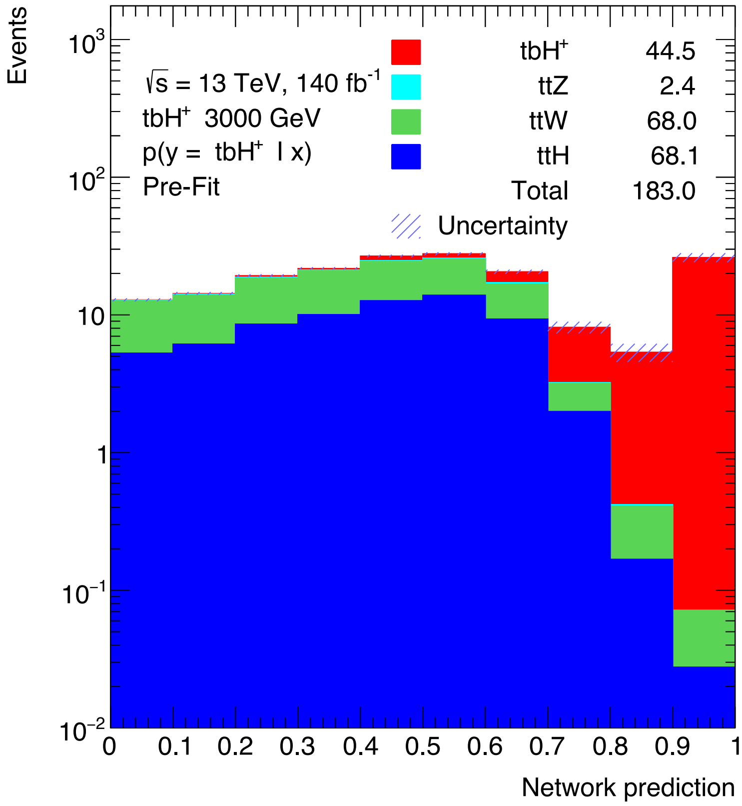

Figure 4.26: Network for 3000 GeV mass. The network training was performed on combined masses with increased weight in low region. $\sigma(tbH^{+}) = 0.1$ pb assumed.

The result for the high mass region shows great separation between signal and background. The signal distribution peaks sharply in the highest value bins (0.9–1.0) with minimal background processes. This suggests that the model can confidently identify high mass tbH^{+} events, because of their strong difference from typical background events.

The results demonstrates that even when using a combined mass training strategy with increased weighting applied to the low mass region it remains challenging for the model to efficiently separate those events from the background. However, the classifier shows better performance for the mid and high mass regions.

4.5.1 ROC curves

In this section we again evaluate the performance of the model, but on a combined dataset. The figures 4.27, 4.28 and 4.29 are the ROC curves.

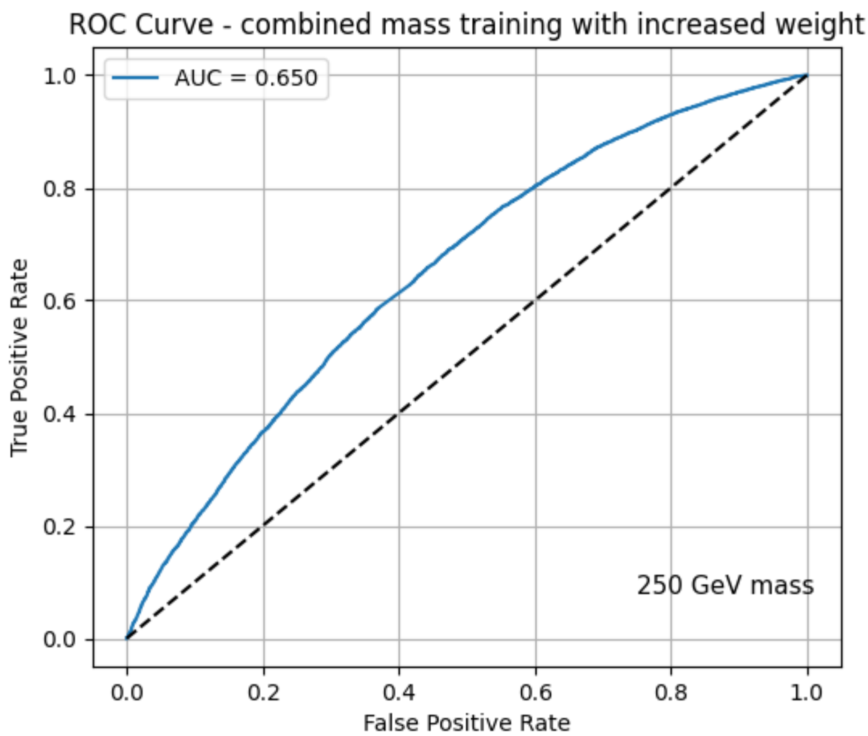

Figure 4.27: ROC curve for 250 GeV in combined mass training with increased weights.

The ROC curve in Figure 4.27 shows the model's performance on 250 GeV. The resulting AUC is 0.650. This value indicates that the model struggles to separate low mass signal events from the background even though those events were up weighted during training. The features of low mass signals appear less distinguishable from those of background processes resulting in higher overlap and reduced discrimination power. However, the AUC increased a bit compared to Figure 4.20.

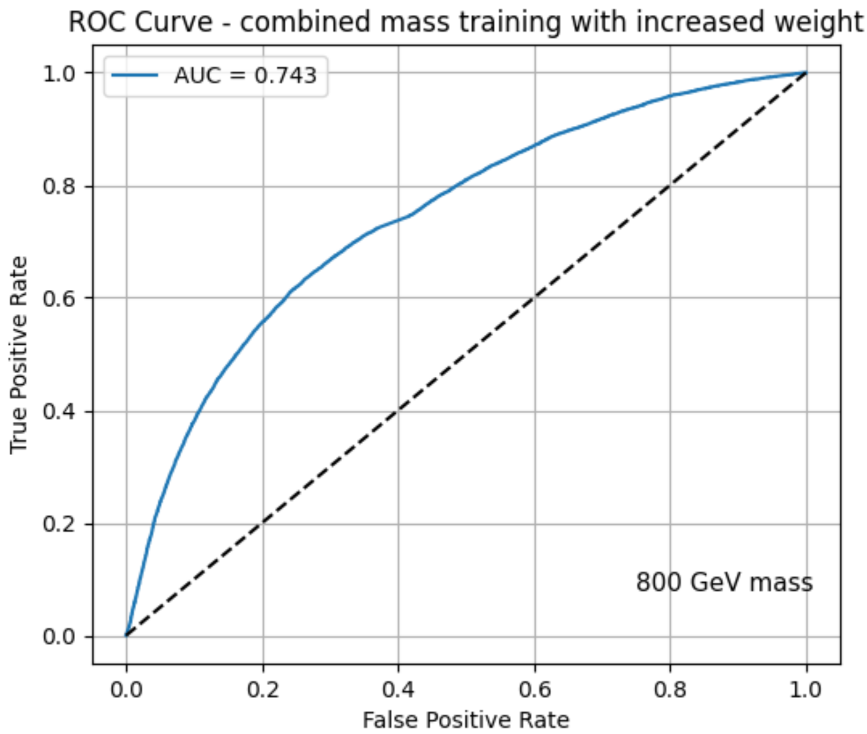

Figure 4.28: ROC curve for 800 GeV in combined mass training with increased weights.

