Naturwissenschaftlich-Technische Fakultät

Universität Siegen

Universität Siegen

Bachelorarbeit

# Machine Learning Approaches for the Classification of $t\bar{t}H(c\bar{c})$ Events against $t\bar{t}H(b\bar{b})$ and $t\bar{t}$ Backgrounds

**Ole–Samuel Benedikt Witt**

Arbeit zur Erlangung des akademischen Grades
Bachelor of Science

vorgelegt der

Experimentellen Teilchen- und Astroteilchenphysik
Department Physik

**Gutachter**
Prof. Dr. Markus Cristinziani

25. September 2025

Hiermit versichere ich, dass ich die vorliegende schriftliche Arbeit selbstständig verfasst und keine anderen als die von mir angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die anderen Werken dem Wortlaut oder dem Sinne nach entnommen sind, wurden in jedem Fall unter Angabe der Quellen (einschließlich des World Wide Web und anderer elektronischer Text- und Datensammlungen) kenntlich gemacht. Dies gilt auch für die beigegebenen Zeichnungen, bildlichen Darstellungen, Skizzen und dergleichen. Mir ist bewusst, dass jedes Zuwiderhandeln als Täuschungsversuch zu gelten hat, der die Anerkennung der Arbeit ausschließt und weitere angemessene Sanktionen zur Folge haben kann.

Siegen, ..... .....

Datum Unterschrift

1. Gutachter: Prof. Dr. Markus Cristinziani

# Acknowledgements

I would like to thank my supervisor Dr. Elisabeth Schopf for her valuable guidance and consistent support throughout this thesis. She always had an open ear for my questions and took the time to help me whenever I needed it, which I truly appreciated.

I would also like to express my gratitude to Prof. Dr. Markus Cristinziani for giving me the opportunity to work on this topic, and to Dr. Wolfgang Walkowiak for his technical assistance and for the lectures that provided an important foundation for this work.

Finally, I would like to thank my parents for their continuous encouragement and support during my studies.

# Contents

CHAPTER 1

# Introduction

## 1.1 The Higgs Boson and the Standard Model

The Standard Model (SM) of particle physics is a well-established theory that describes the known fundamental particles and three of the four fundamental interactions: the electromagnetic, weak, and strong forces, while gravity is not included. The particle content is divided into fermions (quarks and leptons, which make up matter) and bosons (force carriers such as the photon, gluon, and the $W$ and $Z$ bosons). Despite its predictive power, the SM leaves several profound questions unanswered, for example the origin of dark matter or the matter–antimatter asymmetry in the universe [1].

One of the key achievements of the SM is the Higgs mechanism, which explains how elementary particles acquire mass through spontaneous electroweak symmetry breaking [2]. In 1964, Peter Higgs and others proposed the existence of a scalar particle—the Higgs boson—as a manifestation of this mechanism [3].

The discovery of the Higgs boson in 2012 by the ATLAS and CMS collaborations at CERN marked a historic milestone in modern physics [4]. Its observation confirmed the last missing piece of the Standard Model (Fig. 1.1), validating decades of theoretical and experimental work. Nevertheless, measuring the properties of the Higgs boson with high precision remains one of the primary objectives of the High Energy Physics (HEP) research program, as any deviation from SM predictions could provide evidence for new physics beyond the Standard Model.

Among the various production and decay modes of the Higgs boson, associated production with a top–antitop quark pair ($t\bar{t}H$) plays a crucial role, as it allows for a direct measurement of the top–Higgs Yukawa coupling—a fundamental parameter that controls the interaction between the Higgs boson and the heaviest known elementary particle, the top quark.

The measured mass of the Higgs boson is $125.09 \pm 0.21(\text{stat.}) \pm 0.11(\text{syst.})$ as reported in the latest combination of measurements from the ATLAS and CMS collaborations [5]. For a Higgs boson with this mass it is expected that the decay into a pair of bottom quarks ($H \to b\bar{b}$) is the most common. In contrast, the decay into charm quarks ($H \to c\bar{c}$) is much rarer and experimentally challenging due to large background contributions. Improving the sensitivity to this decay is therefore essential to study the coupling of the Higgs boson to second-generation quarks, a largely unexplored sector of the SM [6].

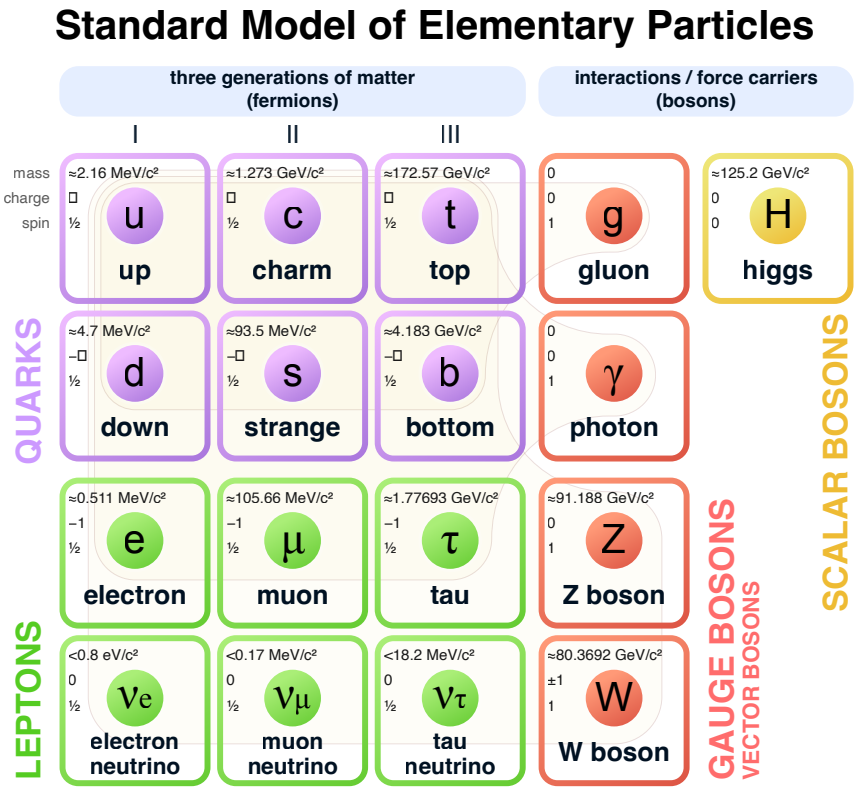



Figure 1.1: The fundamental particles whose interactions are described in the Standard Model of particle physics [1]

## 1.2 The Large Hadron Collider and the ATLAS Detector

To explore rare processes, such as $H \to c\bar{c}$ decays, experiments require a high-energy, high-luminosity particle collider combined with a sophisticated detector system. The Large Hadron Collider (LHC) at CERN is currently the world's most powerful particle accelerator, designed to deliver proton–proton collisions at a center-of-mass energy of up to 14 TeV. It consists of a 27 km circular tunnel located 45–170 meters underground, reusing the former LEP tunnel. Two counter-rotating proton beams are accelerated using superconducting magnets and brought to collision at four interaction points, which host the major experiments ATLAS, CMS, ALICE, and LHCb [7].

ATLAS is one of two general-purpose detectors optimized for a wide range of physics goals, from precision measurements of the Standard Model to searches for new phenomena, such as supersymmetry. As illustrated in Fig. 1.2, ATLAS employs a layered detection system consisting of an inner tracking detector inside a solenoidal magnetic field, electromagnetic and hadronic calorimeters located outside the solenoid, and a muon spectrometer embedded in large toroidal magnetic fields. This configuration enables precise measurement of particle trajectories, energies, and momenta across a broad kinematic range [8].

The experimental environment at the LHC presents significant challenges: at design luminosity,

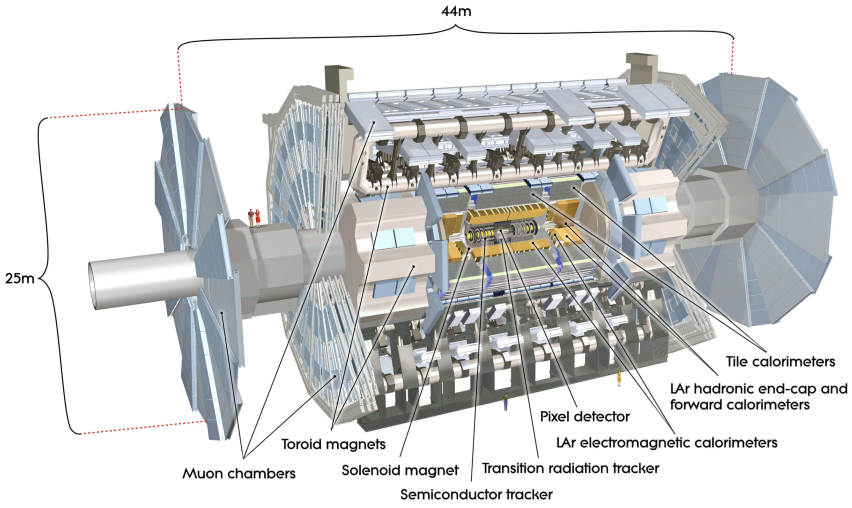



Figure 1.2: Cut-away view of the ATLAS detector [8, p. 4]

there are up to 40 million proton–proton collisions per second, resulting in extremely high event rates. In high-energy hadron collisions, it is primarily the quarks and gluons inside the protons that interact, leading to a large variety of final states. A dominant contribution comes from strong-interaction (Quantum Chromodynamics, QCD) processes such as multijet production (e.g. $gg \to q\bar{q}$), which form a major background for many analyses [8]. Since quarks and gluons cannot be observed as free particles, they appear in the detector as collimated sprays of hadrons, known as jets. To extract rare signatures with jets, such as $H \to c\bar{c}$, and to distinguish them from the large amount of expected background processes with similar final states, advanced techniques in event reconstruction, flavor tagging (identifying the quark flavor that initiated a jet), and machine–learning–based classification are essential.

## 1.3 Thesis Scope and Objectives

This thesis focuses on machine learning approaches for classifying $t\bar{t}H$ events with Higgs decays into charm quarks ($H \to c\bar{c}$) against dominant backgrounds from $H \to b\bar{b}$ and $t\bar{t}$ production. To suppress the immense amount of expected QCD multi-jet background, this work is only considering processes with exactly one electron or muon in the final state.

The aim of this thesis is to evaluate the performance of different classification models, to investigate the relative importance of input features, and to explore how well these models can separate signal from background in a setup similar to current ATLAS analyses.

