Gradient boosting based estimates for trigger efficiency in the CMS experiment

Master's thesis
University of Turku
Physics
2024
B.Sc. Meeri Harkki
Examiners:
Dr. Santeri Laurila
Prof. Jyrki Piilo

UNIVERSITY OF TURKU
Department of Physics and Astronomy

Harkki, Meeri Gradient boosting based estimates for trigger efficiency in the CMS experiment

Master's Thesis, 83 pp.
Physics
August 2024

The Standard Model of particle physics describes the nature of fundamental physics, and it is so far the most precise theory regarding this topic. All the particles predicted by this theory have been experimentally discovered, with the Higgs boson being the latest in 2012. The Standard Model, however, cannot explain some of the known phenomena, such as the dark matter or the asymmetry between matter and antimatter, and therefore studies regarding extended versions of the Standard Model are being conducted.

The European Organization for Nuclear Physics (CERN) operates the world's leading particle physics laboratory. This thesis is conducted in the context of the Compact Muon Solenoid (CMS) experiment, which is one of the four main experiments in CERN's Large Hadron Collider (LHC). The particle beams accelerated in the LHC collide inside the CMS nearly 40 million times per second, meaning that the amount of data these collisions produce would be impossible to save in its entirety. This is why the CMS implements a trigger system, which is used to discard the majority of the collision events, while still keeping the most interesting ones.

In this thesis, a new method for measuring the trigger efficiency using a gradient boosting algorithm is presented. Trigger efficiency measures how well the trigger is able to select the events it is supposed to select, and it is defined as the ratio between the events the trigger accepted and all of the events it is supposed to accept. In many ongoing analyses in the CMS, a combination of several trigger algorithms is used, and this may cause complicated relationships between the variables, which the efficiency can be dependent on. These dependencies are difficult to estimate using traditional trigger efficiency measurement methods, so this new method targets these cases, where the efficiency can depend on several partially correlated variables. This method is developed using data and simulation samples from two ongoing analyses in the CMS: a boosted Higgs boson pair production analysis, and a charged Higgs boson search.

Keywords: Gradient boosting, trigger efficiency

UNIVERSITY OF TURKU
Fysiikan ja tähtitieteen laitos

Harkki, Meeri Gradienttitehostuspohjaiset arviot liipaisutehokkuudelle CMS-kokeessa

Pro Gradu, 83 pp.
Fysiikka
Elokuu 2024

Hiukkasfysiikan standardimalli kuvaa perusvuorovaikutusten ja alkeishiukkasten luonnetta, ja se on toistaiseksi tarkin teoria kyseisestä aiheesta. Kaikki standardimallin ennustamat hiukkaset on havaittu kokeellisesti, viimeisimpänä Higgsin bosoni vuonna 2012. Standardimalli ei kuitenkaan kykene selittämään joitakin tunnettuja ilmiöitä, kuten pimeää ainetta tai epätasapainoa aineen ja antiaineen välillä, ja tästä syystä tutkitaan myös laajennettuja versioita standardimallista.

European Organization for Nuclear Physics (CERN) on maailman johtava hiukkasfysiikan tutkimuslaitos. Tämä tutkielma on toteutettu CMS-koeaseman (engl. Compact Muon Solenoid) kontekstissa. CMS on yksi CERNin suuren hadronitörmäyttimen (engl. Large Hadron Collider, LHC) pääkokeista. LHC-törmäyttimessä kiihdytetyt hiukkaset törmäävät CMS-hiukkasilmaisimessa lähes 40 miljoonaa kertaa sekunnissa johtaen siihen, että näiden törmäysten tuottama datamäärä olisi mahdotonta tallentaa kokonaisuudessaan. Tämän vuoksi CMS-koeasema hyödyntää liipaisujärjestelmää (engl. trigger system), jota käytetään hylkäämään suurin osa törmäystapahtumista säilyttäen silti kaikista kiinnostavimmat tapahtumat.

Tässä tutkielmassa esitellään uusi tapa mitata liipaisutehokkuutta (engl. trigger efficiency) käyttämällä gradienttitehostuspohjaista (engl. gradient boosting) algoritmia. Liipaisutehokkuudella mitataan liipaisimen kykyä valita niitä törmäystapahtumia, joita sen on tarkoitus valita, ja liipaisutehokkuus määritellään liipaisimen valitsemien törmäystapahtumien ja kaikkien törmäystapahtumien suhteena. Monissa CMS-koeaseman analyyseissä käytetään useiden liipaisinalgoritmien yhdistelmiä, mikä voi aiheuttaa monimutkaisia suhteita muuttujien välillä, joista myös liipaisutehokkuus voi olla riippuvainen. Nämä riippuvuudet ovat vaikeita arvioida käyttäen perinteisiä liipaisutehokkuuden arviointimenetelmiä, joten tässä tutkielmassa esitelty uusi menetelmä kohdistuu erityisesti näihin tilanteisiin, joissa liipaisutehokkuus voi riippua useista osittain keskenään korreloivista muuttujista. Tämä menetelmä on kehitetty käyttäen dataa ja simulaatiota kahdesta eri CMS-kokeessa käynnissä olevasta analyysistä: kahden korkean liikemäärän Higgsin bosonin tuotannon analyysistä, sekä sähköisesti varatun Higgsin bosonin etsinnästä.

Avainsanat: Gradienttitehostus, liipaisutehokkuus

Acknowledgements

I would like to express my sincere gratitude to the Helsinki Institute of Physics for the opportunity to write this thesis. I am particularly grateful to my supervisor Dr. Santeri Laurila for providing excellent guidance and feedback throughout my thesis journey. I would also like to thank my family and friends for their support and interest in my thesis.

July 25, 2024
Meeri Harkki

Contents

Acronyms

AUC	Area under the receiver operating characteristic curve
CERN	European Organization for Nuclear Research
CMS	Compact Muon Solenoid
DNN	Deep neural network
GB	Gradient boosting
ggF	Gluon fusion
HLT	High Level Trigger
LHC	Large Hadron Collider
QCD	Quantum chromodynamics
ROC	Receiver operating characteristic curve
SM	Standard Model
VBF	Vector boson fusion

Introduction

The European Organization for Nuclear Research (CERN) is the world's leading organization in studying phenomena predicted by the Standard Model (SM) of particle physics, as well as searching phenomena beyond the SM. The SM describes the fundamental particles and their interactions, but it does not include gravity. All of the particles predicted by the SM have been observed, the latest one being the Higgs boson discovered in 2012 [1–3]. However, the SM is known to be incomplete, and studies regarding physics beyond the SM continue on many fronts.

CERN operates the world's largest and most powerful particle accelerator, the Large Hadron Collider (LHC). In the LHC, particle beams going in opposite directions are accelerated until they reach high energies, and then they are collided in four crossing points. The LHC has four main experiments next to the points where the particle beams collide, and this thesis project has been carried out in the context of the Compact Muon Solenoid (CMS) experiment [4]. The CMS itself is a multipurpose apparatus, which is designed to study many different physics phenomena observed at the LHC, including the Higgs boson [2]. The data and simulation samples used in this thesis are from two ongoing analyses at CMS: a boosted di-Higgs analysis, which targets a pair of highly Lorentz-boosted Higgs bosons, and a charged Higgs boson analysis, which searches for electrically charged Higgs bosons.

This thesis is focused on the CMS trigger system, which is a crucial component in making sure the data from the most interesting particle collisions is saved and thus is available for detailed analysis. The idea of this thesis is to develop an algorithm to predict the trigger efficiency, which is a metric describing how well the trigger performs. The algorithm aims to estimate the trigger efficiency as a function of several variables at the same time, which is difficult to achieve using current trigger efficiency estimation methods. The new method will be demonstrated in the two example analyses, the boosted di-Higgs analysis and the charged Higgs boson analysis, to examine how well the method performs in analyses targeting different processes. The results given by this new method will be compared to the results given by traditional trigger efficiency measurement methods, and the systematic uncertainties of the new method will also be estimated.

1 Theoretical background

The theoretical background of this thesis is based on the Standard Model of particle physics, which describes three of the four known fundamental forces, as well as the known elementary particles in the universe. We will also introduce the Higgs mechanism, which explains how some of the elementary particles acquire mass. We will discuss the main concepts in these theories in the following sections. The focus is on the boosted Higgs pair and the charged Higgs boson production and decay processes, since we will use these analyses to demonstrate the new method presented in this thesis.

1.1 The Standard Model

The SM describes three of the known fundamental forces: the electromagnetic, weak, and strong interactions, not including gravity. The SM also presents the elementary particles, some of them mediating these fundamental interactions, and some of them being the building blocks of matter. The elementary particles in the model are divided into three groups, which are quarks, leptons, and gauge bosons. Besides these groups, there is also one scalar boson in the model, which is the Higgs boson. [5] An overview of the model is given in Figure 1.

Quarks and leptons are fermions, meaning that they have a half integer spin. This results in the fermions following the Pauli exclusion principle, which means that two fermions cannot occupy the same quantum state at the same time [6]. All matter is made out of these quarks and leptons, and they can be categorized into three generations depending on their properties. Each fermion has also an antiparticle, which has the same mass and spin, but opposite electric charge.

Gauge bosons, also known as vector bosons, have a spin value of 1. They can be described as force carriers, since they mediate the electromagnetic, weak and strong interactions. Photons mediate the electromagnetic interactions, W and Z bosons mediate the weak interactions, and gluons mediate the strong interactions [5]. Photons and gluons are massless, but W and Z bosons acquire mass via the Higgs mechanism, which will be explained in Section 1.2.