The ROC curve for the 800 GeV demonstrates improved performance compared to the 250 GeV. The AUC reaches 0.743 indicating better separation between signal and background. However, the AUC decreased compared to Figure 4.21

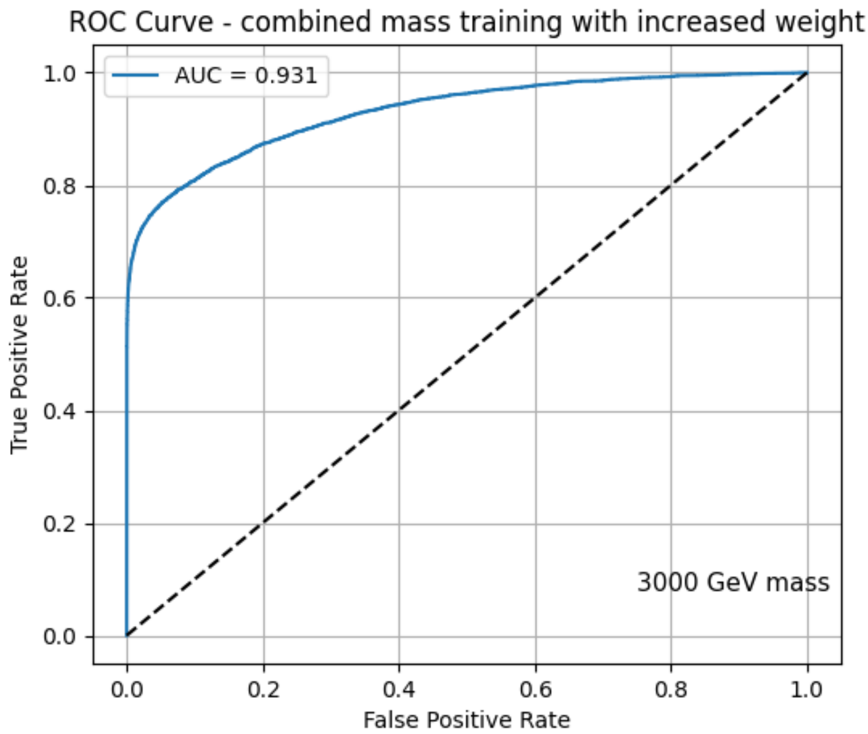

Figure 4.29: ROC curve for 3000 GeV in combined mass with increased weights.

The final plot shows the ROC curve for the high-mass region. The model performs better here with an AUC of 0.931, but still decreased value compared to Figure 4.22.

Compared to other two experiments where signal masses were trained and tested separately and combined without increasing weights of low mass signal we see a change in model performance. In this training scenario the AUC for the low-mass region slightly increases compared to previous experiment indicating a small improvement in the performance. This is likely due to the added weight during training. However, for the mid mass and high mass regions we see a decrease in AUC values.

4.5.2 Expected Limit

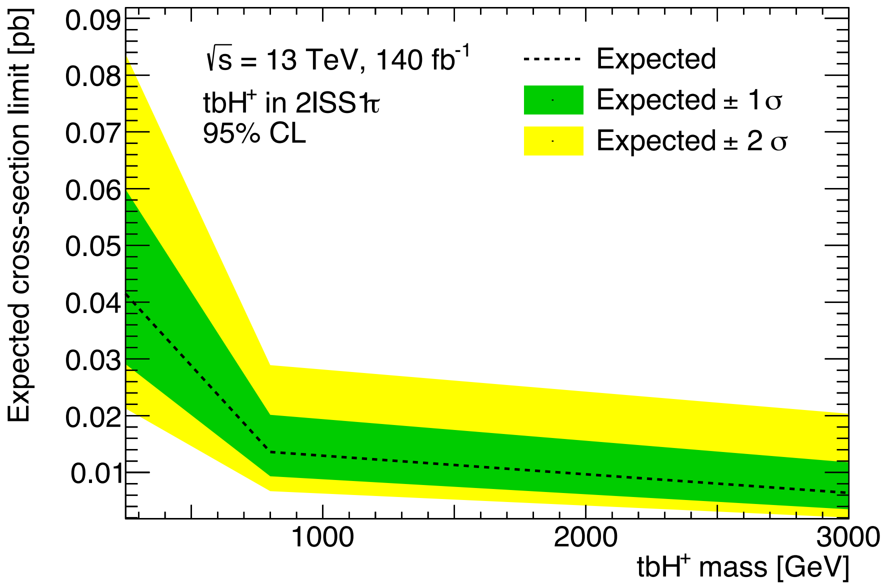

Figure 4.30: Expected 95% confidence level (CL) upper limits on the cross-section for tbH^+ production with statistical uncertainties for release 22 for combined mass training with increased weights.

The expected limits are:

- **250 GeV:** 0.041 pb
- **800 GeV:** 0.014 pb
- **3000 GeV:** 0.006 pb

As we can observe the expected limit for the 250 GeV signal improved to 0.041 pb, the limit for 800 GeV slightly increased to 0.014 pb and the limit for 3000 GeV improved again to 0.006 pb.

4.6 Comparison of training methods

Training strategy	**250 GeV**	**800 GeV**	**3000 GeV**
Combined mass training	0.611	0.851	0.977
Combined mass (increased 250 weight)	0.650	0.743	0.931
Separate mass training	0.680	0.890	0.990

Table 4.2: AUC values for different training strategies and signal masses.

- **Separate mass training** yields the highest AUC values across all signal masses. When the model is trained only on one mass it focuses on distinguishing that specific signal from background. Since the task is clearly defined the model learns to confidently separate signal events assigning high values to signal like features.
- **Combined mass training** reduces overall performance especially for low mass signals. When all masses are combined during training the signal becomes a broad heterogeneous class. Due to the fact that input feature distributions differ significantly between masses the model is forced to generalize. In training features of the low mass signal do not stand out as strongly against the background. Due to that they are less distinct than those of higher mass signals.
- **Combined training with increased weight on 250 GeV** improves the AUC and expected limit for 250 GeV, but at the cost of decreasing performance for 800 and 3000 GeV. While reweighting reduces the training loss for 250 GeV it does not lead the model to assign higher values to these events. The model simply focuses more on reducing the loss for events without making them appear more signal-like in output.