By applying modern machine learning techniques, this thesis contributes to efforts to measure the Higgs couplings more precisely, especially for second-generation quarks, where experimental knowledge is still limited.

CHAPTER 2

# Physics Processes and Input Data

## 2.1 Signal Process

The signal process studied in this thesis is the associated production of a Higgs boson with a top–antitop quark pair, where the Higgs boson decays into a charm–anticharm quark pair:

$$pp \to t\bar{t}H, \quad H \to c\bar{c} \; . \tag{2.1}$$

This process is referred to as $t\bar{t}H(c\bar{c})$. It is particularly challenging to identify due to the small branching ratio of 3 % for the Higgs–boson decay to charm quarks [9] and the difficulty of identifying charm jets, referred to as $c$–tagging.

Fig. 2.1(a) shows a representative leading-order sketch illustrating this signal process. Both top quarks decay into a bottom quark and a $W$ boson, which has a branching ratio of 99.7 % [10]. According to the event preselection, which will be described in more detail later, each event contains exactly one electron or muon. A total of six jets is expected in the final state.

## 2.2 Background Processes

The $t\bar{t}H(c\bar{c})$ signal process is challenged by two major background processes. The dominant one is

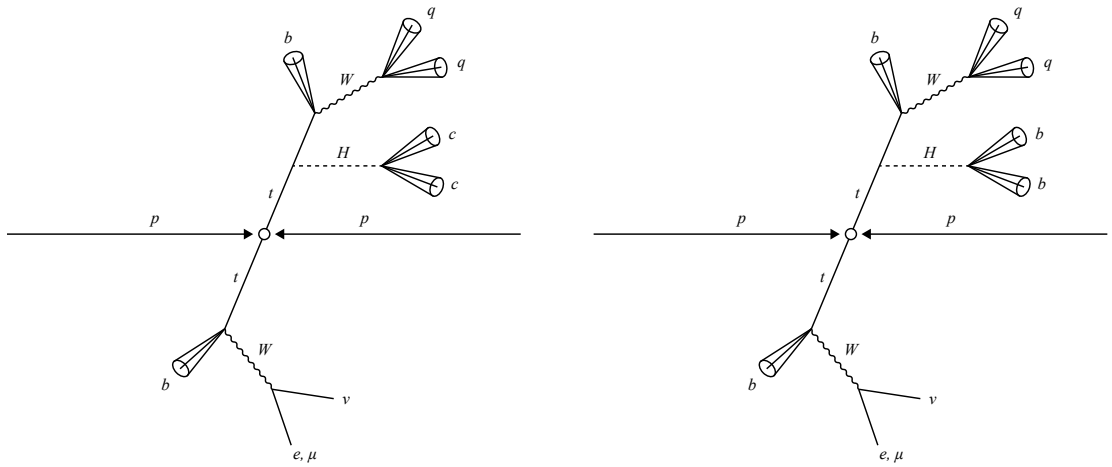
$$pp \to t\bar{t} \; , \tag{2.2}$$

commonly referred to as $t\bar{t}$ production. It also contains two top quarks but no Higgs boson is produced. This typically results in four jets in the final state. A representative leading-order sketch is shown in Fig. 2.2(a), while Fig. 2.2(b) illustrates a next-to-leading-order (NLO) process in which a gluon is emitted from one of the top quarks and subsequently splits into a pair of charm quarks. This NLO process leads to six jets in the final state, making it indistinguishable from the signal process in terms of jet multiplicity and jet flavor.

The second relevant background process considered in this analysis is

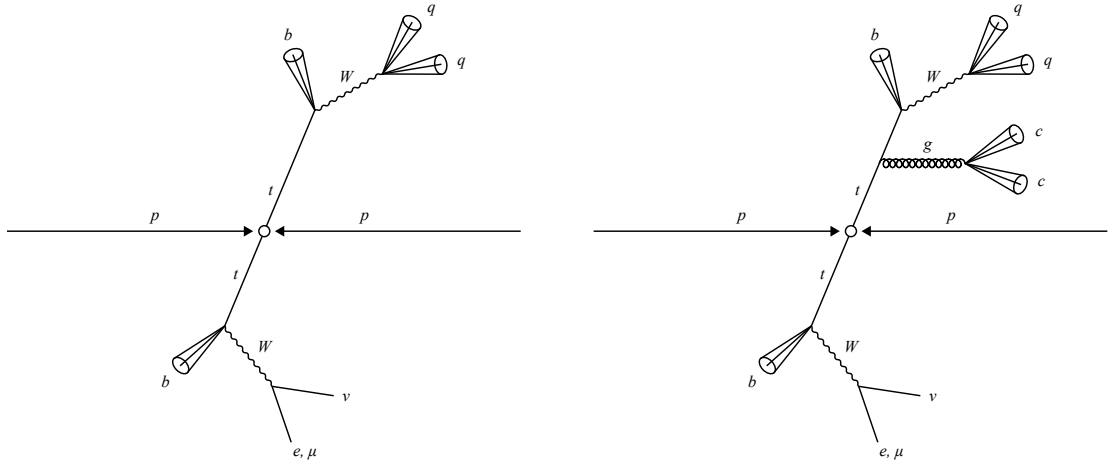
$$pp \to t\bar{t}H, \quad H \to b\bar{b} \; , \tag{2.3}$$

referred to as $t\bar{t}H(b\bar{b})$. In contrast to the signal, the Higgs boson decays into a pair of bottom quarks rather than charm quarks, but the total number of expected jets remains six. A leading-order sketch is shown in Fig. 2.1(b).




(a) $t\bar{t}H(c\bar{c})$ (b) $t\bar{t}H(b\bar{b})$

Figure 2.1: Leading-order sketches of $t\bar{t}H(c\bar{c})$ and $t\bar{t}H(b\bar{b})$




(a) $t\bar{t}$ (b) $t\bar{t}$ (NLO)

Figure 2.2: Leading-order and next-to-leading-order sketches of $t\bar{t}$

Other minor background contributions—such as $t\bar{t}Z$, $t\bar{t}W$, $tW$, and $W$ + jets—were neglected in this study. Based on independent simulations passing the same preselection, these processes are expected to contribute on the order of $2.35 \cdot 10^5$ events, compared to $2.87 \cdot 10^6$ events from the dominant $t\bar{t}$ background [11]. Their overall impact is therefore subdominant, and, in contrast to the challenging separation between $t\bar{t}H(c\bar{c})$ and $t\bar{t}H(b\bar{b})$, whose only difference is the flavor of the jets produced, these backgrounds are generally easier to distinguish from the signal process.

## 2.3 Monte Carlo Data Set and Preselection

The data set used in this analysis consists of Monte Carlo-simulated events, labeled as either $t\bar{t}H(c\bar{c})$, $t\bar{t}H(b\bar{b})$, or $t\bar{t}$. These events are weighted to correspond to the full Run–2 integrated luminosity of 140 fb$^{-1}$ at a center-of-mass energy of 13 TeV [12].

The cross section for $t\bar{t}H$ production is

$$\sigma(t\bar{t}H) = 0.51 \text{ pb} , \tag{2.4}$$

as reported in [13], compared to

$$\sigma(t\bar{t}) = 843 \text{ pb} \tag{2.5}$$

for inclusive $t\bar{t}$ production [14]. Taking into account the branching ratios

$$\text{BR}(H \to c\bar{c}) \approx 3 \text{ \%} , \tag{2.6}$$
$$\text{BR}(H \to b\bar{b}) \approx 58 \text{ \%} , \tag{2.7}$$

for the Higgs–boson decays into charm and bottom quarks, respectively [9], the following weighted number of events is obtained after applying the preselection defined below:

$$N_{t\bar{t}H(c\bar{c})} = 124 , \tag{2.8}$$
$$N_{t\bar{t}H(b\bar{b})} = 4\,424 , \tag{2.9}$$
$$N_{t\bar{t}} = 2\,869\,099 , \tag{2.10}$$

$$N_{\text{sig}} = N_{t\bar{t}H(c\bar{c})} = 124 , \tag{2.11}$$
$$N_{\text{bkg}} = N_{t\bar{t}H(b\bar{b})} + N_{t\bar{t}} = 2\,873\,523 . \tag{2.12}$$

These numbers are affected by a statistical uncertainty of less than 1 % [11]. The resulting signal-to-background ratio is extremely small:

$$\frac{N_{\text{sig}}}{N_{\text{bkg}}} = \frac{124}{2\,873\,523} \approx 4.3 \cdot 10^{-5} . \tag{2.13}$$

All events in the data set are required to satisfy the following preselection criteria:

- Exactly one electron or muon must be present in the event.
- Each event must contain at least five reconstructed jets.
- At least two jets must fulfill a $b$–tagging requirement (at least *loose b–tagging*), as defined in Section 2.5.

## 2.4 Feature Definitions

Each event in the data set contains up to ten reconstructed jets. These jets are ordered by their transverse momentum $p_{\text{T}}$ in descending order; i.e., jet 1 has the highest $p_{\text{T}}$ and jet 10 the lowest.

For each jet, the following features are provided:

- **kinematic features:** energy $E$, transverse momentum $p_{\mathrm{T}}$, pseudorapidity $\eta$, and azimuthal angle $\phi$ [1]
- **flavor–tagging probabilities:** probability for the jet to originate from a $b$–quark ($p_b$), $c$–quark ($p_c$), light-flavor quark ($p_u$), or $\tau$–lepton ($p_\tau$)
- **true flavor:** the ground–truth jet flavor from simulation.

In addition, the following event–level features are included:

- number of reconstructed jets $n_{\mathrm{jets}}$
- number of $b$-tagged jets ($n_b$) and light jets ($n_u$) based on flavor tagging
- missing transverse energy (MET) $E_T^{\mathrm{miss}}$ and azimuthal angle $\phi^{\mathrm{miss}}$ of the MET vector [2]
- true process label (either $t\bar{t}H(c\bar{c})$, $t\bar{t}H(b\bar{b})$, or $t\bar{t}$).