The SM is a quantum field theory, which means that it describes the forces and particles as excited states of quantum fields, which span across the whole universe. Quantum electrodynamics describes the interactions that are mediated by photons, and quantum chromodynamics (QCD), is a theory describing the strong interactions, which bind quarks together to form hadrons with the interactions mediated by gluons. [5]

The SM, however, is known to be incomplete, since it cannot explain several observed phenomena, such as the origin of dark matter or the asymmetry between matter and antimatter. This is why there are several broader theories, which attempt to expand the SM by adding new terms to the Lagrangian equation describing the fields and their interactions in the SM. These theories are known as beyond the Standard Model (BSM) theories. There exist, for example, models which have two Higgs doublets instead of only one (as in the SM). These models predict the existence of five scalar bosons, two of which would be electrically charged [7]. The discovery of a charged Higgs boson would be one way to find evidence of the existence of new physics. There might also exist other, heavier particles predicted by these BSM theories, which are not yet discovered. One way to find evidence of these particles might be to study the rate of two Higgs bosons being produced at the same time, and compare this production rate to the SM prediction, since this rate could be modified by the presence of the new heavy particles in the process. To summarize, there are several ways of searching for phenomena indicating physics beyond the SM, and experiments for discovering new physics are continuing actively.

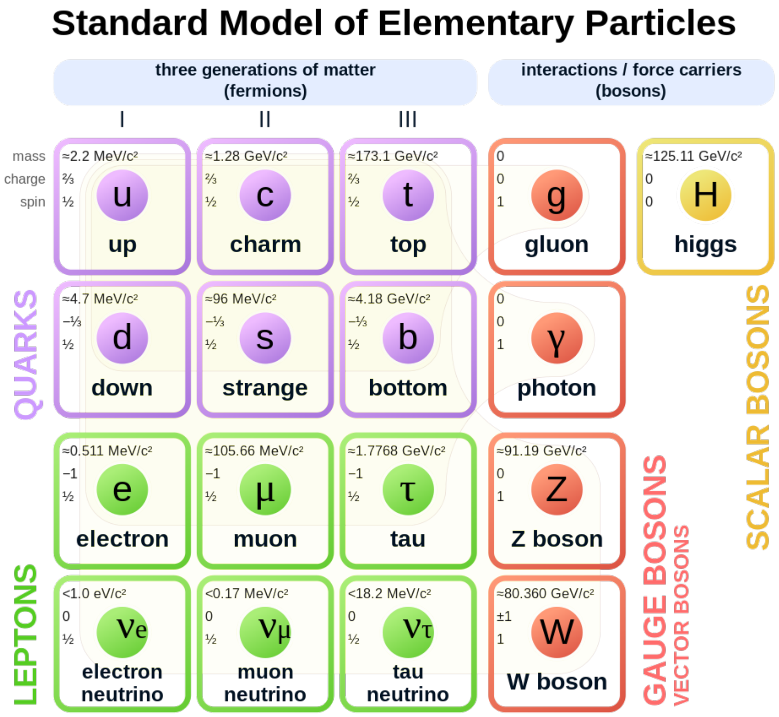


Figure 1: The elementary particles of the Standard Model [8]. The particles are divided into groups that are quarks, leptons, gauge bosons and scalar bosons. The quarks and leptons are the constituents of matter, and the bosons are force carriers.

1.2 The Higgs Mechanism

The Higgs mechanism is crucial in explaining how the W and Z bosons have mass, since without this theory all the gauge bosons would be considered massless. However, measurements show that the W and Z bosons do have mass, while photons and gluons do not. Three independent research groups predicted the Higgs mechanism in 1964, and therefore the mechanism is also known as the Englert–Brout–Higgs–Guralnik–Hagen–Kibble mechanism. [9–11]

The Higgs mechanism introduces a field known as the Higgs field, which induces a spontaneous symmetry breaking below a certain temperature. The symmetry breaking subsequently initiates the Higgs mechanism, and the W and Z bosons gain mass by interacting with the Higgs field. The Higgs field is also able to generate masses for all the fermions, but the mechanism is different than in the case of gauge bosons.

The Higgs boson (H) is an excited state of the Higgs field, and after decades of research a Higgs boson with a mass of approximately 125 GeV was detected in the ATLAS and CMS experiments at the LHC in 2012 [1–3]. For this discovery Peter Higgs and Francois Englert were awarded a Nobel Prize in 2013 [9–11]. The Higgs boson has a spin value of zero and the discovered boson has no electric charge.

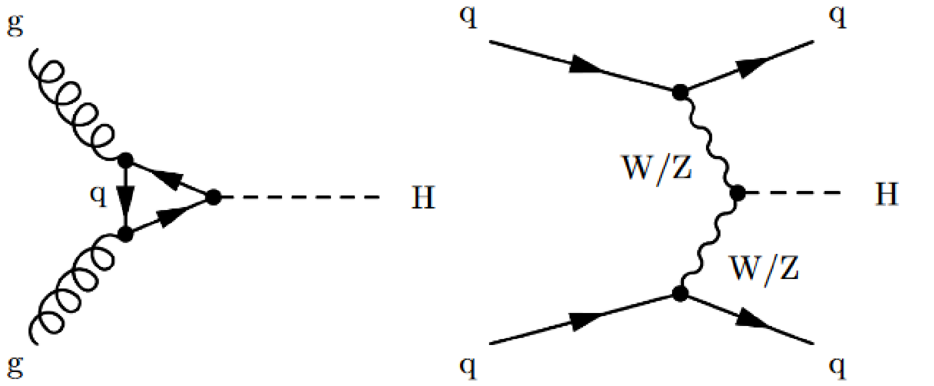
1.3 Higgs boson pair production

A Higgs boson is produced at the LHC mainly in two different ways: via gluon fusion (ggF) or via vector boson fusion (VBF). Examples of the Feynman diagrams of these production modes are presented in Figure 2. In gluon fusion, two gluons from two protons in a bunch crossing interact with each other to produce a new Higgs boson. In vector boson fusion, the quarks interacting with each other when the protons collide radiate a vector boson, and these vector bosons then interact with each other. The result is a new Higgs boson and two final-state quarks.

A Higgs boson pair (HH) can also be produced via these two production modes by a phenomenon known as the Higgs boson self-interaction. This results in two Higgs bosons being produced at the same time, and this process is called the Higgs boson self-coupling. Examples of the Feynman diagrams depicting this self-coupling in the ggF and VBF production modes are presented in Figure 3.

1.3.1 Higgs boson pair decay modes

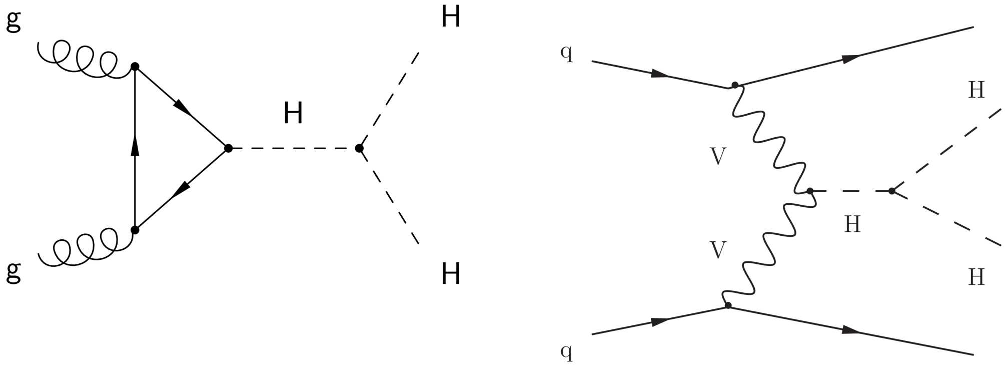
We cannot observe the Higgs boson directly at the LHC, since the Higgs boson decays rapidly into other particles. This is why we need to detect the Higgs pair production through different decay modes. One largely studied decay mode is the



(a) ggF

(b) VBF

Figure 2: Example diagrams for two Higgs boson production modes: gluon fusion (ggF) and vector boson fusion (VBF). [12]



(a) Self-coupling via ggF

(b) Self-coupling via VBF

Figure 3: Example diagrams for two Higgs boson pair production modes: gluon fusion (ggF) and vector boson fusion (VBF). [13]

Higgs boson pair decay into a pair of bottom quarks and a pair of τ leptons, and the data and simulation samples related to this decay are utilized in this thesis.

The dominant decay modes for the Higgs boson are the boson decaying into b quark-antiquark pair, and the boson decaying into a W boson pair. The probability of a decay mode is called a branching fraction. The branching fraction of the $\mathrm{H} \to \bar{\mathrm{b}}\mathrm{b}$ decay mode is 58% [14], and the branching fraction of the $\mathrm{H} \to \mathrm{WW}$ decay is 21%. The boson can also decay into two τ leptons with a branching fraction of 6.3% [15]. The combined branching fraction for the decay where one of the Higgs bosons in a Higgs boson pair decays into a b quark-antiquark pair and the other into a τ lepton pair is 7%, and this is the decay mode we will focus on in this thesis.

When the collisions produce quarks and gluons, it is not possible to observe the quarks and gluons separately, since they become jets of hadrons, which then interact with the detector. In the case when a Higgs boson has a low momentum, the bottom quarks are separated into their own small-radius jets, but when the momentum is high, the two b-quarks are Lorenz-boosted and form together one large-radius jet. In this thesis, we use the data and simulation samples targeting this boosted state, where we have high-momentum jets which have a large radius. In developing the new machine learning method presented in this thesis, we will use a background sample that simulated multijet production via QCD interactions, since this is the dominant background process related to this analysis. We will also use simulated signal samples related to the boosted Higgs pair production modes ggF and VBF. Besides the simulated samples we will use data collected in the CMS experiment in 2018.