In summary, separate mass training provides the most accurate classification. Combined training introduces generalization that is beneficial only at high masses. Attempts to reweight low mass signals improves sensitivity a bit, but it cannot fully compensate for the feature overlap with background.

4.7 Feature importance

In this section we will determine the most important features which influenced our classifier. To achieve this we use SHAP analysis, permutation importance and correlation matrices to study both the individual impact and interrelationships among the features.

4.7.1 The SHAP analysis

Figures 4.31, 4.32 and 4.33 show the SHAP (SHapley Additive exPlanations) [49] values for the top 10 most important features determined by the classifier for the signal mass points. The SHAP values provide a measure of feature importance by evaluating how much each feature contributes to the predicted output of the model for a given event, where

- A positive SHAP value indicates that the feature increases the predicted score of the event being classified as signal
- A negative value mean a contribution towards the background classification.

Each boxplot shows the distribution of SHAP values for one feature across the validation dataset. Key components of each boxplot

- **Box width**: The width of the boxes represents the interquartile range (IQR) spanning from the first quartile (Q1) to the third quartile (Q3) and thus covering the middle 50% of the SHAP values.
- **Orange line**: The orange horizontal line in each box is the median SHAP value for that feature.
- **Whiskers**: The whiskers extend to 1.5 times the IQR beyond Q1 and Q3 and any values outside this range are plotted as outliers.

The different widths of the boxes represent the variability of the feature's contribution: wider boxes indicate greater variability in how the feature influences the model's predictions across different events.

The features are sorted based on the **median** of their SHAP values. This is because we want to capture the typical contribution of each feature to the model's prediction. This approach highlights features that consistently contribute to the classification decision.

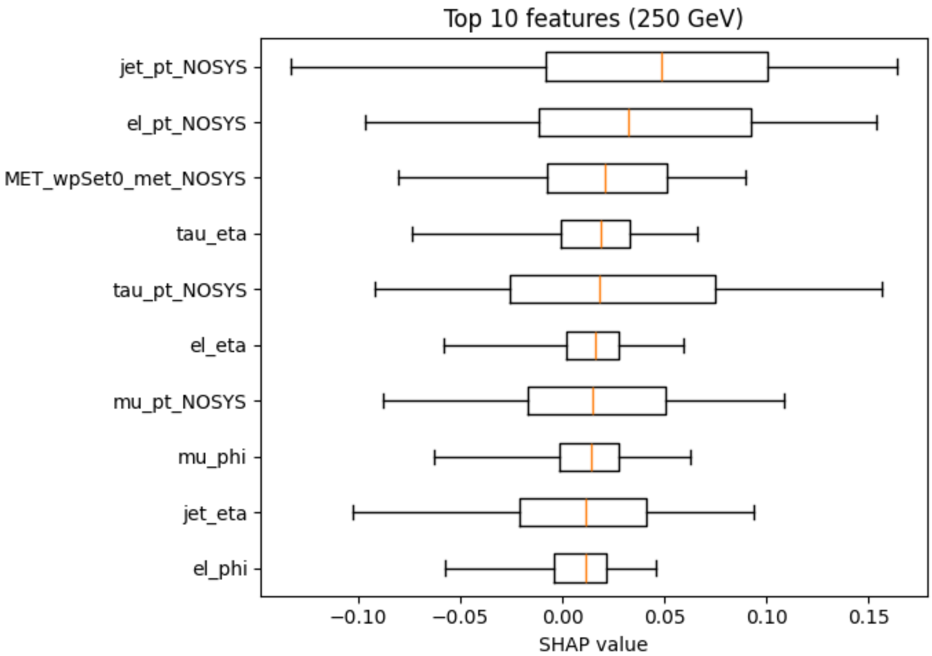

Figure 4.31: Top 10 most important features for 250 GeV using SHAP analysis.

Key Observations for 250 GeV:

- **Dominant Features**:
  - `jet_pt_NOSYS` shows the most significant contribution with a strong positive SHAP value. This makes it the top feature for the signal.
  - `el_pt_NOSYS` is the second-most influential feature also with stable positive impact across events.
- **MET Characteristics**:
  - `MET_wpSet0_met_NOSYS` shows moderate positive influence. This indicates that the transverse MET magnitude contributes meaningfully to signal classification.
  - The direction of MET (`phi`) is not among the top features, suggesting that the signal relies more on MET magnitude than direction.
- **Lepton and Tau Kinematics**:
  - `tau_eta`, `tau_pt_NOSYS`, and `el_eta` appear mid-ranked, with noticeable but weaker contributions.
  - `mu_pt_NOSYS` and `mu_phi` show some relevance but are less consistent across events.

- **Lesser Impact Features**:
  - `jet_eta` and `el_phi` complete the top-10 but exhibit relatively small median SHAP values, suggesting minor influence.
  - Notably, `tau_phi` and `jet_GN2v01_FixedCutBEff_70_select` do not appear in the top features, indicating low importance in this mass region.

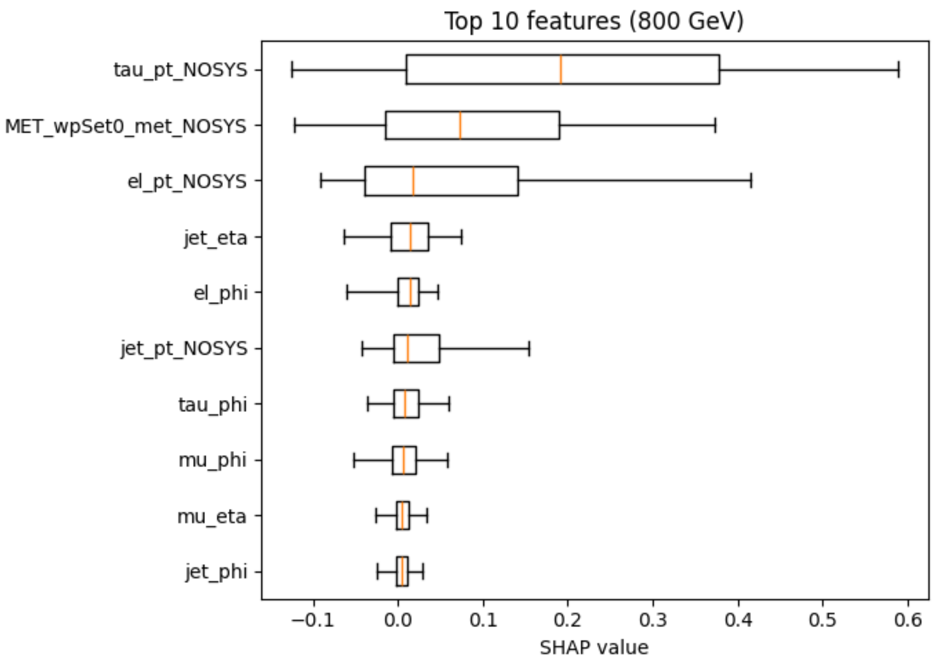

Figure 4.32: Top 10 most important features for 800 GeV using SHAP analysis.