## 2.5 Flavor–Tagging Information

The flavor–tagging probabilities, $p_b$, $p_c$, $p_u$, and $p_\tau$, are provided by the transformer-based flavor–tagging algorithm GN2v01 [17]. Two discriminants, $D_b$ and $D_c$, are defined as

$$D_b = \log\left(\frac{p_b}{f_c \cdot p_c + f_\tau \cdot p_\tau + (1 - f_c - f_\tau) \cdot p_u}\right), \tag{2.14}$$

$$D_c = \log\left(\frac{p_c}{f_b \cdot p_b + f_\tau \cdot p_\tau + (1 - f_b - f_\tau) \cdot p_u}\right), \tag{2.15}$$

where

$$f_c = 0.2, \quad f_\tau = 0.01 \tag{2.16}$$

are parameters used for $b$–tagging and

$$f_b = 0.3, \quad f_\tau = 0.05 \tag{2.17}$$

for $c$–tagging [17].

The discriminants are employed to assign each jet to one of seven categories, defined in Table 2.1 and illustrated in Fig. 2.3.

[1] "ATLAS uses a right-handed coordinate system with its origin at the nominal interaction point (IP) in the centre of the detector and the $z$-axis along the beam pipe. The $x$-axis points from the IP to the centre of the LHC ring, and the $y$-axis points upwards. Polar coordinates ($r$, $\phi$) are used in the transverse plane, $\phi$ being the azimuthal angle around the $z$-axis. The pseudorapidity is defined in terms of the polar angle $\theta$ as $\eta = -\ln\tan\theta/2$ and is equal to the rapidity $y = \frac{1}{2}\ln\frac{E+p_z}{E-p_z}$ in the relativistic limit. Angular distance is measured in units of $\Delta R \equiv \sqrt{(\Delta y)^2 + (\Delta\phi)^2}$." [15]

[2] The MET vector is defined as the negative vector sum of all reconstructed transverse momenta in the event, and it provides sensitivity to non-interacting particles such as neutrinos [16].

| Condition | Category | Number |
|---|---|---|
| $D_b > 2.669$ | tight $b$-tagged | 6 |
| $1.892 < D_b < 2.669$ | medium $b$-tagged | 5 |
| $0.884 < D_b < 1.892$ | loose $b$-tagged | 4 |
| $D_b < 0.884$ and $D_c > 3.958$ | tight $c$-tagged | 3 |
| $D_b < 0.884$ and $2.090 < D_c < 3.958$ | medium $c$-tagged | 2 |
| $D_b < 0.884$ and $0.503 < D_c < 2.090$ | loose $c$-tagged | 1 |
| $D_b < 0.884$ and $D_c < 0.503$ | untagged | 0 |

Table 2.1: GN2v01 flavor–tagging categories [17]

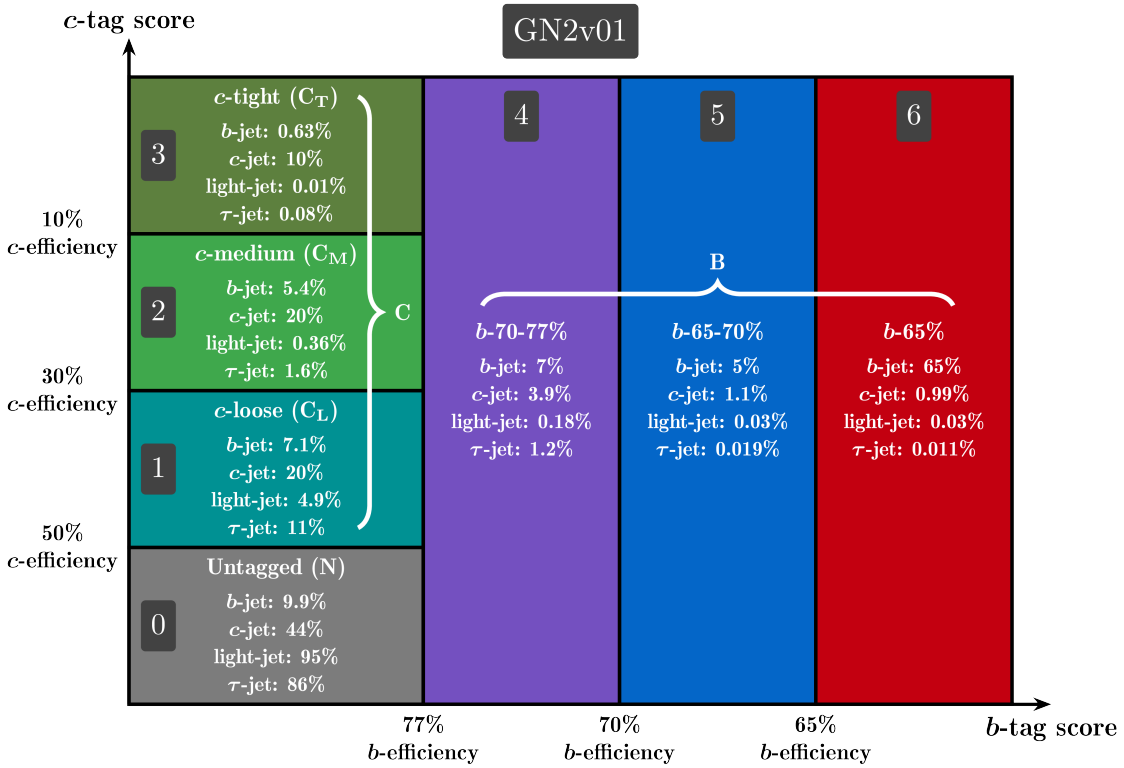



Figure 2.3: Binning scheme in the GN2v01 $c$–tag score ($D_c$) vs. $b$–tag score ($D_b$) plane (efficiency for jets of a given true flavor to be categorized into the given category) [18]

CHAPTER 3

# Initial Manual Approaches for Process Separation

## 3.1 Exploratory Feature Analysis

Before applying machine–learning techniques, an exploratory analysis of selected features is performed to identify variables with potential discriminative power between the signal and background processes. Visualizing these features provides insight into their distributions and possible separability.

As discussed in Chapter 2, the processes differ in the number of jets of specific flavors. Therefore, the multiplicities of $b$–jets ($n_b$) and light jets ($n_u$) are considered as initial separation criteria. Their normalized distributions are shown in Fig. 3.1.

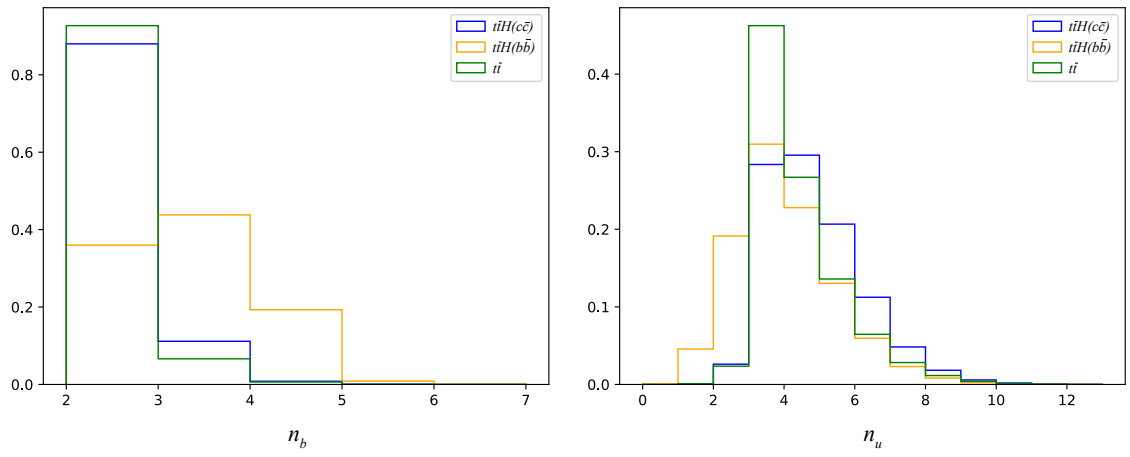



Figure 3.1: Distributions of $b$–jet and light jet multiplicities for signal and background processes, normalized to unity

The distribution of $n_u$ is similar across the three processes, indicating limited discriminative power.

For both $t\bar{t}H(c\bar{c})$ and $t\bar{t}$, most events contain exactly two $b$–jets, consistent with Chapter 2, since each top quark is expected to decay into a bottom quark and a $W$ boson. The absence of events with

fewer than two $b$–jets is a direct consequence of the preselection requirement that every event must contain at least two $b$-tagged jets.

In contrast, the $t\bar{t}H(b\bar{b})$ distribution peaks at $n_b = 3$, with a substantial number of events containing four $b$–jets. The additional $b$–jets originate from the Higgs–boson decay. These observations suggest that $n_b$ can serve as a simple first criterion to separate $t\bar{t}H(b\bar{b})$ events from $t\bar{t}H(c\bar{c})$ and $t\bar{t}$ events.

While many kinematic features exhibit similar distributions across the three processes, the transverse momenta $p_{\mathrm{T},1}$ and $p_{\mathrm{T},2}$ of the leading and subleading jets—i.e., the two jets with the highest $p_{\mathrm{T}}$—show noticeable differences. Their normalized distributions are presented in Fig. 3.2.

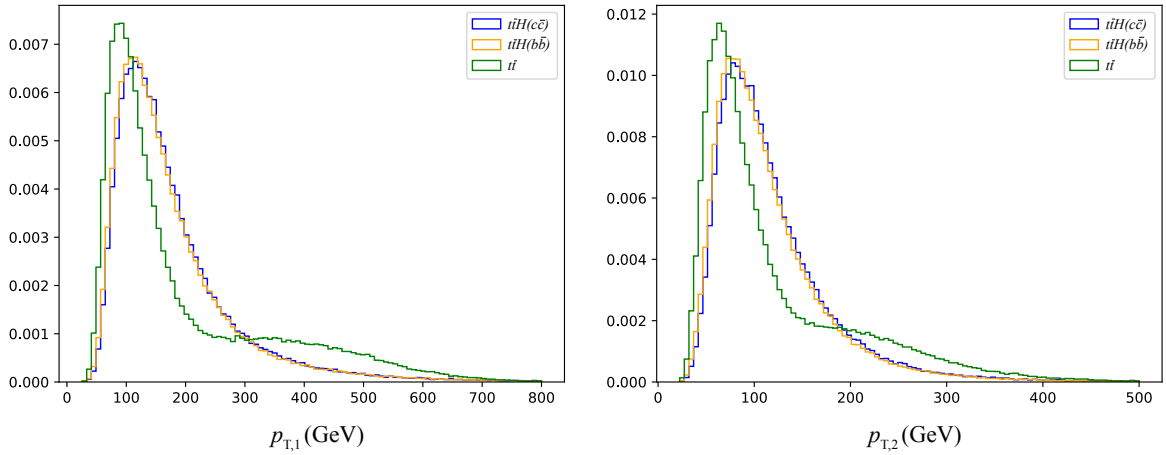



Figure 3.2: Distributions of $p_{\mathrm{T},1}$ and $p_{\mathrm{T},2}$ for signal and background processes, normalized to unity

For both $t\bar{t}H(c\bar{c})$ and $t\bar{t}H(b\bar{b})$, the $p_{\mathrm{T},1}$ distribution peaks at approximately 110 GeV. In contrast, the $t\bar{t}$ distribution peaks at a lower value of about 90 GeV and exhibits an additional shoulder at higher transverse momenta in the range of 300–500 GeV, for which no clear explanation could be found. The $p_{\mathrm{T},2}$ distributions follow similar shapes but are shifted to lower momenta, as the subleading jet always has a lower $p_{\mathrm{T}}$ than the leading jet.