1.4 Charged Higgs boson production

The charged Higgs boson ($\mathrm{H}^{\pm}$) is a hypothetical particle predicted in the two-Higgs-doublet models, and it can be either positively charged (H^{+}) or negatively charged

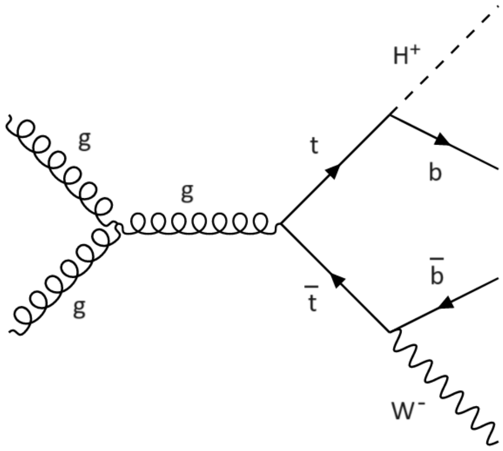
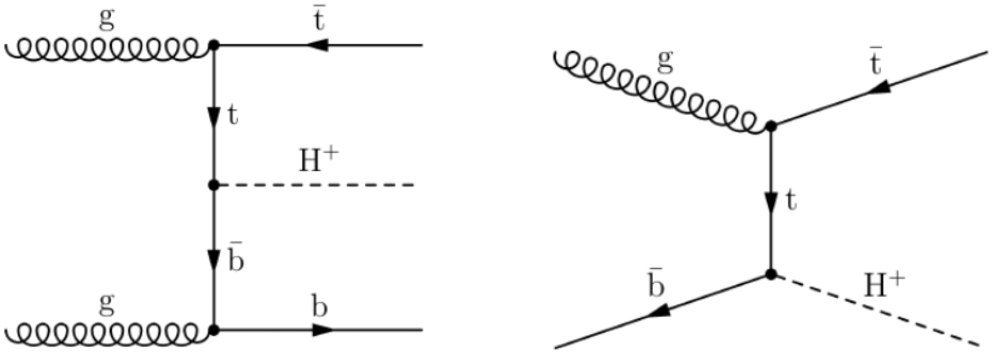


Figure 4: The light charged Higgs boson production.

(H^-), with these two particles being antiparticles to each other [7]. The $\mathrm{H}^\pm$ has different possible mass values, which have an impact on how the $\mathrm{H}^\pm$ can possibly be produced. The charged Higgs boson can either be light, meaning that its mass is smaller than the mass difference between a top and a bottom quark ($m_{H^\pm} < m_t - m_b$), or it can be heavy, if its mass is larger than the top and bottom quark mass difference ($m_{H^\pm} > m_t - m_b$). The light charged Higgs boson is mostly produced via decays of top quarks ($\mathrm{t} \rightarrow \mathrm{bH}^\pm$), which is presented in Figure 4. The heavy charged Higgs boson is mainly produced via $\mathrm{tb} \rightarrow \mathrm{bH}^\pm$, and this process can be understood in two ways. We can represent the process through a four-flavour scheme or a five-flavour scheme, and the order of calculations is different in these schemes. The Figure 5a represents the four-flavour scheme, where the dominant production mode is $\mathrm{gg} \rightarrow \mathrm{tbH}^\pm$ and the secondary mode is $\mathrm{qq}' \rightarrow \mathrm{tbH}^\pm$. In the five-flavour scheme, the dominant production mode is $\mathrm{gb} \rightarrow \mathrm{tH}^\pm$, which is presented in Figure 5b. [16]

There is also an intermediate region for the charged Higgs boson, where its mass is close to the top quark mass ($m_{\mathrm{H}^\pm} \sim m_{\mathrm{t}}$). In this case, both of the production modes introduced earlier are contributing to the process and their interference needs to be taken into account. [17]



(a) Four-flavour scheme production of the heavy charged Higgs boson.

(b) Five-flavour scheme production of the heavy charged Higgs boson.

Figure 5: Heavy charged Higgs boson production modes. [18]

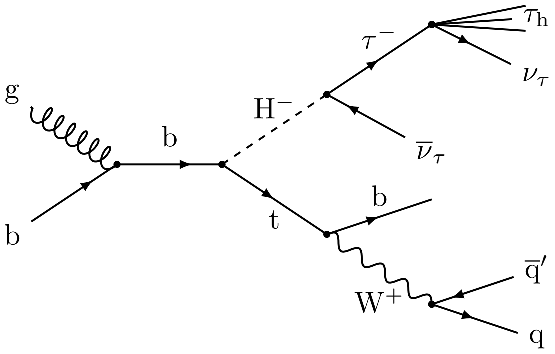


Figure 6: Decay of the heavy charged Higgs boson into a final state with a hadronic τ lepton.

1.4.1 Charged Higgs boson decay modes

The decay modes and the branching fractions of a charged Higgs boson again depend on the mass of the boson. If the charged Higgs boson is heavy, $H^\pm \to \tau^\pm\nu_\tau$ and $H^\pm \to tb$ can be among the dominant decay modes [16]. The Feynman diagram of this heavy charged Higgs boson decay into the $\tau\nu$ final state is presented in Figure 6. Different decay modes are also possible, such as the decay to $\mu\nu$ and cs, but the branching ratios for these decay modes are smaller than for the $\tau^\pm\nu_\tau$ and tb.

For the light charged Higgs boson, the dominant decay mode depends on the cho-

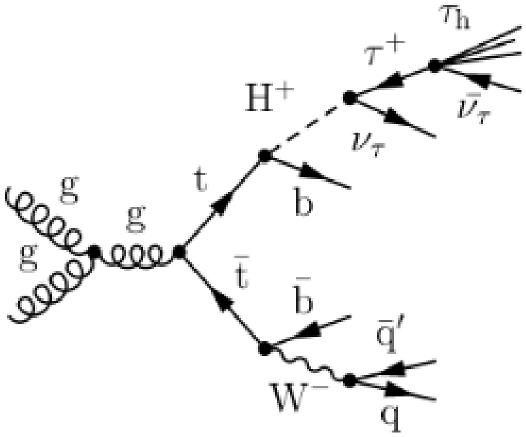


Figure 7: Decay of the light charged Higgs boson into a final state with a hadronic τ lepton. [18]

sen two-Higgs doublet model and its parameters. In some scenarios, the $H^\pm \to \tau^\pm \nu_\tau$ decay is significantly the dominant decay mode, but in other scenarios decay modes such as $H^\pm \to cs$ and $H^\pm \to \tau^\pm cb$ can also be significant [14]. The Feynman diagram of the decay into the $\tau\nu$ final state is presented in Figure 7.

Since the hadronic τ lepton is a heavy particle, it can decay further into hadrons. In the charged Higgs boson analysis, the targeted hadronic decay modes are usually either the τ decaying into one charged pion and one or two neutral pions (one-prong) or three charged pions (three-prong) [16].

In this thesis, we will use simulation and data samples related to an ongoing analysis used to search for the charged Higgs boson as another example case for the new machine learning based trigger efficiency estimation method. We will use simulated signal samples related to the mass values where the charged Higgs boson is considered heavy. The background sample we use simulates the $t\bar{t}$ production, since this is the dominant background process for the hadronic final state [16]. Similar to the boosted Higgs boson pair case, we will use data collected in the CMS experiment in 2018 in this charged Higgs boson analysis.

2 The CMS experiment

In this section, we will discuss the experimental setup at CERN, which is used to measure particles and their interactions predicted by the SM and search for phenomena beyond the SM. This section introduces the Large Hadron Collider, as well as the Compact Muon Solenoid experiment, from which the data used in this thesis are from. We will also use simulation samples based on this data.

2.1 Large Hadron Collider

The LHC is a 27-kilometer long circular particle accelerator located in the Geneva region at the European Organization for Nuclear Research, CERN. The LHC is located between 45 and 175 meters underground, in a tunnel that was originally built for CERN's Large Electron Positron collider [19]. The LHC project was approved by the CERN Council in 1994, and the first particle collisions in the LHC happened in 2008. The LHC is at the moment the largest and the most powerful particle accelerator in the world.

The LHC consists of two circular beam pipes where the particle beams can be accelerated in opposite directions, before they collide in four collision points. Before the particle beams reach the beam pipes in the LHC, they are accelerated gradually in several smaller linear and circular accelerators that are connected into the LHC.

Since the LHC has such a high circumference, the particle collisions in it can reach very high energies. This is because the momentum of a charged particle in a circular orbit is directly proportional to the radius of the orbit. Therefore, by increasing the radius of the collider we can increase the highest possible momentum the charged particles can reach. However, a higher momentum of a charged particle also requires a stronger magnetic field to keep the particle on the circular path. This is why the trajectory of the particle beams is controlled by superconducting electromagnets, which ensure that the beams travel steadily in the circular beam pipes, and that the beams don't fall apart in the process [19].

The energy that the collisions in the LHC reach currently is $\sqrt{s}$= 13.6 TeV, where $\sqrt{s}$ means the center-of-mass energy. In this thesis, we will use data from proton-proton collisions at a center-of-mass energy of 13 TeV. The luminosity that the LHC reached during the data taking run where the data used in this thesis was acquired is around $L \approx 2 \cdot 10^{34}$ cm^{-2}s^{-1}, where luminosity tells how many collisions happen in the accelerator in a given amount of time. [19]

In the four collision points along the LHC ring there are four major experiments: Compact Muon Solenoid (CMS) [4], A Toroidal LHC Apparatus (ATLAS) [20], A Large Ion Collider Experiment (ALICE) [21] and LHC-beauty (LHCb) [22]. The CMS and ATLAS can be called general-purpose detectors, since they can be used to detect a broad range of physics phenomena, including the study of the Higgs boson [1–3]. The ALICE and LHCb have more specific purposes: ALICE is designed to study quork-gluon plasma, which can be produced by colliding heavy ions, and LHCb specializes in questions concerning antimatter by studying the interactions of the b hadrons.

Besides these experiments next to the collision points, the LHC has five more experiments: LHCf (The Large Hadron Collider forward) [23], which simulates cosmic rays, TOTEM (Total cross section, Elastic scattering and diffraction dissociation Measurement at the LHC) [24], which is used for precise measurements of the proton-proton interaction cross-section, MoEDAL-MAPP (Monopole and Exotics Detector at the LHC and MoEDAL Apparatus for Penetrating Particles) [25], which is used to search the magnetic monopole, FASER (Forward Searching Experiment) [26], which is used for searching light and extremely weakly interacting particles, and SND@LHC (Scattering and Neutrino Detector at the LHC) [27], which is designed to study neutrinos.