Key Observations for 800 GeV:

- **Dominant Features**:
  - `tau_pt_NOSYS` stands out as the most impactful feature. It is consistently contributing positively to signal classification with a high median SHAP value.
  - `MET_wpSet0_met_NOSYS` is the second most important feature, also showing a strong and stable positive effect.
  - `el_pt_NOSYS` completes the top three, with a slightly lower but still noticeable contribution.
- **Jet Kinematics**:

  - `jet_eta` and `jet_pt_NOSYS` appear in the top 10 with smaller SHAP values. `jet_pt_NOSYS` in particular shows some asymmetric contribution, with both positive and negative values.

- **Lepton and Tau Angles**:
  - Angular variables such as `el_phi`, `tau_phi`, `mu_phi`, `mu_eta`, and `jet_phi` appear with low SHAP magnitudes, indicating minor influence on the prediction.
  - Their presence in the top 10 suggests that although they contribute, their effect is small and less consistent across events.

- **Overall Structure**:
  - The top features at 800 GeV show larger SHAP values than at 250 GeV, indicating stronger feature driven separation at this mass point.

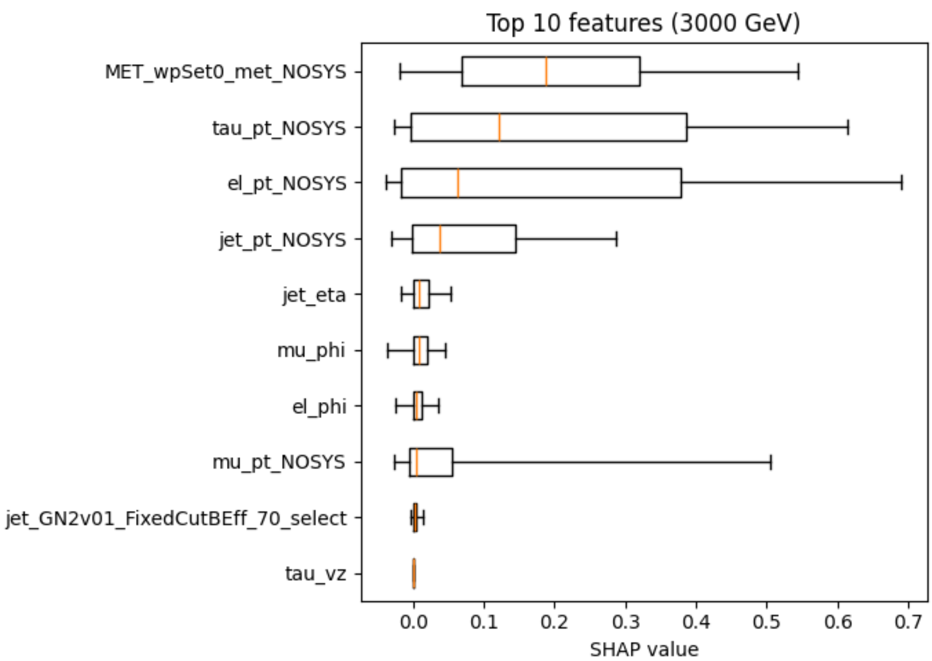

Figure 4.33: Top 10 most important features for 3000 GeV using SHAP analysis.

Key Observations for 3000 GeV:

- **Dominant Features**:

  - MET_wpSet0_met_NOSYS is the most important feature, showing a strong and stable positive SHAP value clearly distinguishing high-mass signal events from background.
  - tau_pt_NOSYS and el_pt_NOSYS also contribute significantly and consistently toward signal classification, with wide SHAP distributions and high medians.
  - jet_pt_NOSYS continues to be a relevant feature although its impact is slightly less than the lepton and MET-related features.

- **Moderate Kinematic Influence**:

  - jet_eta, mu_phi, and el_phi appear in the middle of the top 10, with relatively narrow SHAP distributions and low medians, suggesting minor but non-negligible influence.

- **Low-Impact Features**:

  - mu_pt_NOSYS and jet_GN2v01_FixedCutBEff_70_select show broader variance but lower typical SHAP contributions.
  - tau_vz appears at the bottom of the ranking with almost zero SHAP impact, shows negligible contribution to the classifier's decisions at this mass.

- **Overall Structure**:

  - Feature importance is sharply peaked for a few top variables, with the rest showing minimal influence.
  - The dominance of MET and high-p_T features is expected for heavy signal masses, where energy deposits and event level kinematics are more distinctive.

4.7.2 Permutation analysis

Figures 4.34, 4.35 and 4.36 show the top 10 most important features ranked by their permutation importance [50]. The importance is evaluated as the mean decrease in the area under the ROC curve (AUC) caused by random shuffling of each feature's values. To compute this the original AUC was first calculated using the unshuffled validation dataset.

Then for each input feature the following procedure was performed: the feature's values were randomly permuted breaking the relationship between that feature and the target labels while keeping all other features unchanged. The model then made predictions on this permuted dataset and the AUC was recalculated. The decrease in AUC relative to the original AUC indicates how much the model relies on that specific feature for classification between signal and background.

This process was repeated 30 times per feature to consider for random variability and the mean AUC decrease (in percentage points) was recorded. The resulting plots show horizontal bars corresponding to these average AUC drops. The length of each bar represents the mean importance value with percentages at the end of each bar. As in the case with SHAP values the features are ordered in decreasing importance with the most influential at the top. Features whose permutation causes a large decrease in AUC are considered to be more important as their break impacts the model's classification performance. In contrast features with short bars have little effect when permuted indicating limited predictive contribution.