This high-$p_{\mathrm{T}}$ shoulder could act as a discriminating feature between $t\bar{t}$ events and those from $t\bar{t}H(c\bar{c})$ or $t\bar{t}H(b\bar{b})$ processes.

## 3.2 Jet–Pair Combinations for Higgs Reconstruction

The invariant mass of two jets $i$ and $j$ is defined as

$$m_{ij} = \sqrt{(p_i + p_j)^{\mu}(p_i + p_j)_{\mu}}\,, \tag{3.1}$$

where $p_i$ and $p_j$ denote the four–momenta of the jets. In terms of their energies $E_i$, $E_j$ and three–momenta $\vec{p}_i$, $\vec{p}_j$, this expression becomes

$$m_{ij} = \sqrt{(E_i + E_j)^2 - \left|\vec{p}_i + \vec{p}_j\right|^2}\,. \tag{3.2}$$

Fig. 3.3 shows the dijet invariant mass of the two leading true $c$–jets in $t\bar{t}H(c\bar{c})$, and, for comparison, the two leading true $b$–jets in $t\bar{t}H(b\bar{b})$.

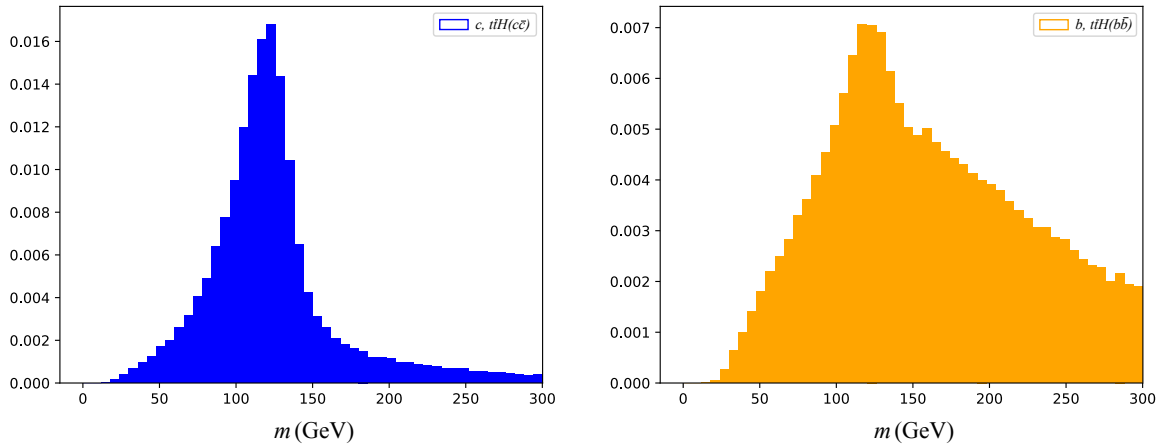



Figure 3.3: Distributions of the dijet invariant mass of the two leading true $c$–jets in $t\bar{t}H(c\bar{c})$ and the two leading true $b$–jets in $t\bar{t}H(b\bar{b})$, normalized to unity

The $c$–jet–mass distribution exhibits a pronounced peak around the Higgs–boson mass of $\sim 125$ GeV, since almost all $c$–jets in the signal process originate from Higgs decays.

In contrast, the $b$–jet–mass distribution shows the same Higgs peak overlaid with an additional high-mass component. This tail is likely caused by $b$–jets from top–quark decays ($t \to Wb$), which tend to have higher transverse momenta.

To study which jet pairs $(i, j)$ are consistent with a Higgs decay, all possible combinations

$$i, j = 1, \ldots, 10, \quad i < j \ , \tag{3.3}$$

are considered in Fig. 3.4. For each combination, the number of events with

$$110 \text{ GeV} < m_{ij} < 140 \text{ GeV} \ , \tag{3.4}$$

is counted. This mass window corresponds to the Higgs–boson region.

In general, the decay of a high-mass particle such as the Higgs boson tends to produce jets with relatively high transverse momenta. Consequently, jet pairs falling into the Higgs–mass window are often dominated by high-$p_{\mathrm{T}}$ jets (low indices $i$). The overall patterns are similar for $t\bar{t}H(c\bar{c})$ and $t\bar{t}H(b\bar{b})$, but small differences compared to $t\bar{t}$ are visible—for example, for the combinations (1, 6), (2, 3), and (2, 4). Interestingly, the (1, 2) combination is not the most promising for reconstructing the Higgs boson in either $t\bar{t}H(c\bar{c})$ or $t\bar{t}H(b\bar{b})$ events, since the leading jets often originate from the top–quark decays rather than the Higgs decay.

These observations indicate that the differences between the processes are very subtle, and no single jet combination provides both a high efficiency for reconstructing the Higgs boson and a strong separation between $t\bar{t}H$ and $t\bar{t}$ events. Selected dijet masses from this study will be included as additional features in the machine–learning models introduced later.

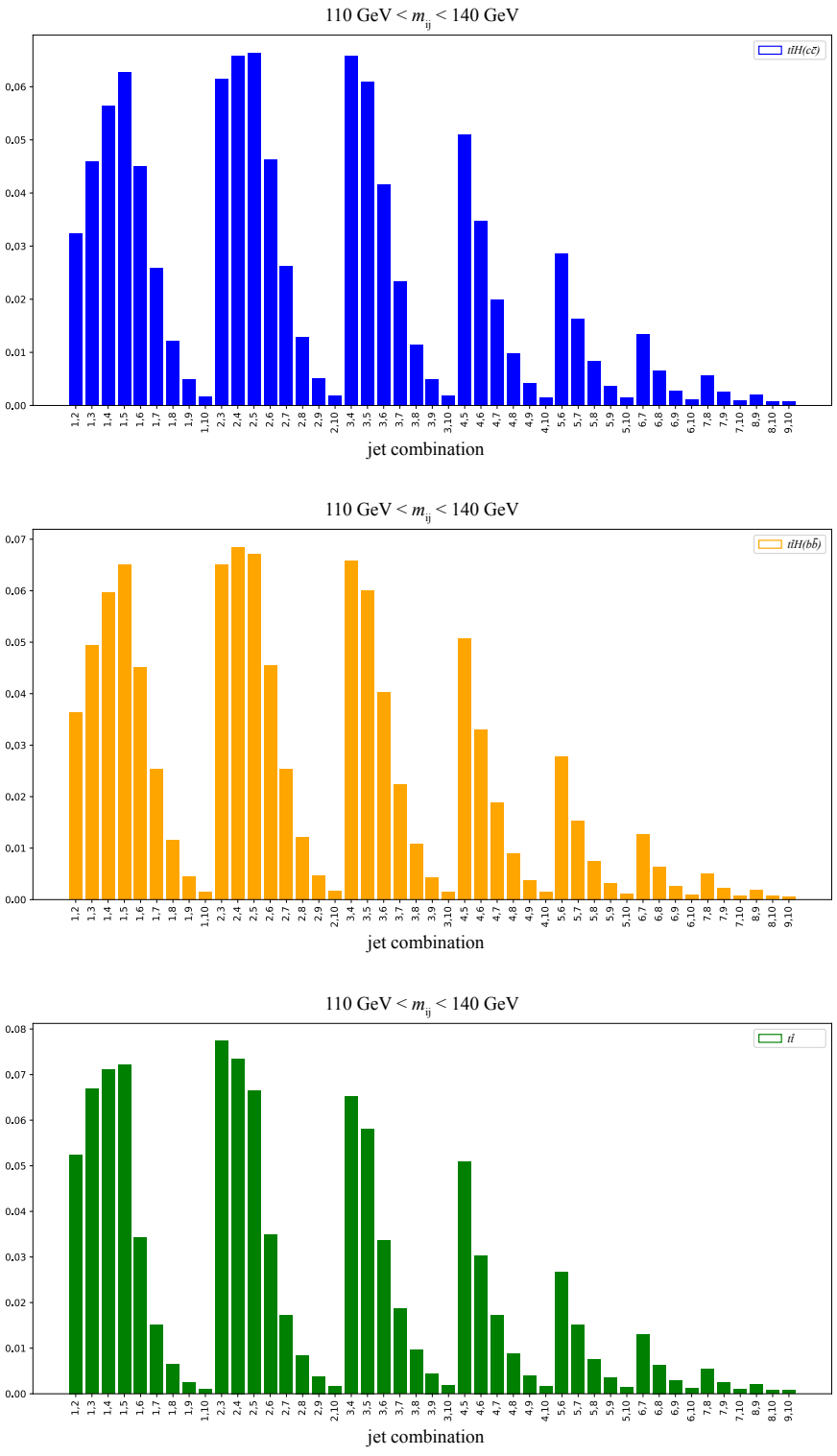



Figure 3.4: Number of events with dijet masses in the Higgs–mass window for all jet combinations, normalized to unity

## 3.3 Performance Metrics

Evaluating and comparing the performance of classification algorithms requires clearly defined performance metrics. For binary classification, the following quantities are used:

- **true positives** ($TP$): signal events correctly classified as signal
- **false negatives** ($FN$): signal events incorrectly classified as background
- **false positives** ($FP$): background events incorrectly classified as signal
- **true negatives** ($TN$): background events correctly classified as background.

From these quantities, the following metrics can be computed:

- **signal efficiency:**

$$\epsilon_{\text{sig}} = \frac{TP}{TP + FN} \tag{3.5}$$

- **signal purity:**

$$P_{\text{sig}} = \frac{TP}{TP + FP} \tag{3.6}$$

- **binary accuracy:**

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \,. \tag{3.7}$$

For multiclass classification with $K$ classes, $k = 1, \ldots, K$, the **multiclass accuracy** is defined as

$$\text{Acc}_{\text{mult}} = \frac{\sum_{k=1}^{K} N(\text{predicted} = k \wedge \text{true} = k)}{N_{\text{total}}} \,. \tag{3.8}$$

To also calculate a binary accuracy for multiclass classification, both $t\bar{t}$ and $t\bar{t}H(b\bar{b})$ are treated as background in this case.