2.2 Compact Muon Solenoid

The CMS is a general-purpose detector located at one of LHC's four collision points around a hundred meters underground [4]. Like mentioned earlier, the detector is designed to observe a wide range of physics phenomena that can be produced at the LHC. Inside the detector, the proton bunches each containing around 10^{11} protons collide up to 40 million times in one second, and in each of these bunch crossings we get around 40–60 proton-proton interactions. We refer to these bunch crossings containing several proton-proton interactions as events.

The name of the detector comes from the fact that although the detector weights 14 000 tonnes, has a diameter of 15 meters and a length of 29 meters, the detector is quite compact compared to other detectors with similar weight. The detector is also designed to detect muons very accurately, and it uses a powerful solenoid magnet to bend the tracks of the charged particles.

The CMS consists of several layers each designed to detect different particles. A view of the layers of the detector is given in Figures 8 and 9. As seen from Figure 8, the detector is cylindrical in shape, and it consists of several different components used to measure the properties of particles produced in the collisions, as detailed below. A slice of the detector and its different components are shown in Figure 9.

2.2.1 Tracking system

The purpose of the silicon-based tracking detectors located at the innermost part of the CMS is to record the trajectories of the charged particles produced in the collisions as accurately as possible. The tracking system can be used to reconstruct the trajectories of muons, electrons and charged hadrons. The trajectories of the charged particles are affected by the solenoidal magnet that bends their paths, which makes it possible to measure the charge and the momentum of the particle. [4]

The tracker has to be made so that it disturbs the path of the particles as

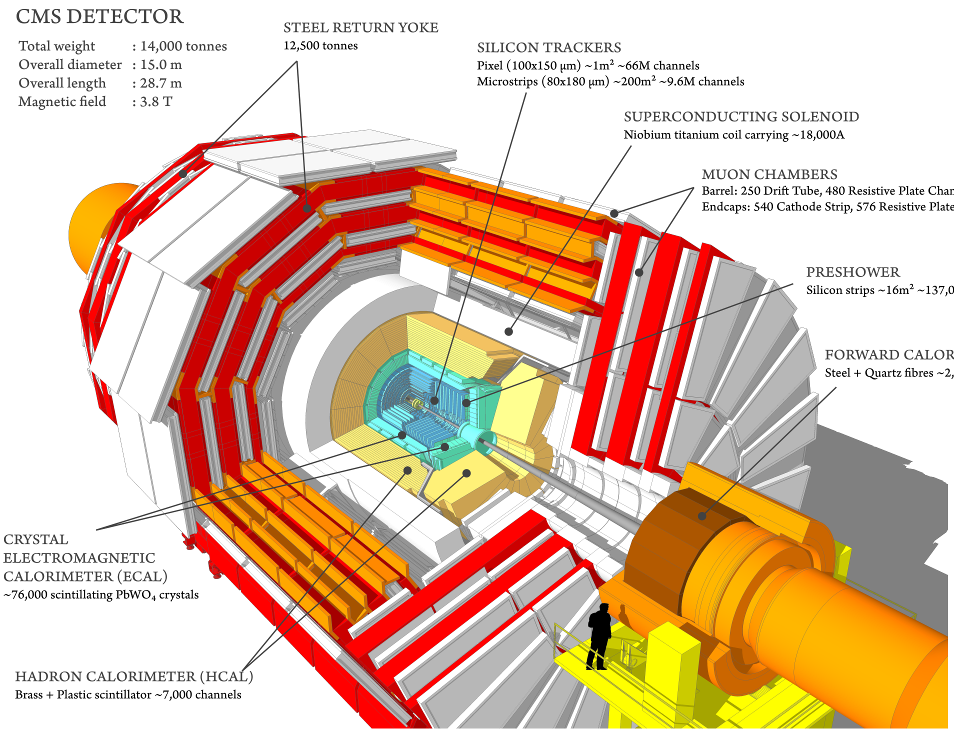

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~1m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 8: A view inside the CMS detector. The different parts of the detector, such as the calorimeters, trackers and the muon chambers are labeled. [28]

little as possible, and so the tracker is constructed to measure the positions of the particles so accurately that the trajectories can be reconstructed based on only a few measurement points. The design of the tracker therefore requires high granularity, and because there will be a large amount of particles produced in each collision, the tracker has to have a fast response time to attribute the trajectories to the right bunch crossing. Since the tracker system is located near the collision point in the detector, the tracker also faces severe radiation damage caused by the intense particle flux. [4]

All of these requirements mentioned previously have lead to the tracker being made of entirely silicon components. The tracker consists of a pixel detector with three barrel layers which have a radius of 4.4 cm and 10.2 cm, as well as a silicon strip tracker with 10 barrel detection layers, which extend to a radius of 1.1 m. The tracker also contains endcaps, which extend the tracking area on each side of the

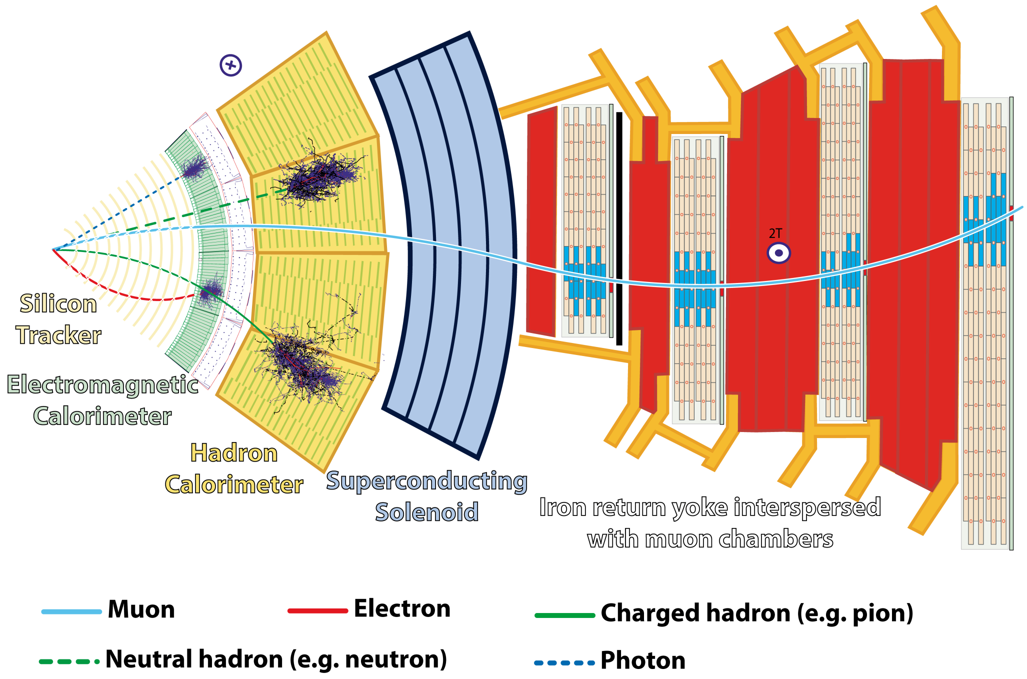

Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with muon chambers
2T
Muon
Electron
Charged hadron (e.g. pion)
Neutral hadron (e.g. neutron)
Photon

Figure 9: A slice of the CMS detector, illustrating the detector components where different particles are detected. [29]

barrel up to $|\eta| = 2.5$. [4]

2.2.2 Calorimeters

The CMS contains two calorimeters: closer to the tracker is the electromagnetic calorimeter (ECAL) and outside of that is the hadron calorimeter (HCAL). The ECAL is used to measure the energy of electrons and photons, and the HCAL is used to measure the energy of hadrons, as well as to enable indirect measurements of non-interacting particles like neutrinos. The measurement of non-interacting particles is possible since the HCAL is almost hermetic, which is why there isn't any region where the energy of the charged particles produced in the collisions is able to go undetected. [4]

The ECAL is made of 61 200 lead tungstate ($PbWO_4$) crystals, which are located in the central barrel section. In both of the endcaps there are also 7324 crystals in each of them. The lead tungstate crystals have a high density, a short radiation length, which represents the energy loss of high-energy particles when they interact electromagnetically with certain material, and a small Molière radius, which describes the spread of a particle shower initiated by a high-energy particle traversing a material. These properties result in the crystals having high granularity and making them a compact calorimeter.

The HCAL sits between the ECAL and the solenoid magnet, and it consists of a barrel (HB), endcap (HE) and forward (HF) regions. The HFs are positioned on both ends of the CMS, and they receive the majority of energy produced in the particle collisions. This is why the HFs are designed to withstand higher levels of radiation than the other parts of the HCAL. Since the volume of the HCAL is restricted by the solenoid, there is also an outer hadron calorimeter (HO), also known as a tail catcher placed outside the solenoid to complement the barrel calorimeter. The HCAL is mostly made out of brass.

The HCAL is designed to detect the position, energy and arrival time of the particles, and it does this by using alternating layers of absorbent and scintillator materials. When a particle passes through the calorimeter, the scintillator layers produce a rapid light pulse, which is then collected by optical fibers and fed forward into readout boxes. The measure of a particle's total energy is then the amount of photons in a given region summed over several layers of tiles.

2.2.3 Muon chambers

Muon detection is an important feature of the CMS experiment, as the name suggests. Muon detection can be used to recognize signatures of several interesting processes, since muons can be produced in various SM and BSM processes. For example, the Higgs boson can decay into four muons. Muons are also produced at a high rate at the LHC, and since they can be a signature of many interesting events they are examined in multiple different analyses in the CMS. [4]

The muon chambers are located at the outer part of the detector, since muons can penetrate through the calorimeters without losing much of their energy. The muon system is designed to do muon identification, momentum measurement and triggering. These are performed using three types of gaseous particle detectors: drift tubes (DTs), cathode strip chambers (CSCs) and resistive plate chambers (RPCs). The muons can be detected by measuring the current flow produced by the ionization of the gas when a particle enters in the detector.