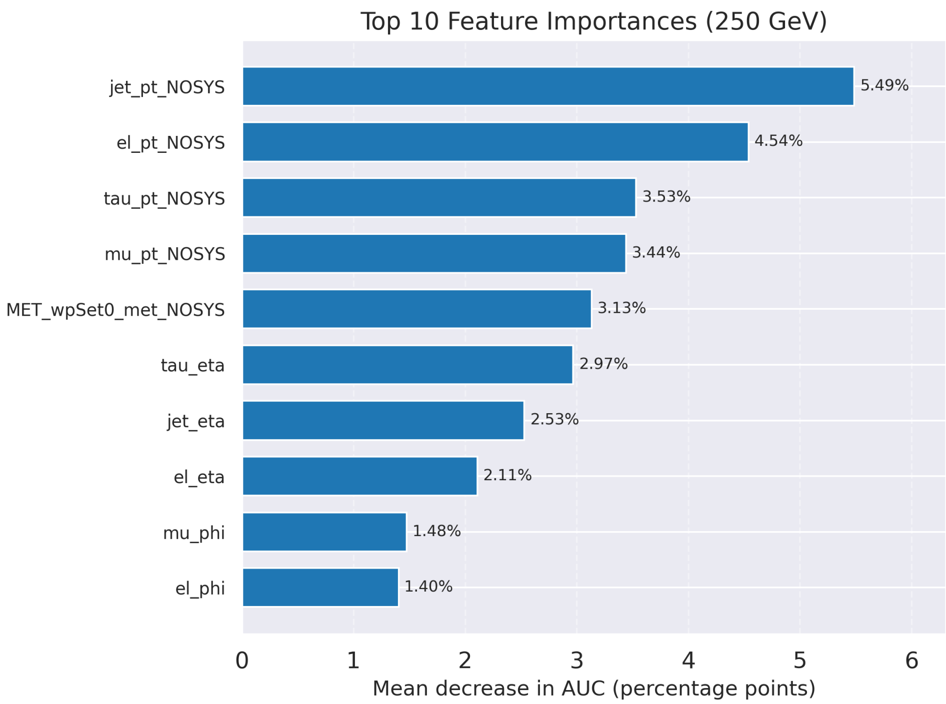

Figure 4.34: Top 10 most important permuted features for 250 GeV.

- **Dominant Features**:
  - `jet_pt_NOSYS` (5.49%) and `el_pt_NOSYS` (4.54%) are the most influential features, indicating strong dependence on high-momentum

jets and electrons.

  - tau_pt_NOSYS (3.53%) and mu_pt_NOSYS (3.44%) also play critical roles, completing the dominant p_T group.

- **Subdominant Features**:

  - MET_wpSet0_met_NOSYS (3.13%) shows moderate influence, less than expected compared to SHAP ranking.
  - Angular variables such as tau_eta, jet_eta, and el_eta contribute between 2–3%.
  - mu_phi (1.48%) and el_phi (1.40%) are the least impactful among the top 10, consistent with the idea that ϕ angles carry limited separation power at this mass scale.

Feature	**Importance (%)**
jet_pt_NOSYS	5.49
el_pt_NOSYS	4.54
tau_pt_NOSYS	3.53
mu_pt_NOSYS	3.44
MET_wpSet0_met_NOSYS	3.13

Table 4.3: Top 5 feature importances by mean decrease in AUC at 250 GeV.

Physics Meaning:

- The most influential features are transverse momenta (p_T) of jets and leptons.
- MET is moderately important, possibly due to overlaps with tau or jet kinematics.
- Angular variables (η, ϕ) play secondary roles; their distributions between signal and background are more similar at this energy scale.

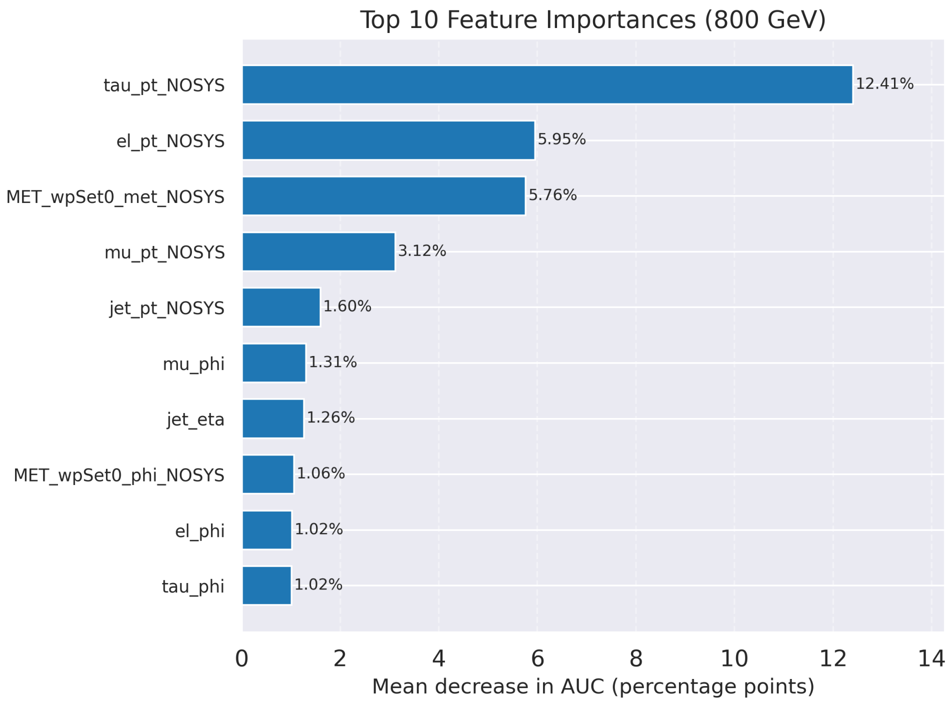

Figure 4.35: Top 10 most important permuted features for 800 GeV.

- **Dominant Features**:
  - `tau_pt_NOSYS` shows a high impact (12.41%), making it the most important feature in distinguishing signal from background.
  - `el_pt_NOSYS` (5.95%) and `MET_wpSet0_met_NOSYS` (5.76%) follow with strong contributions, indicating that both electron momentum and MET are critical at this mass point.
- **Moderate Contributors**:
  - `mu_pt_NOSYS` (3.12%) and `jet_pt_NOSYS` (1.60%) provide noticeable but smaller influence.
- **Lower-Impact Features**:
  - Angular and directional variables such as `mu_phi`, `jet_eta`, `el_phi`, `tau_phi`, and `MET_wpSet0_phi_NOSYS` contribute only around 1–1.3%, showing their relatively minor role in classification at 800 GeV.

Physics Meaning:

- High-p_T objects dominate: `tau_pt_NOSYS`, `el_pt_NOSYS`, and `mu_pt_NOSYS` are the most distinctive likely due to boosted kinematics at this mass.

Feature	Importance (%)
tau_pt_NOSYS	12.41
el_pt_NOSYS	5.95
MET_wpSet0_met_NOSYS	5.76
mu_pt_NOSYS	3.12
jet_pt_NOSYS	1.60

Table 4.4: Top 5 feature importances by mean decrease in AUC at 800 GeV.

- MET remains critical, consistent within signal.
- Directional variables again show minor influence as energy/momentum remains the dominant separating factor.