The **significance** $Z$ of a binary classification is calculated from the number of signal events $S$ correctly classified as signal and the number of background events $B$ misclassified as signal. The approximated formula reads

$$Z \approx \frac{S}{\sqrt{B}} = \frac{TP}{\sqrt{FP}} \,, \tag{3.9}$$

yielding a number in units of standard deviations [19].

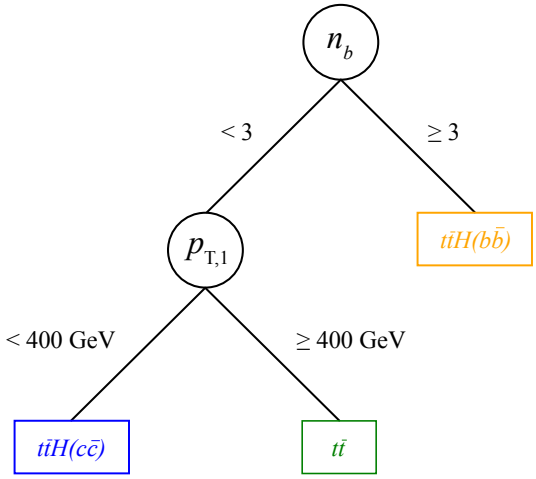



Figure 3.5: Baseline manual decision tree for process classification

## 3.4 Baseline Classification without Machine Learning

Based on the feature observations discussed in this chapter, a simple manual classification approach is implemented as a reference for later comparison with machine–learning methods. The structure of the resulting two-step decision tree is illustrated in Fig. 3.5.

In the first step, $t\bar{t}H(b\bar{b})$ events are separated from $t\bar{t}H(c\bar{c})$ and $t\bar{t}$ events using the $b$–jet multiplicity $n_b$, as motivated by Fig. 3.1. Since most $t\bar{t}H(c\bar{c})$ and $t\bar{t}$ events contain exactly two $b$–jets, while $t\bar{t}H(b\bar{b})$ events often contain three or more, the threshold

$$n_{b,\text{cut}} = 3 \tag{3.10}$$

is applied.

In the second step, $t\bar{t}H(c\bar{c})$ events are distinguished from $t\bar{t}$ events by exploiting the differences in the leading-jet transverse-momentum distribution shown in Fig. 3.2. Leading jets in $t\bar{t}$ events tend to have higher $p_\text{T}$ values than those in $t\bar{t}H(c\bar{c})$ events, motivating the cut

$$p_{\text{T},1,\text{cut}} = 400 \text{ GeV} \, . \tag{3.11}$$

This baseline approach achieves an accuracy of 0.42 for signal–background classification. The full set of performance metrics is summarized in Table 3.1.

Given the small signal-to-background ratio of $4.3 \cdot 10^{-5}$ obtained in Section 2.3, the resulting significance of 0.08 is very low. This value will serve as a baseline for evaluating the improvement achieved by machine–learning–based classifiers in later sections.

| **Significance** | **0.08** |
|---|---|
| Binary accuracy | 0.42 |
| Multiclass accuracy | 0.12 |
| Signal efficiency | 0.84 |
| Signal purity | $6.3 \cdot 10^{-5}$ |

Table 3.1: Performance metrics of manual decision tree

CHAPTER 4

# Random Forest Classification and Feature Importance Analysis

## 4.1 Random Forest with Full Feature Set

The random forest (RF) algorithm is an ensemble learning method that combines multiple decision trees to improve classification performance and reduce overfitting compared to individual decision trees. Each tree is trained on a bootstrap sample of the data, and feature selection at each split introduces additional randomness, resulting in a diverse set of trees. The final prediction is obtained by majority voting over all trees in the forest [20].

One key advantage of random forests is their ability to provide an estimate of feature importance, which reflects how strongly each input variable contributes to the classification. In practice, this is computed from how often and how effectively a feature is used to split the data within the trees. The values are normalized so that all importances add up to unity, meaning that a value of 0.035 indicates that the feature accounts for about 3.5 % of the overall importance across all features. This makes RF particularly suitable for analyzing which features are most discriminative for the processes considered in this study.

First, a random forest is trained to classify the three studied processes using the complete set of features introduced in Section 2.4 (excluding true flavor). The RF classifier is implemented using the *scikit–learn* library [21] with 100 decision trees which is the default setting in recent versions. This value is chosen as it provides a good trade-off between model stability and computational efficiency. Increasing the number of trees typically improves performance slightly but at the cost of longer training time, while reducing the number may lead to less robust results.

The performance metrics of this process classification are summarized in Table 4.1. Both binary and multiclass classification yield a significance of 0.11, which represents an improvement over the baseline value of 0.08 from Section 3.4.

Fig. 4.1 illustrates the feature importance for the full feature set. The distribution is dominated by flavor–tagging–based features, with the top six being $p_{c,1}$, $n_b$, $p_{c,2}$, $p_{c,3}$, $p_{c,4}$ and $p_{c,5}$. This aligns with the heuristic criteria used in the manual decision tree (Section 3.4), where the $b$–jet multiplicity $n_b$ served as the primary discriminant.

Furthermore, feature importance generally decreases with increasing jet index. This is expected, as many events do not contain all ten jets, resulting in features of higher-index jets being set to zero and thus contributing little to the classification.

| | **Binary** | **Multiclass** |
|---|---|---|
| **Significance** | **0.11** | **0.11** |
| Binary accuracy | 0.95 | 0.88 |
| Multiclass accuracy | | 0.82 |
| Signal efficiency | 0.32 | 0.52 |
| Signal purity | $2.9 \cdot 10^{-4}$ | $1.8 \cdot 10^{-4}$ |

Table 4.1: Performance metrics of random forest with full feature set

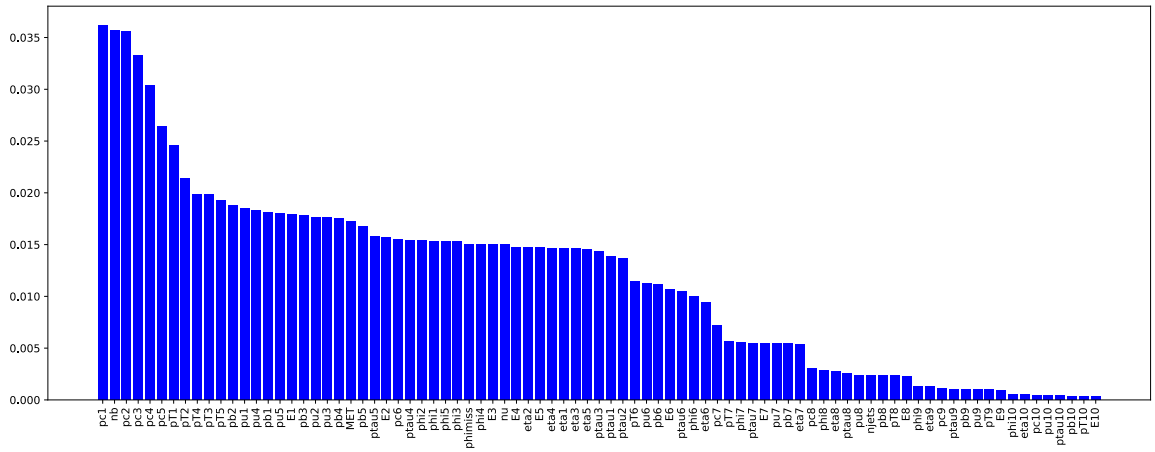



Figure 4.1: Relative feature importance of random forest with full feature set (binary classification)

## 4.2 Random Forest without Flavor–Tagging Features

To assess the discriminative power of purely kinematic features, all flavor–tagging–based variables are excluded from training. Specifically, the following features are removed:

$$n_b,\ n_u,\ \ p_{b,i},\ p_{c,i},\ p_{u,i},\ p_{\tau,i},\ \ i = 1, \ldots, 10\ . \tag{4.1}$$

The feature–importance distribution shown in Fig. 4.2 highlights the transverse momenta of the leading and subleading jets, $p_{\mathrm{T},1}$ and $p_{\mathrm{T},2}$, among the three most important features. This observation is consistent with the shape differences in their distributions discussed in Fig. 3.2 and supports the heuristic use of $p_{\mathrm{T},1}$ as a discriminant in the second step of the manual decision tree (Section 3.4).

However, random forests are generally limited in their ability to exploit certain physics features, such as invariant masses of jet combinations, which are often crucial for distinguishing signal from background in high-energy physics. Therefore, the subsequent analysis focuses on optimizing the performance of neural network classifiers, which are better suited to capture complex, non-linear correlations among features.

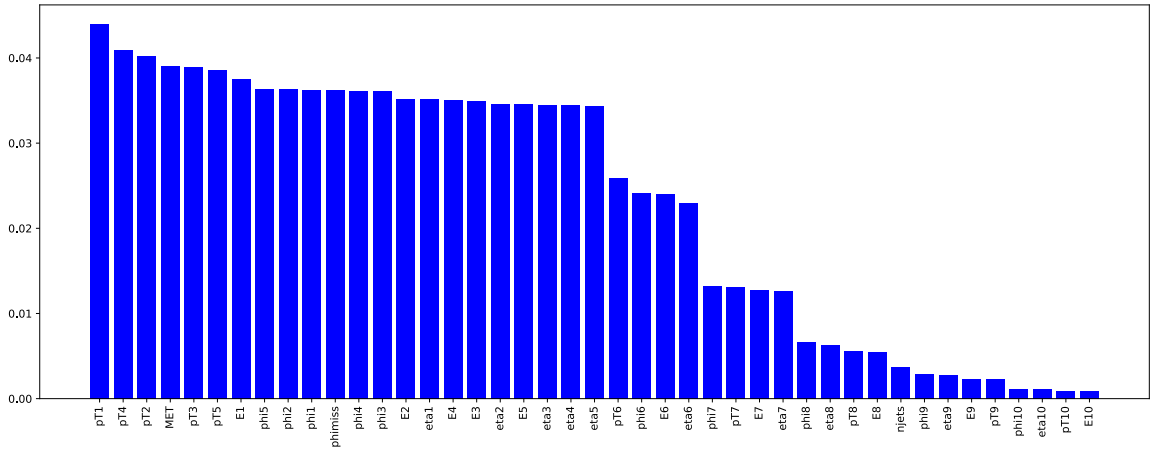



Figure 4.2: Relative feature importance of random forest without flavor–tagging features (binary classification)

CHAPTER 5

# Neural Network Classification

## 5.1 Neural Network Architecture and Training Performance

Artificial neural networks (NNs) are a class of machine–learning algorithms inspired by the structure and functioning of biological neural systems. They consist of interconnected layers of nodes (neurons), where each connection is associated with a weight that is iteratively adjusted during training to minimize a given loss function. Through this process, NNs are capable of modeling highly non-linear relationships between input features and target labels, making them particularly suitable for complex classification tasks [22].