The DTs are used in the barrel region, where the muon rate is low and the magnetic field is uniform. A DT chamber consists of cells filled with a mixture of argon and CO_2 gas, as well as an anode wire in the middle and cathode surfaces on the sides, which makes it possible for the ionization to produce the current flow.

In the two endcap regions where the muon rates are high and the magnetic field is non-uniform, the CSCs are used to identify the muons. The CSCs are suitable for this region, since they have a fast response time, fine segmentation and they are radiation resistive. The cells of the CSCs consist of several anode wires in the middle and cathodes at the top and bottom of the cell.

The RPCs act as a complementary system. The RPCs are located both at the barrel region and in the endcaps, and they are especially important in the overlap region between these parts. The RPCs are double-gap chambers which produce a fast response with good time resolution, but less precise position resolution than the DTs or CSCs.

The muon detectors and calorimeters are also used together to measure the missing transverse momentum (p_T^{miss}), which is used to infer the presence of neutrinos or other weakly interacting particles. Neutrinos interact so weakly with the detector that they cannot be observed directly, so to look for neutrinos we will have to observe the momentum imbalance in the plane perpendicular to the beamline, which is known as the transverse plane. The particle beams accelerated in the LHC have momentum only in their direction of movement at first. According to the conservation of momentum, since there are no momenta in the transverse plane before the collision, the momenta of the particles produced in the collision must also add up to zero. When neutrinos or other weakly interacting particles are produced, they escape the detector without interacting, causing an imbalance in the measured transverse momentum. The missing transverse momentum needed to balance the sum of the momenta in the transverse plane corresponds to the momenta of the neutrinos or other weakly interacting particles, and this is referred to as missing transverse momentum. Instead of all of the reconstructed particles we can also use jets to calculate the missing transverse momentum, which is then referred to as missing H_T, where H_T is the sum of jets.

3 The CMS trigger system

In CMS, the particle beams collide every 25 ns, and every event in the collision produces about 1 MB of data. In just one second, the beams collide up to 40 million times. With this output rate, it would be completely impossible to store all the events in the collisions, so we need a system that is able to reduce the amount of events to save drastically, while still making sure we save the most important events. To do this, we have a trigger system, which is designed to filter out events based on pre-made criteria, and save only the events that fit the criteria in question. The trigger system in CMS is designed to reduce the output rate in two steps: first the Level-1 (L1) trigger reduces the output rate to around 100 kHz, and then the High Level Trigger (HLT) reduces the rate even further to around 1 kHz [30].

3.1 Level-1 trigger

The L1 trigger system is physically situated in the service cavern, right next to the experimental cavern where the detector resides, and it receives information directly from the calorimeters and the muon detectors. The L1 trigger performs the first online event selections, where "online" refers to the events being processed in real time as the collisions are happening inside the detector. The system is completely hardware-based. In practice, this means that the trigger system consists of processors which are largely based on Field Programmable Gate Arrays (FPGAs). The L1 trigger has also a so called trigger menu, which consists of a large number of different trigger algorithms designed to look for different kinds of physics events. [31]

The L1 trigger system makes the decisions to whether accept or reject an event in only around 4 microseconds, so the trigger is designed to look out for quite simple signs that might indicate that there is something interesting or unusual in the event in question. The L1 trigger is divided into three different trigger systems:

calorimeter, muon and global triggers.

3.1.1 Calorimeter trigger

The calorimeter trigger receives information about the energy and quality of an event from the ECAL and HCAL. The local energy deposits are known as trigger primitives. The calorimeter trigger is two-layered, where the first layer is responsible for the data pre-processing, which includes, for example, summing the transverse energies of the trigger primitives and calculating the ECAL/HCAL energy ratio. The data is then transmitted to the second layer, where the processor boards receive the full information regarding each event calculated in the first layer. In the second layer, the particle candidates are reconstructed based on the information received from the first layer, and the energy sums, such as H_T or p_T^{miss} are calculated. Finally, the data from the second layer is forwarded into the de-multiplexing board, which combines the partial vector sums calculated in the previous layer, and sends the data to the global trigger. [31]

3.1.2 Muon trigger

The muon trigger receives information from the muon chamber: the DTs, the CSCs and the RPCs. The muon trigger system consists of three different muon track finders that correspond to different areas of the detector. In the barrel region, the Barrel muon track finder takes input from the DTs and RPCs. The Endcap muon track finder receives input from the CSC and RPC chambers in the endcap regions. Besides these muon track finders there is also the Overlap muon track finder that covers the region between the barrel and endcap muon track finder regions, and it takes input from all the muon detection systems, which are the DTs, the CSCs and the RPCs.

3.1.3 Global trigger

The global trigger combines the information received from the muon and calorimeter triggers, synchronizes the input data arriving at different times and makes the final L1 trigger decision. The global trigger applies a threshold, or a combination of thresholds, to accept and sort the events, and the decision is made by using logical combinations from the calorimeter and muon triggers. The thresholds can be related, for example, to the energy of the events. A combination of these thresholds is referred to as a seed, and a combination of the seeds forms a trigger menu.

The decision made by the global trigger is transmitted to the Data Acquisition system (DAQ), which can read out the data for offline storage. Offline analysis is then performed on this stored data, where the data collected is processed to perform event reconstruction. The information used by the HLT to make the trigger decisions is received from the DAQ.

3.2 High Level Trigger

The HLT receives the full event information, including information from the tracker, so it is able to make more precise decisions about which events to save. The decision time for the HLT system is around 100 milliseconds, meaning that it takes a significantly longer time to make the decisions compared to the L1 system. The HLT is physically a CPU farm with tens of thousands of cores capable of processing the events in parallel. The HLT reduces the output rate of 100 kHz down to around 1 kHz, which is an output rate that we can manage. [32]

The HLT is able to identify the physics objects and reconstruct the events by using several layers of filtering. This also ensures that the system's bandwidth is not saturated with the large amount of data received in the HLT. To reconstruct the events, the HLT also uses a streamlined version of an algorithm called Particle Flow (PF), which combines information from all the subsystems to measure the properties of each particle in the event and creates a global description of the event [33].

Similarly to the Level-1 trigger, the HLT consists of a trigger menu with hundreds of different trigger algorithms, which in this case are referred to as trigger paths. The trigger paths also have different thresholds used to filter different kinds of events, and in this case the thresholds are limited by the thresholds used at the L1 stage, as well as by the total HLT output rate and the HLT timing.

The events that are accepted by the HLT are sent forward to the the CERN Tier-0 computing centre, where they will be stored permanently.

4 Trigger efficiency

Trigger efficiency is a metric used to measure how well the trigger in question is able to select the events that it is supposed to select [34]. Trigger efficiency is essentially defined as the number of events passing the trigger divided by the number of all targeted events, which we can also denote as

$$E = \frac{N_{pass}}{N_{all}}, \tag{1}$$

where E is the trigger efficiency, N_{pass} is the number of events that pass the reference trigger and N_{all} is the number of all events of interest. However, since we don't know the number of all events because the events that don't pass any triggers are discarded, we have to use a different method to get a reliable estimate of the event distribution across the whole sample. In this section, measuring trigger efficiency using a traditional procedure called the reference trigger method is explained, as well as the choice of a reference trigger, which we need to measure the trigger efficiency.

4.1 Reference trigger method

With simulated events, the efficiency calculation is straightforward, since we can simulate events regardless of whether they would pass some trigger or not. This means that we know the amount of all the events, and we're able to calculate the trigger efficiency as previously defined. On the contrary, with real data we don't have access to all the events, since the events that don't pass any triggers are lost permanently. To address this problem, we can utilize a reference trigger that selects events independently of the trigger we're interested in, which we call the signal trigger. The purpose of the reference trigger is to obtain unbiased data that represent the distribution of the events accurately, so we could get a reliable trigger efficiency even if we don't have access to all the events in the sample. With the reference trigger, the trigger efficiency can now be defined as the ratio of the events passing both the reference and signal trigger to the events passing the reference trigger:

$$E = \frac{N_{sig\&ref}}{N_{ref}}, \tag{2}$$

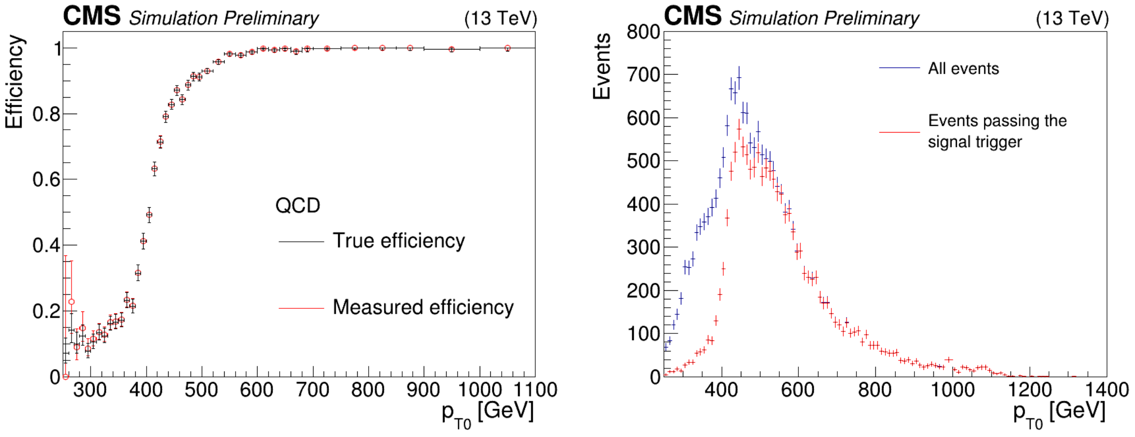
where $N_{sig\&ref}$ is the number of events passing the signal and the reference trigger, and N_{ref} is the number of events passing the reference trigger.