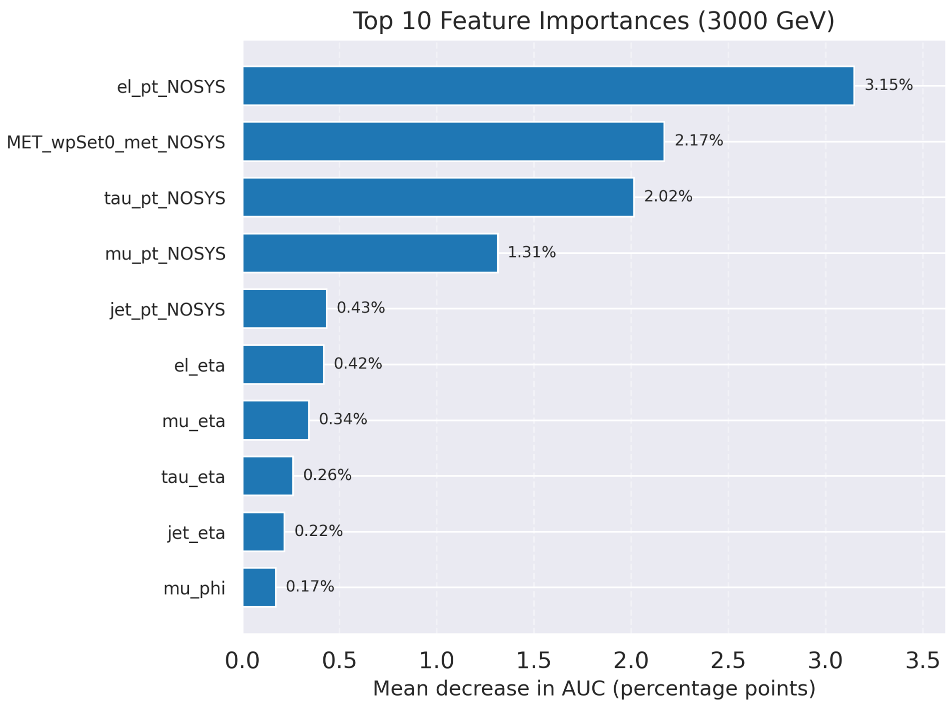

Figure 4.36: Top 10 most important permuted features for 3000 GeV.

- **Dominant Features**:
  - `el_pt_NOSYS` is the most important feature (3.15%), contributing significantly to the classifier's performance at high mass.
  - `MET_wpSet0_met_NOSYS` (2.17%) and `tau_pt_NOSYS` (2.02%) also show high influence.
- **Moderate Contributors**:

- mu_pt_NOSYS (1.31%) shows smaller, but non-negligible contribution.
- jet_pt_NOSYS (0.43%) and el_eta (0.42%) have low influence, but still appear in the top 10.

- **Low-Impact Features**:
  - Variables such as mu_eta (0.34%), tau_eta (0.26%), jet_eta (0.22%), and mu_phi (0.17%) provide minimal influence at this signal mass.

Feature	**Importance (%)**
el_pt_NOSYS	3.15
MET_wpSet0_met_NOSYS	2.17
tau_pt_NOSYS	2.02
mu_pt_NOSYS	1.31
jet_pt_NOSYS	0.43

Table 4.5: Top 5 feature importances by mean decrease in AUC at 3000 GeV.

Physics Meaning:

- At 3000 GeV the most important features remain related to transverse momenta and MET.
- Compared to lower masses, the overall importance values are lower, likely due to the model achieving high confidence from fewer features.
- Angular features (eta, phi) show minimal influence, suggesting that directional information is less influential than energy based quantities in this region.

4.7.3 Comparison of SHAP and Permutation

We saw that the top features identified by SHAP analysis did not always coincide with those obtained from permutation importance. This is due to the differences in how each method measures importance.

- **SHAP** looks at how much each feature affects the model's prediction for each individual event. Even if a feature is similar to another one SHAP can still find it important if it helps the model make better predictions in specific cases.

- **Permutation importance** checks how much the overall performance of the model drops when we randomly shuffle the values of one feature. If the model can still make good predictions using other similar features the drop in performance is small and the feature gets a lower score.

Even though a feature might be useful for the model (as SHAP shows), permutation importance might not see it as critical if the model has other ways to get the same information. That is why it is useful to look at both methods to understand the full picture.

4.7.4 Feature correlation matrices

If we want to know the linear relationships between the input features we need to compute the Pearson correlation matrices [51] for signal and background events combined. The Pearson correlation coefficient ρ_{ij} between two features x_i and x_j is defined as

$$\rho_{ij} = \frac{\mathrm{Cov}(x_i, x_j)}{\sigma_i \sigma_j}$$

where $\mathrm{Cov}(x_i, x_j)$ is the covariance between x_i and x_j, and σ_i, σ_j are their standard deviations. The value of ρ ranges from -1 to $+1$, where

- $\rho = 1$ indicates a perfect positive linear relationship,
- $\rho = -1$ indicates a perfect negative linear relationship,
- $\rho = 0$ indicates no linear correlation.

Only features with at least one correlation coefficient above 0.1 in absolute value kept to make the plots readable. The axis labels show the index of the feature in the selected feature list. These indices correspond to

Index	Feature
0	el_phi
1	tau_phi
2	tau_pt_NOSYS
3	tau_eta
4	tau_vz
5	mu_pt_NOSYS
6	mu_eta
7	jet_GN2v01_FixedCutBEff_70_select
8	mu_phi
9	el_pt_NOSYS
10	el_eta
11	jet_pt_NOSYS
12	jet_phi
13	jet_eta
14	MET_wpSet0_phi_NOSYS
15	MET_wpSet0_met_NOSYS

Table 4.6: Mapping between feature indices and their names used during the training.

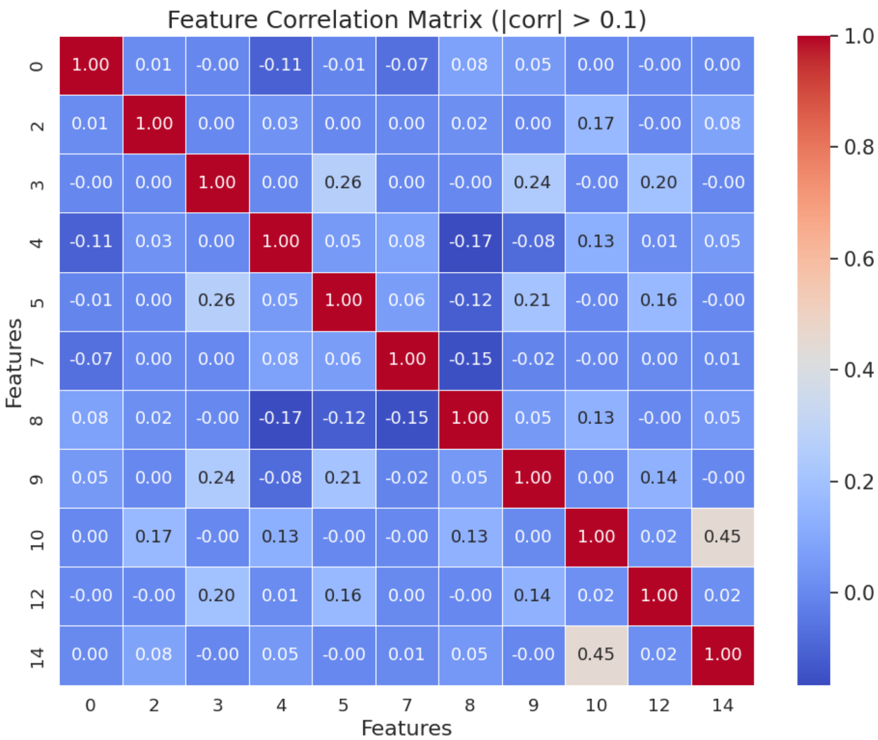

Figure 4.37: Feature correlation matrix for 250 GeV.