In this work, a feed-forward neural network is implemented using the *scikit–learn* library [21]. The optimized architecture consists of three hidden layers with 120, 60, and 30 neurons, respectively. Between each pair of hidden layers, dropout layers are inserted to reduce overfitting by randomly deactivating a fraction of neurons during training. This architecture was chosen empirically to provide a good balance between model complexity, stability, and computational efficiency. The network is trained with the Adam optimizer and uses binary cross–entropy or categorical cross–entropy as loss function, depending on whether binary or multiclass classification is performed.

Since the feature analysis in Section 4.1 indicated that features corresponding to jets with high indices contribute negligibly to classification, all features of jets 7–10 are excluded. This reduces dimensionality, accelerates training, and forces the network to focus on more discriminative features.

Fig. 5.1 shows the evolution of training and test score as a function of epochs for multiclass classification. Both scores increase rapidly during the first five epochs and then flatten, indicating convergence. The training is performed for a fixed number of 30 epochs, as no further improvement is observed after this point.

Interestingly, the test score consistently exceeds the training score by around 0.5 %. This effect can be explained by dropout regularization: dropout is applied only during training, which introduces noise and makes the optimization problem more challenging. In contrast, the evaluation on the test set uses the full network without dropout, resulting in more stable predictions and slightly higher performance.

Note that the training and test scores in Fig. 5.1 refer to unweighted batch accuracies during training and should not be confused with the event-weighted accuracy used for physics performance evaluation. The latter, together with other performance metrics, is summarized in Table 5.1.

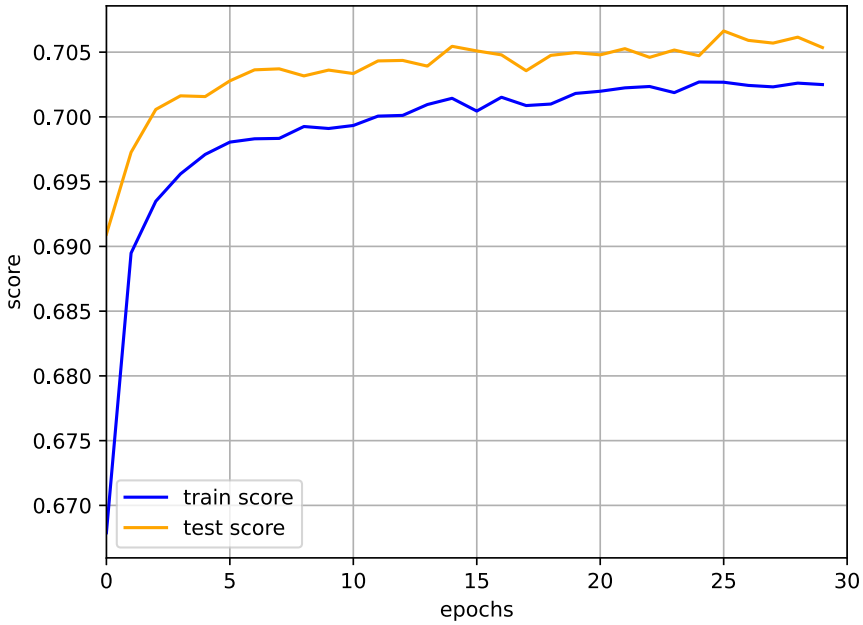



Figure 5.1: Learning curve of neural network (multiclass classification)

| | **Binary** | **Multiclass** |
|---|---|---|
| **Significance** | **0.12** | **0.12** |
| Binary accuracy | 0.93 | 0.88 |
| Multiclass accuracy | | 0.83 |
| Signal efficiency | 0.42 | 0.55 |
| Signal purity | $2.7 \cdot 10^{-4}$ | $2.0 \cdot 10^{-4}$ |

Table 5.1: Performance metrics of neural network

Both binary and multiclass classification achieve a significance of 0.12, which corresponds to a slight improvement compared to the random forest result of 0.11.

## 5.2 Feature Augmentation for Performance Optimization

A common strategy to enhance the performance of neural networks is feature augmentation, where additional input variables are derived from existing features if they are expected to improve the discriminative power between signal and background.

As a first approach, dijet–mass information is considered (Section 3.2). For each event, the invariant mass $m_{ij}$ of selected jet pairs $(i, j)$ is computed and added as an additional feature. The jet pairs chosen for this purpose are (1,2), (1,3), (1,4), (1,5), (1,6), (2,3), (2,4), (2,5), (3,4), (3,5), and (4,5), as they contain most of the combinations relevant for Higgs decays. The selected pairs are highlighted in yellow in Fig. 5.2, which shows the distribution for the $t\bar{t}H(c\bar{c})$ process as an example.

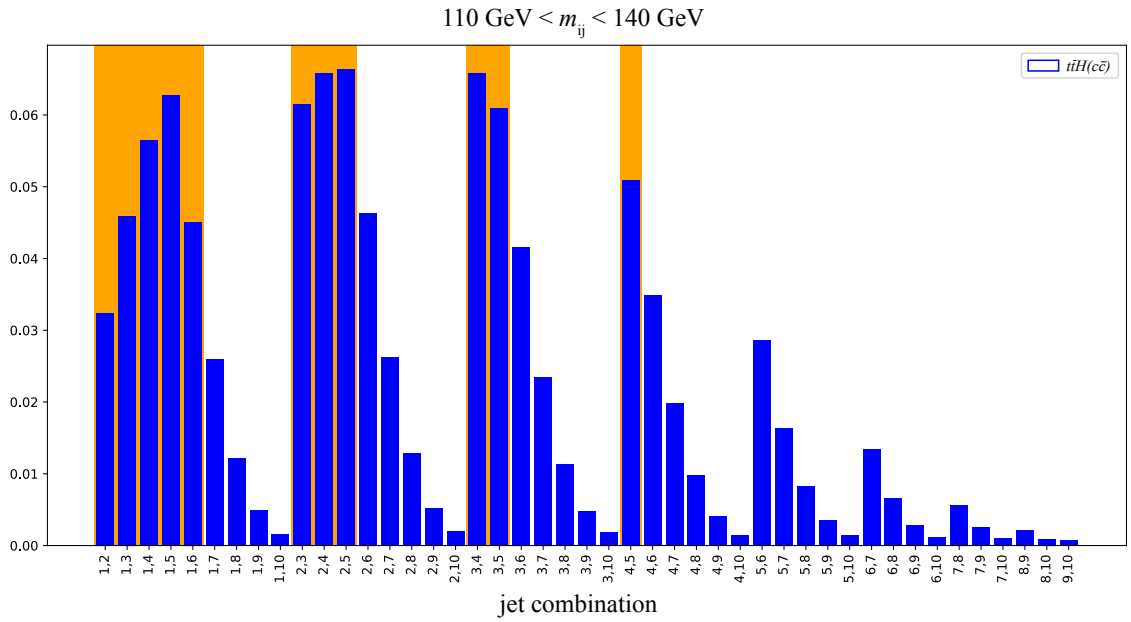



Figure 5.2: Number of events with dijet masses in the Higgs–mass window (selected combinations, normalized to unity)

As a second modification, the logarithm of the leading-jet transverse momentum,

$$\log(p_{\mathrm{T},1}) \,, \tag{5.1}$$

is introduced to further exploit the differences in the transverse-momentum distributions observed in Section 3.1. The high-$p_{\mathrm{T},1}$ shoulder present in the $t\bar{t}$ distribution indicates that events with unusually large $p_{\mathrm{T},1}$ are more likely to originate from the background process.

By taking the logarithm, the dynamic range of $p_{\mathrm{T},1}$ is compressed, allowing the network to more effectively distinguish between low- and high-momentum events. This transformation can enhance the

sensitivity of the classifier to subtle differences in the kinematic distributions, potentially improving its overall performance. Fig. 5.3 compares the distributions of $p_{\mathrm{T},1}$ and $\log(p_{\mathrm{T},1})$ for signal and background processes.

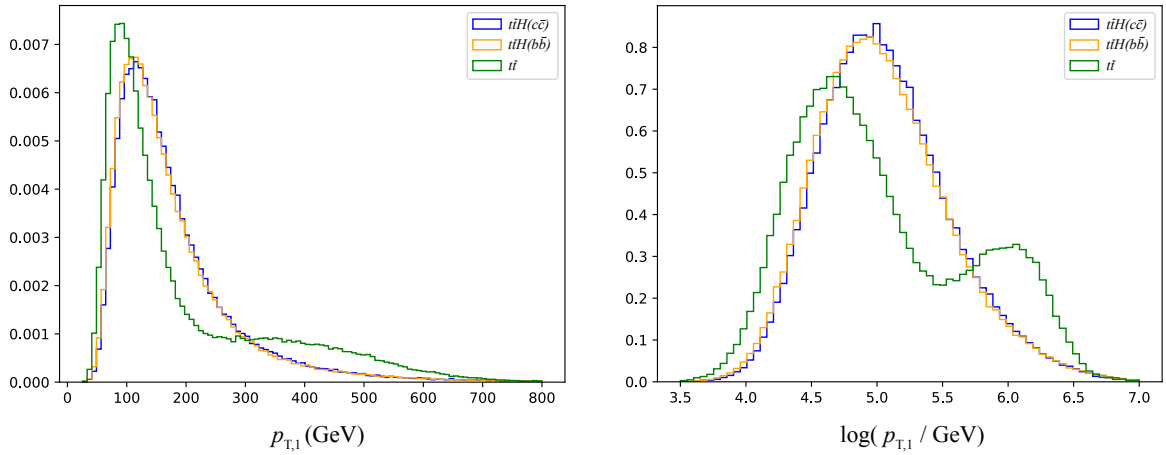



Figure 5.3: Distributions of $p_{\mathrm{T},1}$ and $\log(p_{\mathrm{T},1})$ for signal and background processes, normalized to unity