4.2 Choice of a reference trigger

In order to obtain the most accurate results, we have to choose a reference trigger or a trigger combination that is able to provide us a representative and large enough sample corresponding to the actual distribution of data. When choosing the reference trigger, we have to also take into account the signal trigger and the offline selections used in the analysis, which define the type of events the efficiency measurements are used for. We can do the reference trigger selection with the help of simulated samples.

In Figure 10a, the trigger efficiency is plotted as a function of the transverse momentum p_T of the large-radius jet with the highest transverse momentum in the event (p_{T0}). When we use a simulated sample we can calculate the efficiency in two ways: by using all the events in the simulated sample ("true efficiency") as in Equation 1, and by using a reference trigger ("measured efficiency") as in Equation 2.

From this example plot we can see that the efficiency increases steeply after the value 400 GeV for the p_{T0}, and this is called the turn-on region. The turn-on effect happens because of the energy resolution of the jets reconstructed in the trigger system. After a certain threshold the trigger will rapidly start selecting the events which pass this threshold, but since the resolution is limited, it is not able to pick all the events right after the threshold, so the efficiency will not immediately reach a value close to 1. However, after the turn-on region the efficiency reaches a plateau close to the value 1, which means that the trigger is able to steadily accept nearly



(a)

(b)

Figure 10: (a) Trigger efficiency as a function of the p_T of the large-radius jet having the highest p_T in the event (p_{T0}). The true efficiency is calculated by using all the events in the simulated sample, and the measured efficiency by using a reference trigger. (b) The distribution of events passing the signal trigger compared to the distribution of all events as a function of p_{T0}.

all the events in this p_T range. For analysis purposes, often the values before the turn-on region are not useful, so usually we focus on events close to the plateau. This is because there is a large amount of events close to the plateau, since the events which have larger p_T values are more rare.

This effect can also be observed in Figure 10b, where the distribution of events passing the signal trigger is compared to the distribution of all events in the simulated sample. As seen in the figure, most of the events in the sample fall in the p_{T0} range between 400 GeV and 600 GeV, which corresponds to the curve near the plateau in Figure 10a. As we can also observe in Figure 10b, the events which have a low p_{T0} are largely not accepted by the signal trigger, but the higher the p_{T0} is, the closer the distribution of events accepted by the signal trigger is to the distribution of all events in the sample. This can be compared to the trigger efficiency plot in Figure 10a, where in the low p_{T0} range the efficiency of the signal trigger is low, meaning that only a few events are selected, but in the high p_{T0} range almost every event gets selected, meaning that the efficiency of the signal trigger is almost 1.

As also seen in Figure 10a, the trigger efficiencies measured using all the events and the events passing the reference trigger agree quite well, which means that the chosen reference trigger would be suitable to use when measuring trigger efficiencies in data. The efficiencies have a slight difference regarding lower values, but since the lower values aren't important for analysis purposes, the mismatch between the efficiencies isn't concerning.

5 Modelling trigger efficiencies with machine learning

In this section, we will go through the motivation behind developing a machine learning method to estimate the trigger efficiencies. We will also discuss the theoretical background for the approach and introduce the machine learning architecture chosen for this task, as well as present the performance metrics used to evaluate the machine learning model.

5.1 The motivation and theoretical background behind the method

For many analyses in the CMS, the trigger efficiency can be maximized by using a combination of different trigger algorithms. However, this might lead into the efficiency being dependent on several partially correlated variables, which makes the trigger efficiency complicated to estimate using traditional methods. This effect is known as the curse of dimensionality, which in principle means that when we have data containing many variables, also known as high dimensional data, the data points become sparsely distributed in the data space. This results in a difficulty to obtain any reliable connections between the features of the data points. With the help of modern machine learning methods, our goal is to model these complex trigger efficiencies more reliably, with the algorithm able to consider the effects of these partially correlated variables and model the efficiency simultaneously as a function of these variables.

The idea behind the method is to consider the trigger itself as a binary classifier, which classifies each event into either one that passes or one that fails the trigger selection. Correspondingly, we should be able to train a binary classifier algorithm with the goal of reproducing the trigger selections as accurately as possible, and as a function of several variables. We could then compare the results of the algorithm to the results obtained with traditional methods.

The algorithm is trained by giving the model a set of offline-reconstructed variables as inputs, which are used to train the model to predict whether any given event in the sample either passes or fails the trigger based on the properties of the input variables. As output, the algorithm produces a weight between 0 and 1 for each event, and these weights represent the probability of each event to pass the trigger. To compare with the traditional methods, we can plot the trigger efficiencies by binning the events in the same way the results obtained by the traditional methods are binned, and taking the mean value of the weights of the events in the specific bin. This way we can produce a plot of the efficiency predicted by the algorithm as a function of a given variable of interest, which is comparable with the traditional trigger efficiency plots.

If the method in question works well, we can use the weights it gives to the events to estimate the performance of the triggers regardless of the accuracy of the trigger emulation in the simulation samples, whereas currently the estimation of the performance of triggers relies heavily on the simulation samples.

5.2 Gradient Boosting

Architecture of a machine learning model refers to the structure and components of the model, which affect the way the model handles data, how the model is trained and assessed, and how the predictions are made. There exist various possibilities for the architecture of the model when designing a binary classifier algorithm. The architectures we tried for this project included random forest [35], deep neural network (DNN) [36], multilayer perceptron [37] and gradient boosting [38, 39] algorithms. The algorithm that we chose was the gradient boosting algorithm, since it gave the most accurate results out of the other architectures we tried, it is easy to tune, and it is quite fast to train. The gradient boosting algorithm we utilized is provided by the Scikit-learn library [40].

Gradient boosting [38, 39], in some cases also known as gradient tree boosting or boosted decision trees, represents an ensemble method containing multiple weak learners, in this case decision trees. Decision trees are a type of supervised learning algorithm, which have a tree-like structure. They have a root node, which branches into internal nodes, and they branch into leaves, which represent all the possible outcomes in the dataset. The decision trees choose a specific outcome by splitting the data in these branches by their features, which then lead to specific leaf nodes representing specific outcomes.

In gradient boosting, the model improves the weak learners by focusing on the mistakes made in the previous iteration, which results in the weak learners evolving into strong learners that are able to perform better than the weak learners would on their own. The model does this by minimizing a loss function, which measures the accuracy of the model's predictions compared to the target values. This is done by setting the target of each subsequent model based on the residual error of the prediction. The process of gradient boosting is presented in detail next.

The gradient boosting algorithm can be notated as follows [39]:

Algorithm 1 Gradient Boosting

```
$F_0(\mathbf{x}) = \text{argmin}_\rho \sum_{i=1}^{N} L(y_i, \rho)$
for $m = 1$ to $M$ do:
    $\tilde{y}_i = -[\frac{\partial L(y_i,F(\mathbf{x}))}{\partial F(\mathbf{x}_i)}]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, i = 1, N$
    $\mathbf{a}_m = \text{argmin}_{\mathbf{a},\beta} \sum_{i=1}^{N} [\tilde{y}_i - \beta h(\mathbf{x}_i; \mathbf{a})]^2$
    $\rho_m = \text{argmin}_\rho \sum_{i=1}^{N} L(y_i, F_{m-1}(\mathbf{x}_i) + \rho h(\mathbf{x}_i; \mathbf{a}_m))$
    $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \rho_m h(\mathbf{x}; \mathbf{a}_m)$
end for
```

In the previous algorithm, the $L(y_i, \rho)$ represents the loss function, which can be, for example, the squared error or the absolute error in the case of regression, and the negative binomial log-likelihood for classification. The parameter y_i represents the target values, and the parameter ρ is a constant value that minimizes the loss function for the whole dataset. The model starts by making an initial prediction, which is usually the average of the target variable in the data for regression tasks, and the most frequent class in the case of classification. This is represented in the first line of Algorithm 1, where the $F_0(\mathbf{x})$ is the initial model, and the $\operatorname{argmin}_\rho \sum_{i=1}^{N} L(y_i, \rho)$ indicates the minimization of the loss function by finding a value for ρ that minimizes the total loss across the training samples. In the second line, the model is built with M trees, and the next lines are looped for all the trees. In the third line, the model calculates the pseudo-residuals $\tilde{y}$ from the initial prediction as the negative gradient of the loss function, and the pseudo-residuals represent the errors made by the model at this stage. The model then fits a weak learner $h(\mathbf{x}; \mathbf{a})$ multiplied by a scale factor β on the pseudo-residuals in the fourth line. The parameter $\mathbf{x}$ signifies the input feature vector, and $\mathbf{a}$ represents the parameters of the weak learner, for example the tree's structure and split locations. The result $\mathbf{a}_m$ represents the updated parameters of the weak learner. The parameter ρ_m, signifying a learning rate, is optimized in the next line. A learning rate, also known as a shrinkage, controls the impact each weak learner has on the final decision, and it is a value between 0 and 1. Finally, the model from the previous iteration $F_{m-1}(\mathbf{x})$ and the weak learner $h(\mathbf{x}; \mathbf{a}_m)$ multiplied by the learning rate ρ_m are summed together to form an updated model $F_m(\mathbf{x})$. This is done in the sixth line. This process is then repeated through a number of iterations until a predefined stopping criterion is met, and the final prediction is the sum of all the predictions made by the weak learners. The stopping criterion that the Scikit-learn library recommends for the gradient boosting classifier is the Friedman mean squared error [41].

Gradient boosting models can reach high accuracies, and they are also very flexible regarding the format of the input data. Gradient boosting models can be used in both classification and regression tasks. However, although gradient boosting models can often reach very accurate results, they can also be prone to overfitting, meaning that the model learns the noise and outliers of the training data excessively, and it is not able to generalize into new, unseen data. This can happen especially if we have a large number of trees or the trees have many nodes. However, risks for overfitting can be significantly reduced, when using appropriate performance metrics for the model. The risks can also be reduced when selecting subsets from the original dataset for training, testing and validating the model, and making sure that the model isn't able to use the testing and validation samples in the training phase to ensure that we get a good estimate of the model's capability to generalize into new, unseen data.