Figure 4.37 shows the correlation matrix computed from the 250 GeV dataset together with background. The strongest correlations are seen between feature 10 (`el_eta`) and feature 14 (`MET_wpSet0_phi_NOSYS`) with a coefficient of 0.45. Additionally, feature 3 (`tau_eta`) shows a moderate correlation with features 5 and 9. Overall, we observe weak correlations among a few kinematic features, but most coefficients remain below 0.4, indicating relative independence.

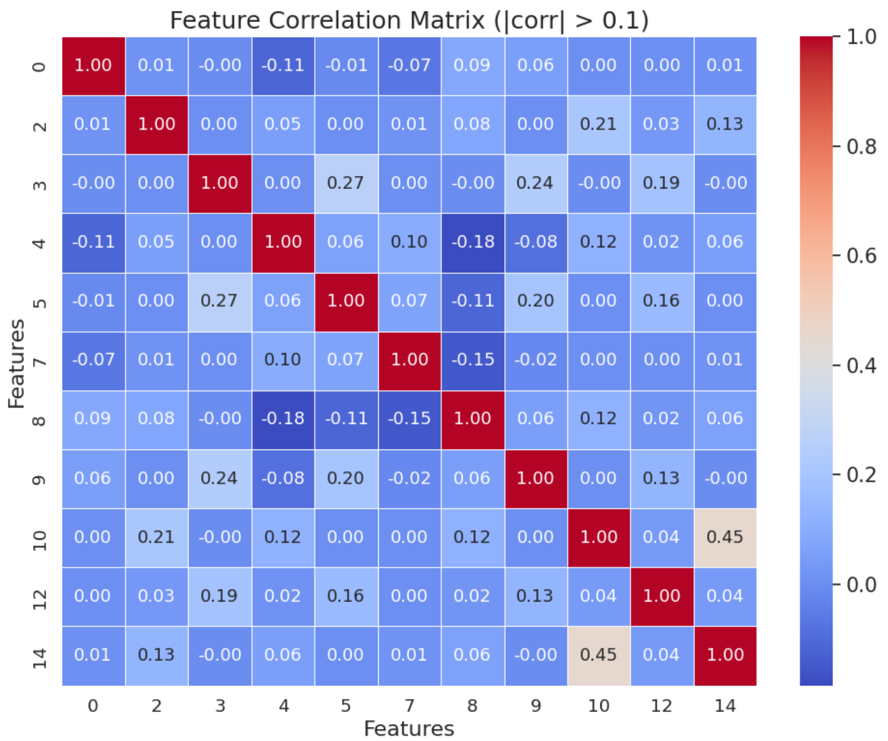

Figure 4.38: Feature correlation matrix for 800 GeV.

Figure 4.38 shows the feature correlation matrix for 800 GeV and background. The overall correlation pattern remains similar to the 250 GeV case. Some dependencies are slightly stronger such as those between features 5 (`mu_pt_NOSYS`) and 3 (`tau_eta`). The maximum observed correlation is again 0.45.

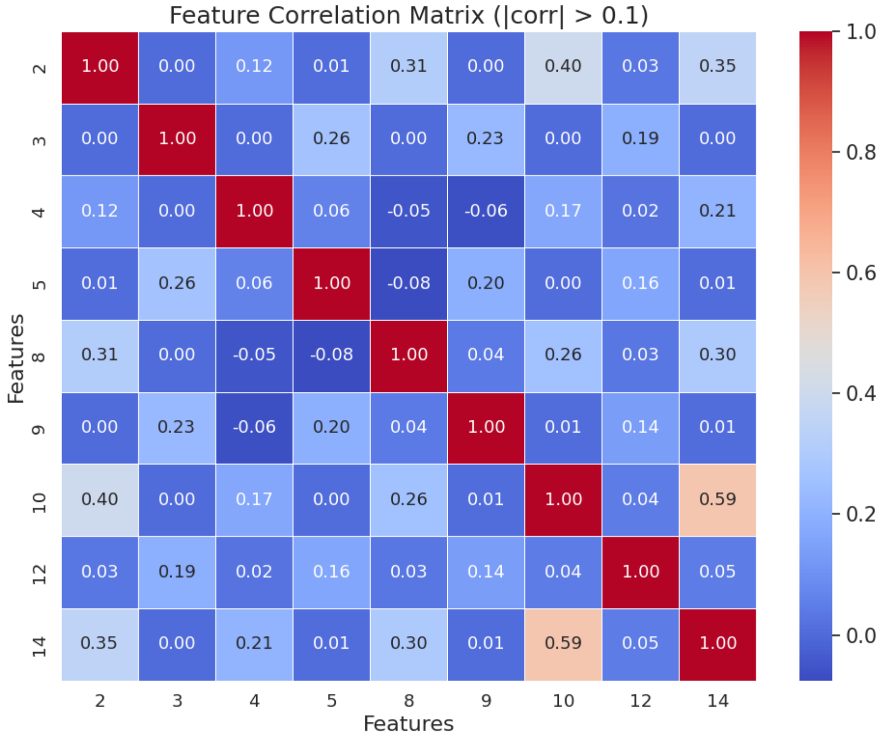

Figure 4.39: Feature correlation matrix for 3000 GeV.

Figure 4.39 illustrates the correlation matrix for 3000 GeV. At higher mass, the correlations between features become more visible. For instance, the correlation between feature 10 and 14 reaches 0.59 compared to the 250 GeV case. The consistency of this correlation across mass points suggests that `el_eta` and `MET_wpSet0_phi_NOSYS` may show a mass independent structural relationship in the events.

Chapter 5

Conclusion

The main objective of this thesis was to develop a machine learning algorithm capable of classifying signal events corresponding to the charged Higgs boson from background events and to understand its performance. At the beginning of the project several challenges were encountered to understand both machine learning techniques and the theoretical foundations of particle physics particularly the Standard Model. Overcoming these obstacles required significant effort to study relevant machine learning methodologies, data preprocessing techniques, and the physics of proton-proton collisions.