The impact of both feature augmentations on classification performance is summarized in Table 5.2. Compared to the initial configuration in Table 5.1, no significant improvement is observed in any metric; in particular, the significance remains at 0.12. This indicates that the neural network is already capable of capturing the relevant correlations between the original kinematic variables and the target classes, even without these explicit feature transformations.

| | **Dijet masses** | **$\log(p_{\mathrm{T},1})$** |
|---|---|---|
| **Significance** | **0.12** | **0.12** |
| Binary accuracy | 0.87 | 0.89 |
| Multiclass accuracy | 0.81 | 0.83 |
| Signal efficiency | 0.57 | 0.53 |
| Signal purity | $2.0 \cdot 10^{-4}$ | $2.1 \cdot 10^{-4}$ |

Table 5.2: Performance metrics of neural network with extended feature set (multiclass classification)

## 5.3 Classification Results

To visualize the separation capability of the unmodified neural network discussed in Section 5.1, the output probabilities for the classes $t\bar{t}H(c\bar{c})$ and $t\bar{t}H(b\bar{b})$ are plotted against each other in Fig. 5.4. For $t\bar{t}H(c\bar{c})$ events, a high density is observed in the lower-right corner of the plot, corresponding to high $t\bar{t}H(c\bar{c})$ probability and low $t\bar{t}H(b\bar{b})$ probability. A smaller cluster of $t\bar{t}H(c\bar{c})$ events is also visible

in the lower-left corner, where both class probabilities are low, reflecting cases where the network assigns low confidence to both Higgs hypotheses. Conversely, most $t\bar{t}H(b\bar{b})$ events are located in the upper-left region, where the predicted $t\bar{t}H(b\bar{b})$ probability is high. Events from $t\bar{t}$ are mostly concentrated in the lower-left corner, where both class probabilities are low. These patterns indicate that the network discriminates $t\bar{t}H(b\bar{b})$ from $t\bar{t}$ more effectively than $t\bar{t}H(c\bar{c})$ from $t\bar{t}$.

Given this observation, the trained network could also be used to design a $t\bar{t}H(b\bar{b})$ measurement, however, this capability will not be explored further within the scope of this thesis, which focuses on $t\bar{t}H(c\bar{c})$.

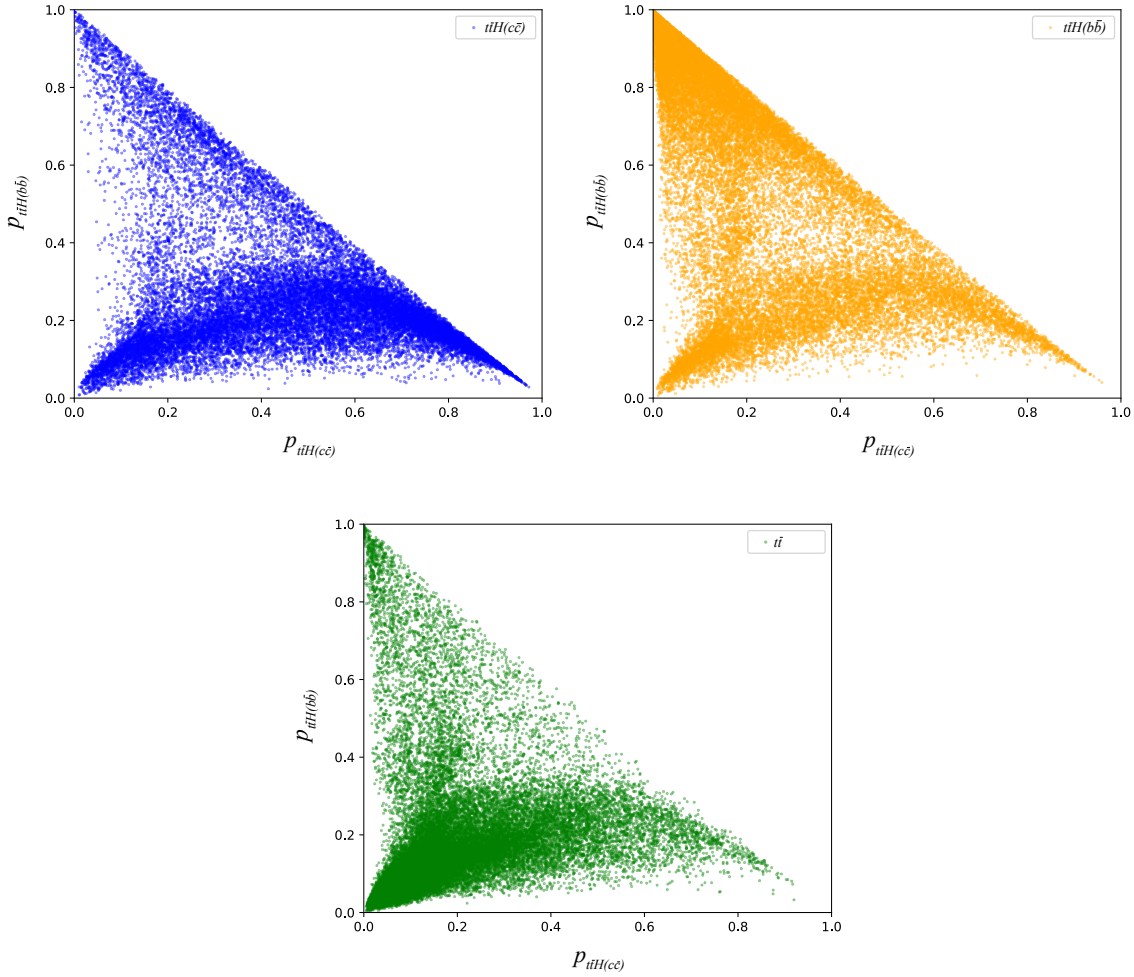



Figure 5.4: Neural network output probabilities for $t\bar{t}H(c\bar{c})$ and $t\bar{t}H(b\bar{b})$ for the three processes

The distribution of $p_{t\bar{t}H(c\bar{c})}$, the neural network output probability for the signal class, is shown in Fig. 5.5 for all three processes. As expected, events from the $t\bar{t}H(c\bar{c})$ process exhibit an approximately uniform distribution across the entire range, reflecting the model's uncertainty in distinguishing some signal-like features from background-like ones. In contrast, $t\bar{t}H(b\bar{b})$ and $t\bar{t}$ events are strongly concentrated at low $p_{t\bar{t}H(c\bar{c})}$ values, indicating that the network effectively assigns them a low

probability of being $t\bar{t}H(c\bar{c})$. The vertical line marks the optimal cut at $p_{t\bar{t}H(c\bar{c})} = 0.50$, which is used to calculate the significance of 0.12.

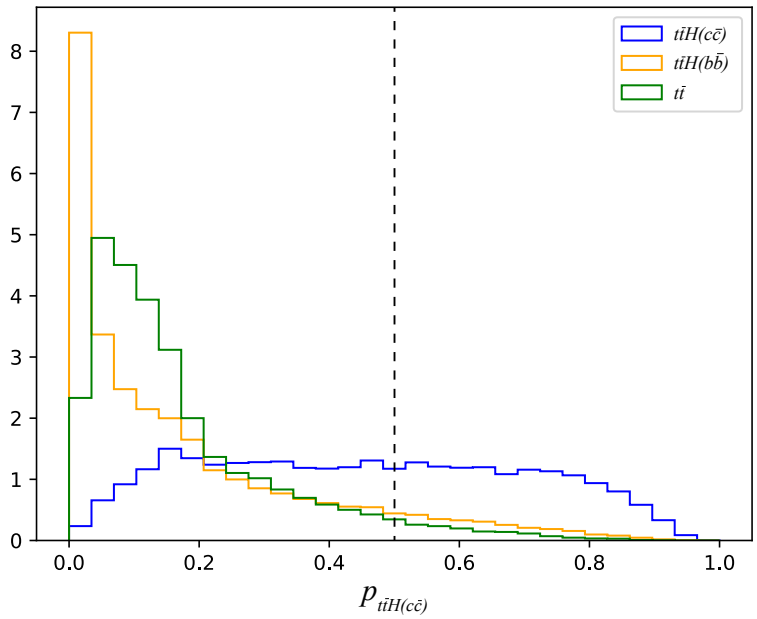



Figure 5.5: Distribution of $p_{t\bar{t}H(c\bar{c})}$ for the three processes, normalized to unity

For a more detailed analysis, the significance can also be evaluated on a bin-by-bin basis using 20 bins. For each bin $i$, the significance contribution is defined as

$$Z_i = \frac{S_i}{\sqrt{B_i}} , \tag{5.2}$$

where $S_i$ and $B_i$ denote the number of signal and background events in bin $i$, respectively.

In many similar ATLAS analyses (e.g., $VH(c\bar{c})$ or $t\bar{t}H(b\bar{b})$), the classifier's output score is often used directly as a discriminant instead of applying a hard cut. This allows the significance to be extracted continuously across the score range, giving a more complete picture of the network's performance [11].

The combined bin-wise significance is then:

$$Z_{\text{bin}} = \sqrt{\sum_{i=1}^{20} Z_i^2} = \sqrt{\sum_{i=1}^{20} \left(\frac{S_i}{\sqrt{B_i}}\right)^2} = 0.16 . \tag{5.3}$$

This bin-based approach improves the overall significance compared to a single cut and yields a result comparable to the binary classification performed in a $W(\ell\nu)H(c\bar{c})$ analysis using a boosted decision tree (neglecting systematic uncertainties) [11], which reported:

$$Z_{WH(c\bar{c})} = 0.16 . \tag{5.4}$$

CHAPTER 6

# Higgs Candidate Classification

## 6.1 Labeling Jets as Higgs Candidates

To identify $c$–jets in $t\bar{t}H(c\bar{c})$ events that most likely originate from the Higgs–boson decay, a jet–level classification task is introduced. Since the dataset does not provide explicit labels indicating whether a jet originates from a Higgs boson, a dedicated labeling procedure is required. This procedure is implemented in three steps:

1. **Event selection:** Only events labeled as $t\bar{t}H(c\bar{c})$ by the simulation are considered.
2. **Jet preselection:** Within these events, jets 1–6 are examined, and only those labeled as true $c$–jets are retained.
3. **Invariant mass criterion:** For events with at least two $c$–jets passing the previous step, the invariant mass of the two leading $c$–jets is calculated. If this dijet mass satisfies

$$110 \text{ GeV} < m_{ij} < 140 \text{ GeV} , \tag{6.1}$$

   the two leading jets are labeled as Higgs candidates.