A large number of trees or nodes can also lead the model being computationally expensive, since the method cannot be parallelized due to the iterative nature of the algorithm. Another setback of gradient boosting, and especially complex gradient boosting models is their limited interpretability, since we might have tens or hundreds of decision trees in the model, and understanding the relationship between a certain input and its corresponding output can be difficult, or even impossible. However, there exist several explainability methods, such as post-hoc methods, which approximate complex models by creating more simpler surrogate models out of them [42]. These methods can be used, for example, in instances where it is ethically important to understand the relationship between certain inputs and outputs. Another explainability method is calculating the feature importance, which means that we calculate how much impact each feature has on the final decision the model makes [43]. These features can be ranked from the most important to the least important, and the importance of each feature gives us insight of how the model makes its decisions based on the different features. We will utilize this explainability method in this thesis.

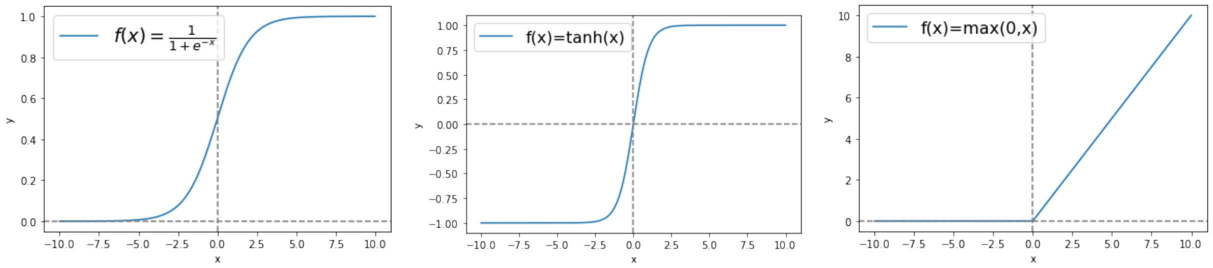
Gradient boosting models offer several hyperparameters which can be tuned in order to optimize the model performance. In our case, we tuned the number of estimators, which determines the amount of decision trees the model has, the learning rate, and the maximum depth of the tree, corresponding to the number of nodes in each tree. In the optimized model for the boosted HH analysis, we employed two hundred estimators, a learning rate of 0.25 and a maximum depth of two. For the charged Higgs boson analysis, we used a model with two hundred estimators, a learning rate of 0.75 and a maximum depth of two. The loss function used in both models was a logistic loss function, which is the negative logarithm of the likelihood function [44]. The logistic loss for a single sample with a true label $y \in \{0,1\}$ and a probability estimate $p = Pr(y = 1)$ can be defined as

$$L_{log}(y,p) = -(y\log(p) + (1-y)\log(1-p)).$$

5.3 Deep Neural Network

We will use another architecture besides the gradient boosting to study the systematic uncertainties related to this method. For this task we have chosen the deep neural network architecture, which will be explained next.

Neural networks are artificial intelligence models inspired by biological neural networks. In simple artificial neural networks (ANNs), there typically exists an input layer, a hidden layer, and an output layer. These layers consist of nodes, or neurons in this case, interconnected with each other. Each node carries a weight and a threshold. If the output of a node exceeds its threshold value, it forwards the data to the subsequent layer, and if the output is less than the threshold, no data is transmitted to the next layer. In practice, the output of a node is calculated using an activation function, which can be selected from several options. Some of the most popular activation functions are the sigmoid function, the hyperbolic tangent



(a) Sigmoid function (b) Tanh function (c) ReLU function

Figure 11: Activation functions for deep neural networks

(tanh) function and the rectified linear unit (ReLU) function. These are visualized in Figure 11. The sigmoid function sets a value between 0 and 1 for the output, the tanh function output ranges between -1 and 1, and the ReLU output is zero when the input is negative and equal to the input value when the input is positive. [45]

Deep neural networks (DNNs) refer specifically to ANNs which contain multiple hidden layers with different properties, for example a convolutional layer, which is usually used in image detection tasks, or a dense layer, where neurons are connected to every neuron in the preceding layer [36]. The activation functions for these layers can also differ. The DNN we trained for the boosted HH analysis case contains an input layer, two hidden layers and an output layer. All the layers are densely connected, and the first one has 12 neurons, the second one 10, the third one 8 and the output layer has one neuron, and the activation function for the first three layers is the ReLU function and the last one the sigmoid function.

5.4 Performance metrics

To assess how well the machine learning model performs in a classification task, there are several performance metrics. In our case, the performance metrics we will use are the accuracy, receiver operating characteristic curve (ROC) and area under the ROC curve (AUC), which will be explained next.

Accuracy is the most basic performance metric in use, and it can simply be

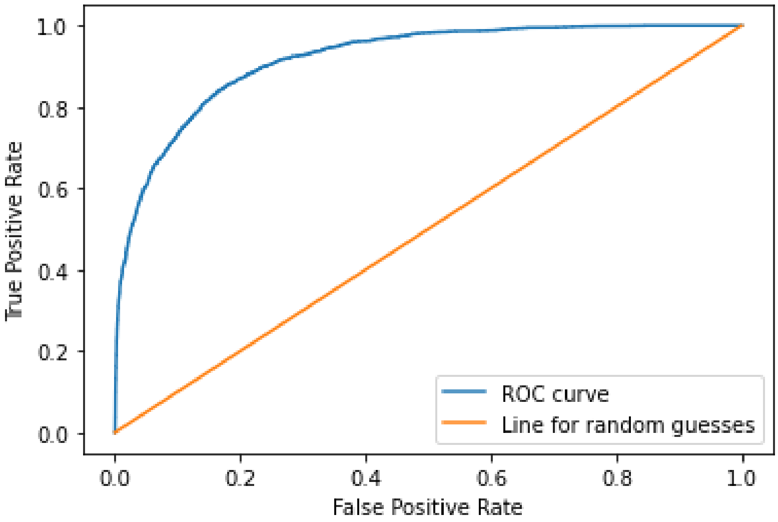


Figure 12: Example of a ROC curve of a classifier algorithm, compared to a curve of a model which makes completely random guesses.

defined as the fraction of correct predictions made by the model out of all predictions it made. It is a performance metric which is easy to understand and interpret. It can be formally denoted as:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}.$$

However, since accuracy just measures the number of correct predictions versus the total number of predictions, the accuracy can be very high in cases where some classes are much more frequent than others. For example, if we have a sample where 90% of the instances are 1 and 10% 0, the model will reach a 90% accuracy even if it classifies all the 0 values wrong. This is why accuracy should never be used as the only performance metric for a classifier model, especially when the dataset is imbalanced. [45]

The receiver operating characteristic curve (ROC) is another metric for assessing binary classifiers [46]. It plots the true positive rate (TPR) as a function of the false positive rate (FPR) at different threshold values. An example of the ROC curve is illustrated in Figure 12.

The true positive rate, also known as recall, corresponds to the amount of positive values the model is able to correctly classify as positive (true positives, TP) divided by the sum of true positive values and the positive values incorrectly classified as negative values, (false negatives, FN):

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

The false positive rate is correspondingly the amount of negative values the model has incorrectly classified as positive (false positives, FP) divided by the sum of false positives and the negative values the model was able to correctly classify as negative (true negatives, TN):

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

The FPR is equal to 1-TNR, where TNR is the true negative rate, which is also known as specificity. Thus, the ROC curve can also be understood as plotting the recall versus 1-specificity.

There is a trade-off between the recall and false positive rate, since the curve represents the true positive rate against the false positive rate at different threshold values, where a lower threshold produces more instances classified as positive (a higher TPR), but it might also classify more instances falsely as positive. A perfect model would have a ROC curve reaching into the top left corner of the plot, corresponding to a 100% TPR and 0% FPR. [46]

The ROC can be easily presented as a single scalar value, when we calculate the area under the curve (AUC). A perfect AUC score would then be 1, corresponding to a 100% amount of correct predictions, and a completely random model would have a 0.5 AUC score. The AUC can also be understood as the probability that the classifier ranks a random positive instance higher than a random negative one. [46]

6 Developing the method

In this section, we will explain the development of the new trigger efficiency estimation method using the boosted HH analysis and the charged Higgs boson analysis as example cases to illustrate the method.

6.1 The variables and the signal triggers

We will present the variables used as inputs for the gradient boosting model in both the boosted HH analysis and in the charged Higgs boson analysis. We will also explain the signal triggers used for both of these analyses, the boosted HH analysis having a combination of several trigger paths and the charged Higgs boson analysis having one signal trigger.

6.1.1 Boosted HH analysis

In the boosted HH analysis case we have a total of 14 variables of interest, which we use as inputs for the machine learning model. We will inspect the efficiency as a function of these variables across all of the three simulated samples as well as the data sample. The variables that we use are presented in Table I. The table includes the names and explanations of these variables. We use several variables related to the features of the large-radius jet having the highest transverse momentum in the event and the large-radius jet having the second highest transverse momentum in the event. We refer to these two large-radius jets as Jet_0 and Jet_1. The variables we use in this thesis are also used in an ongoing boosted HH analysis in the CMS.

The distributions of all variables in all of the three simulation samples are presented in Figure 13. In the figure, the distribution of all events and the distribution of events passing just the signal trigger are presented for all variables. We can note from these distribution curves that the distributions are slightly more similar for the QCD and VBF samples than for the QCD and ggF samples, which might affect the

Table I: The names and explanations of the variables used in the boosted HH analysis case.