The goal of the thesis was successfully achieved. A deep neural network classifier was developed and trained on simulated Monte Carlo data prepared using the FastFrame framework. The classifier demonstrated good performance in distinguishing signal from background events across various charged Higgs boson mass hypotheses. Furthermore, advanced feature importance analyses, including SHAP values and permutation importance, were performed to understand better the decision-making process of the model and to identify the most relevant input features.

There remains room for further development. Future work could focus on improving the classification performance in the low-mass region, where the separation between signal and background is more challenging. Additionally, expanding the training dataset by including new background processes and incorporating fake events would provide a more comprehensive training scenario. The implementation of further features in the novel software will surely improve the network performance. Further optimization of the neural network architecture and hyperparameters could also be explored to enhance performance.

This thesis has not only contributed to the application of machine learning in high-energy physics but has also provided valuable experience in both fields, laying the groundwork for more advanced studies and analyses in the future.

Bibliography

[1] CERN. Views of CERN — Image Gallery, n.d. Accessed: 2025-05-22.

[2] ResearchGate. Cut-away view of the ATLAS detector, 2024. https://www.researchgate.net/figure/Cut-away-view-of-the-ATLAS-detector_fig1_313612874.

[3] Wikipedia. Standard Model (image), 2024. https://en.wikipedia.org/wiki/Standard_Model.

[4] P. Jiri. Application of Machine Learning for the Charged Higgs Boson Search Using ATLAS Data, 2022. Presented 14 Jun 2022.

[5] ATLAS Collaboration. Observation of tth production. *Physical Review Letters*, 120(23):231801, 2018. Letter.

[6] V. Mnih et al. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5:5308, 2014. Nature.

[7] ATLAS Collaboration. Measurement of ttZ and ttW production at ATLAS in 13 TeV data, using trilepton and same charge dimuon final states, 2016. CDS CERN.

[8] Machine Learning Geek. Multi-layer perceptron neural network using python, 2024. MachineLearningGeek.com.

[9] M. Rameš J. Böhm and A. Sopczak. tbH^+ analysis with MultiLeptons using machine learning with Run-2 ATLAS data. Presentation at DPG Spring Meeting, Karlsruhe, 2024. Link to abstract.

[10] CERN. Higgs boson, 2024. https://home.cern/science/physics/higgs-boson.

[11] CERN. ATLAS Experiment, 2024. https://home.cern/science/experiments/atlas.

[12] CERN. CMS Experiment, 2024. Accessed: 2025-05-22.

[13] Encyclopaedia Britannica. Particle physics, 2024. https://www.britannica.com/science/particle-physics.

[14] CERN. Supersymmetry, 2024. https://home.cern/science/physics/supersymmetry.

[15] CERN. Large Hadron Collider, 2024. https://home.cern/science/accelerators/large-hadron-collider.

[16] CERN ATLAS. Discover the ATLAS Detector, 2024. https://atlas.cern/Discover/Detector.

[17] CERN. Concept of luminosity, 2007. https://cds.cern.ch/record/941318/files/p361.pdf.

[18] CERN. The Standard Model of Particle Physics, 2024. https://home.cern/science/physics/standard-model.

[19] Encyclopaedia Britannica. Fermion, 2024. https://www.britannica.com/science/fermion.

[20] Encyclopaedia Britannica. Boson, 2024. https://www.britannica.com/science/boson.

[21] CERN. What is the Higgs Mechanism?, 2024. https://www.home.cern/science/physics/higgs-boson/what.

[22] Proton - subatomic particle, 2024. https://www.britannica.com/science/proton-subatomic-particle.

[23] What is a bottom quark?, 2024. https://people.smu.edu/ssekula/main/research/what-is/what-is-a-bottom-quark/.

[24] Top quark research at lmu, 2024. https://www.etp.physik.uni-muenchen.de/research/atlas/top/index.html.

[25] W Boson | CERN, 2024. https://home.cern/tags/w-boson.

[26] Charged leptons - britannica, 2024. https://www.britannica.com/science/subatomic-particle/Charged-leptons-electron-muon-tau.

[27] What's a neutrino? | fermilab, 2024. https://neutrinos.fnal.gov/whats-a-neutrino/.

[28] Introduction to Monte Carlo Simulation — ATLAS Open Data, 2024. ATLAS Open Data.

[29] Python programming language — official website, 2024. https://www.python.org/.

[30] Tensorflow — machine learning framework, 2024. https://www.tensorflow.org.

[31] Keras. Keras api reference. `https://keras.io/api/`, 2024. Accessed May 2025.

[32] Trexfitter - github repository, 2024. https://github.com/...

[33] FastFrame Documentation — ATLAS Top Reconstruction, 2024. FastFrames ATLAS.

[34] ROOT Data Analysis Framework — CERN, 2024. https://root.cern.

[35] Activation functions in neural networks — geeksforgeeks, 2024. GeeksforGeeks.

[36] What is a Loss Function? — IBM, 2024. IBM Think.

[37] What is an epoch in machine learning — simplilearn, 2024. Simplilearn.

[38] Batch size in neural network — geeksforgeeks, 2024. GeeksforGeeks.

[39] Andrew Dag. Introduction to neural networks: Weights, biases and activation — medium, 2024. Medium.

[40] What is a Learning Rate? — IBM, 2024. IBM Think.

[41] Neural Networks — IBM, 2024. https://www.ibm.com/think/topics/neural-networks.

[42] Binary cross-entropy / log loss for binary classification — geeksforgeeks, 2024. GeeksforGeeks.

[43] Relu activation function in deep learning — geeksforgeeks, 2024. GeeksforGeeks.

[44] Derivative of the sigmoid function — geeksforgeeks, 2024. GeeksforGeeks.

[45] GeeksforGeeks. Adam optimizer. https://www.geeksforgeeks.org/adam-optimizer/, 2021. https://www.geeksforgeeks.org/adam-optimizer/.

[46] B. Janick. Optimization of Machine Learning for the Leptoquark Search using CERN ATLAS. Data, 2023. Prague, Tech. U., presented 21 Jun 2023.

[47] Y. Vladyslav. Optimization of the ttH Selection Including Systematics Using Machine Learning with ATLAS Data, 2023. Prague, Tech. U., presented 05 Sep 2023.

[48] Google Developers. Classification: Roc and auc - machine learning crash course. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc, 2024. https://developers.google.com/.../roc-and-auc.

[49] SHAP Documentation. SHAP — SHapley Additive exPlanations. https://shap.readthedocs.io/en/latest/, 2024. https://shap.readthedocs.io/en/latest/.

[50] GeeksforGeeks. Permutation tests in machine learning. https://www.geeksforgeeks.org/permutation-tests-in-machine-learning/, 2023. https://www.geeksforgeeks.org/...

[51] W3Schools. Correlation matrix in data science. https://www.w3schools.com/datascience/ds_stat_correlation_matrix.asp, 2023. https://www.w3schools.com/...