The full labeling logic is visualized in Fig. 6.1. The resulting labels are used for training a jet–level neural network to distinguish jets originating from Higgs–boson decays from all other jets. This Higgs candidate classification network employs the same architecture as the process–level network introduced in Section 5.1: three hidden layers with 120, 60, and 30 neurons, combined with dropout layers for regularization. Training is performed using the Adam optimizer with binary cross–entropy loss.

## 6.2 Classifying Jets as Higgs Candidates

The classification scheme for identifying Higgs candidate jets mirrors the logic of the labeling procedure and consists of three main steps, illustrated in Fig. 6.2:

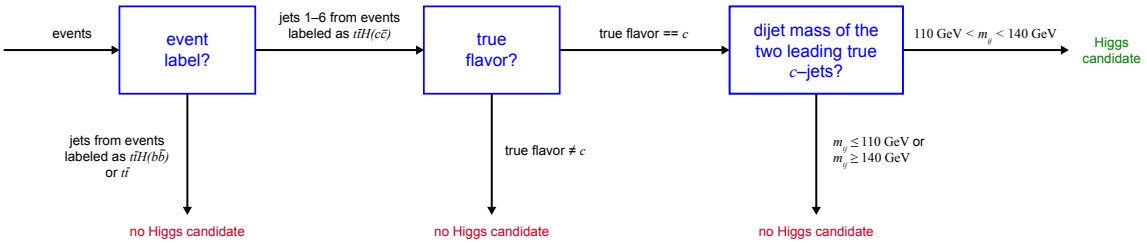



Figure 6.1: Jet–labeling scheme for Higgs candidate identification

1. **Event–level prediction:** The event is first classified using the neural network described in Section 5.1. Only events predicted as $t\bar{t}H(c\bar{c})$ are considered for further processing.

2. **Jet filtering:** For these events, jets 1–6 are required to pass at least a *loose c–tagging* threshold according to the GN2v01 flavor–tagging algorithm introduced in Section 2.5.

3. **Candidate prediction:** The remaining jets are evaluated by the Higgs candidate classification network, which assigns each jet a probability of originating from the Higgs–boson decay.

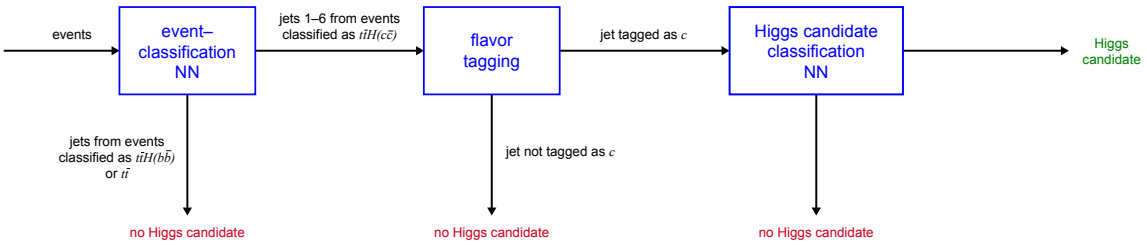



Figure 6.2: Jet–classification scheme for Higgs candidate identification

The performance of the complete Higgs candidate classification procedure is summarized in Table 6.1. In this context, signal refers to jets labeled as Higgs candidates, while background corresponds to all other jets. The resulting significance of 0.06 is substantially lower than in event–level classification, which is expected due to the severe class imbalance: $t\bar{t}H(c\bar{c})$ events are rare, and each event contains at most two Higgs candidate jets, further reducing the signal-to-background ratio.

It is important to note that this evaluation is performed per jet, not per event. Each jet is treated independently, and the significance quantifies how well the network can distinguish true Higgs jets from other jets in the event.

However, the definition of the training labels introduces an additional limitation: there is no exact ground truth for the origin of reconstructed jets. Instead, jets are labeled using a simplified matching procedure, which can only approximate their true origin. This unavoidably creates ambiguities, especially in events with many overlapping jets, and likely explains why standard metrics such as the accuracy reach very high values like 0.98 even though the actual separation power in terms of significance remains low.

| **Significance** | **0.06** |
|---|---|
| Binary accuracy | 0.98 |
| Signal efficiency | 0.40 |
| Signal purity | $1.1 \cdot 10^{-4}$ |

Table 6.1: Performance metrics of Higgs candidate classification

Overall, this analysis demonstrates that while the neural network can identify Higgs candidate jets with some success, the inherently low signal fraction at the jet level limits the achievable significance. This motivates further exploration of combining jet–level predictions with event–level classification in future studies.

CHAPTER 7

# Summary and Outlook

This thesis explored machine–learning approaches for distinguishing $t\bar{t}H(c\bar{c})$ events from the dominant $t\bar{t}H(b\bar{b})$ and $t\bar{t}$ backgrounds. A manually constructed decision tree, based on feature–level observations, served as the initial reference and achieved a significance of **0.08**. This approach was not optimized—for instance, it does not make use of $c$–tagging information or finely tuned jet–selection criteria—but rather provides a simple baseline that exploits only the most obvious discriminating features.

Building upon this, a random forest classifier improved the significance to **0.11**, demonstrating the benefit of machine–learning methods for this classification task. Further enhancement was achieved through a feed–forward neural network, which reached a value of **0.12** using a single threshold cut and **0.16** with a bin-by-bin approach.

This performance is comparable to that reported in a $VH$ analysis employing a boosted decision tree, which achieved a significance of **0.16** (neglecting systematic uncertainties) [11]. While $VH$ has traditionally been regarded as the benchmark channel for probing $H \to c\bar{c}$ decays, the results obtained in this work indicate that the $t\bar{t}H(c\bar{c})$ channel can achieve a similar level of sensitivity. This suggests that the $t\bar{t}H$ production mode represents a promising and complementary channel for future searches targeting the Higgs coupling to charm quarks.

However, these findings are subject to certain limitations, as systematic uncertainties have not been taken into account yet and minor background contributions—such as $t\bar{t}Z$, $t\bar{t}W$, $tW$, and $W+$jets—were neglected (Section 2.2).

The multi-class neural network also demonstrates good performance in identifying $t\bar{t}H(b\bar{b})$ events, which indicates that the same approach could potentially be used to extract the $t\bar{t}H(b\bar{b})$ signal. However, it should be noted that the raw flavor–tagging network output probabilities used in this study are not calibrated and, in ATLAS analyses, using uncalibrated probabilities instead of binned $D_b$ and $D_c$ distributions is currently not recommended. This motivates further investigation into probability calibration [23] to improve the reliability and applicability of this approach.

Overall, analyses of $H \to c\bar{c}$ decays remain challenging due to the very low signal-to-background ratio, making large future data sets essential for improving statistical precision. The study also confirms that flavor–tagging features are among the most discriminative, indicating that advances in charm–tagging algorithms could play a key role in enhancing sensitivity. Additionally, the Higgs candidate neural network could be further improved if truth labels for the origin of jets (from the Higgs boson, top quarks, or $W$ bosons) were available, allowing the network to learn more precise jet–level patterns. Further improvements might also come from exploring more sophisticated model architectures and optimizing hyperparameters.

# Bibliography


[1] Wikipedia contributors, *Standard Model*, URL: https://en.wikipedia.org/wiki/Standard_Model (visited on 08/20/2025).

[2] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13** (1964), URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.321.

[3] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13** (1964), URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.508.

[4] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Physics Letters B **716** (2012), ISSN: 0370-2693, URL: http://dx.doi.org/10.1016/j.physletb.2012.08.020.

[5] ATLAS Collaboration, CMS Collaboration, *Combined Measurement of the Higgs Boson Mass in $pp$ Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments*, Phys. Rev. Lett. **114** (2015), URL: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.114.191803.

[6] ATLAS Collaboration, *Prospects for $H \to c\bar{c}$ using Charm Tagging with the ATLAS Experiment at the HL-LHC*, ATL-PHYS-PUB-2018-016, ATL-PHYS-PUB-2018-016, CERN: CERN, 2018, URL: https://cds.cern.ch/record/2633635.

[7] L. Evans and P. Bryant, *LHC Machine*, Journal of Instrumentation **3** (2008), URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08001.

[8] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, Journal of Instrumentation **3** (2008), URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08003.

[9] *Branching Ratios Higgs $\to 2\,Fermions$*, URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageBR#Higgs_2_fermions (visited on 08/07/2025).

[10] Particle Data Group, *Review of particle physics*, Phys. Rev. D **110** (2024).

[11] Dr. E. Schopf, Private communication, 2025.

[12] ATLAS Collaboration, *Luminosity determination in $pp$ collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC*, Eur. Phys. J. C **83** (2023), URL: https://link.springer.com/article/10.1140/epjc/s10052-023-11747-w.

[13] *Cross Sections $t\bar{t}H$*, URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageAt13TeV#ttH_Process (visited on 08/07/2025).

[14] *Cross Sections $t\bar{t}$*, URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/TtbarNNLO (visited on 08/07/2025).

[15] ATLAS Collaboration, *Measurements of $WH$ and $ZH$ production with Higgs boson decays into bottom quarks and direct constraints on the charm Yukawa coupling in 13 TeV $pp$ collisions with the ATLAS detector*, J. High Energ. Physics. **2025** (2025), URL: https://arxiv.org/pdf/2410.19611.

[16] Wikipedia contributors, *Missing energy*, URL: https://en.wikipedia.org/wiki/Missing_energy (visited on 09/06/2025).

[17] ATLAS Collaboration, *Transforming jet flavour tagging at ATLAS*, 2025, arXiv: 2505.19689 [hep-ex], URL: https://arxiv.org/abs/2505.19689.

[18] R. Bouquet (INFN Genova), Private communication, 2025.

[19] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, The European Physical Journal C **71** (2011), URL: http://dx.doi.org/10.1140/epjc/s10052-011-1554-0.

[20] L. Breiman, *Random Forests*, Machine Learning **45** (2001).

[21] F. Pedregosa et al., *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research **12** (2011), URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016 164, URL: http://www.deeplearningbook.org.

[23] ATLAS Collaboration, *A continuous calibration of the ATLAS flavour-tagging classifiers via optimal transportation maps*, 2025, URL: https://arxiv.org/abs/2505.13063.

# List of Figures

# List of Tables