Name	Explanation
p_{T0}	The p_T of the large-radius jet having the highest p_T in the event (Jet$_0$)
p_{T1}	The p_T of the large-radius jet having the 2nd highest p_T (Jet$_1$)
m_0	The mass of Jet$_0$
m_1	The mass of Jet$_1$
m_{HH}	The invariant mass between Jet$_0$ and Jet$_1$
η_0	The eta direction of Jet$_0$
η_1	The eta direction of Jet$_1$
$\Delta\eta$	The eta angle between Jet$_0$ and Jet$_1$
$\Delta\phi$	The phi angle between Jet$_0$ and Jet$_1$
H_T	Sum of all the jets in the event
H_T^{LR}	Sum of all the large-radius jets in the event
p_T^{miss}	Missing transverse momentum
$m_{HH} + p_T^{miss}$	Invariant mass added to the missing transverse momentum
$p_T^{miss} + p_{T0} + p_{T1}$	Sum of the missing transverse momentum and p_{T0} and p_{T0}

way the model trained with the QCD sample is able to perform in these two signal samples.

The distributions of events passing the reference trigger and the events passing both the signal and reference trigger in the data sample are presented in Figure 14. From this figure we can observe that the event distributions across the different variables are similar to the event distributions in the QCD sample, which is reasonable since the background mostly consists of QCD multijet events and therefore this simulation sample is supposed to resemble the distribution of events in actual data.

The trigger efficiency we want to measure as a function of all of these variables is

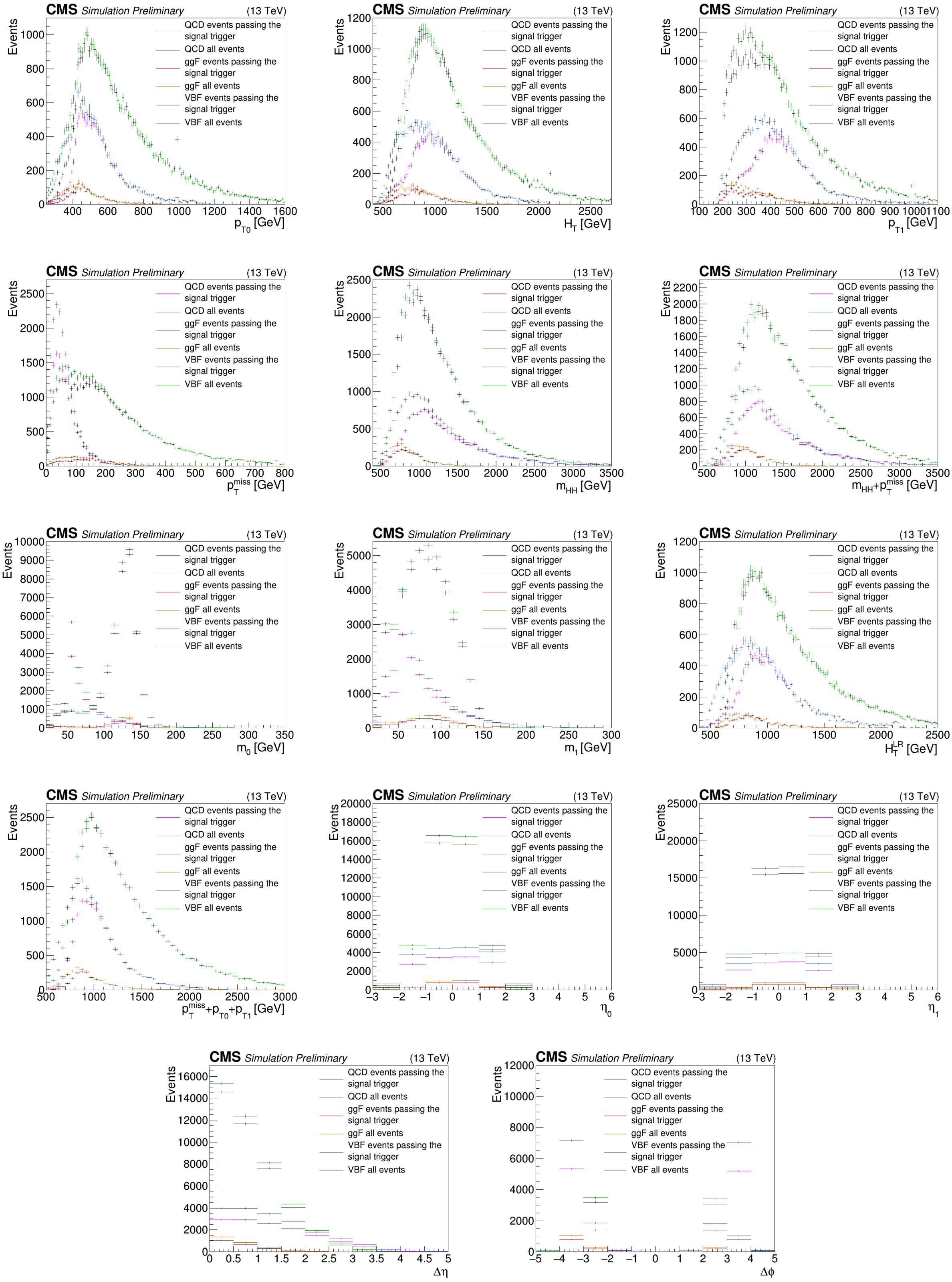


Figure 13: The distribution of all events and events passing the signal trigger in simulation samples QCD, ggF and VBF for all of the variables of interest in the boosted HH analysis.

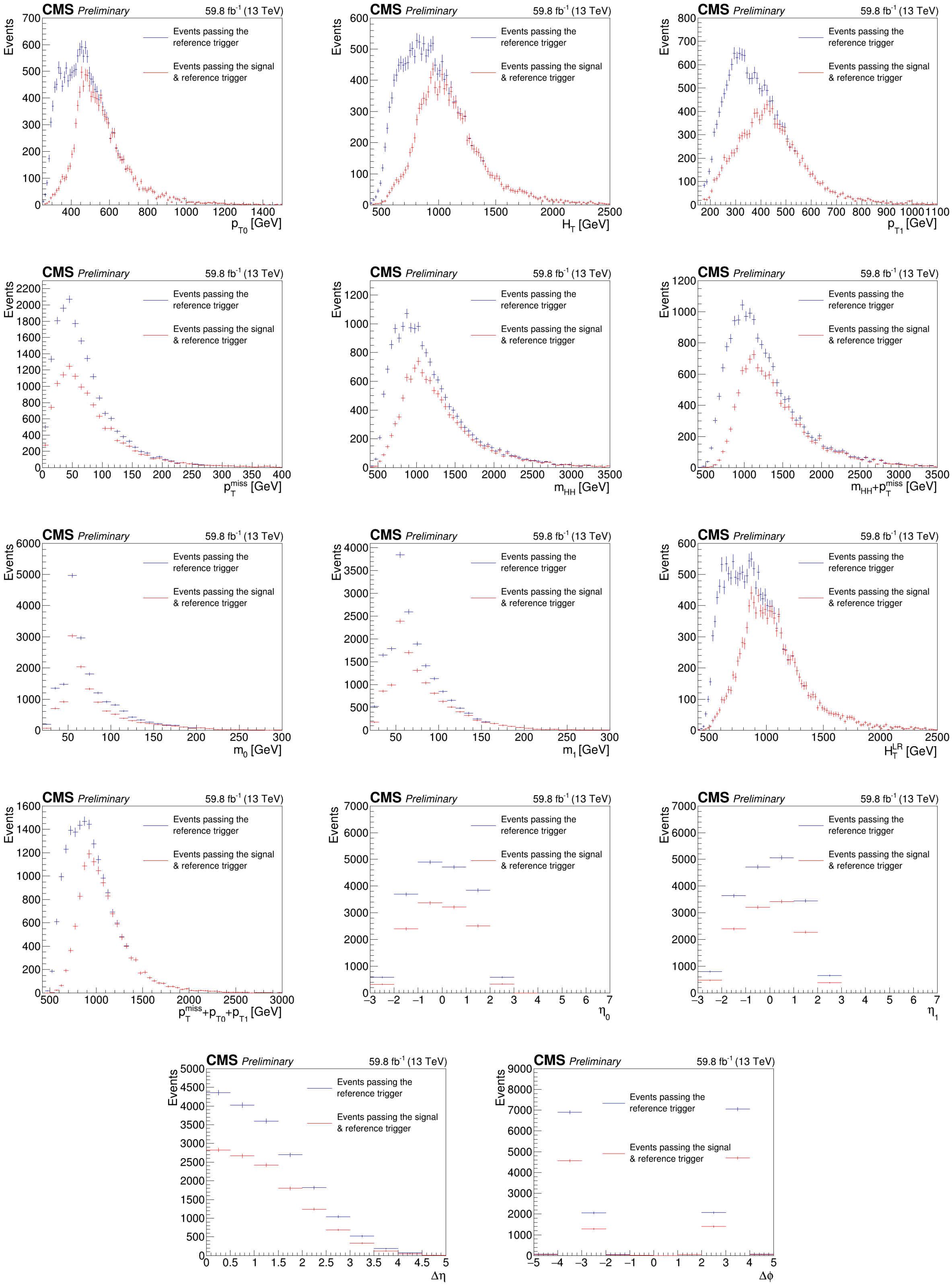


Figure 14: The distribution of events passing the reference trigger and events passing both the signal and reference trigger in the data sample for all of the variables of interest in the boosted HH analysis.

a combination of individual efficiencies of several HLT trigger paths. As mentioned earlier, the total trigger efficiency can be improved when we use a combination of different trigger algorithms, since this increases the number of events we are able to use in the analysis. For the boosted HH case, we use a combination of trigger paths targeting the events with high H_T, jet p_T, p_T^{miss}, mass and H_T^{LR} values. The signal triggers we use are

- HLT_AK8PFHT800_TrimMass50
- HLT_AK8PFJet400_TrimMass30
- HLT_AK8PFJet500
- HLT_PFJet500
- HLT_PFHT1050
- HLT_PFHT500_PFMET100_PFMHT100_IDTight
- HLT_PFHT700_PFMET85_PFMHT85_IDTight
- HLT_PFHT800_PFMET75_PFMHT75_IDTight

We will use an or-condition of all of these trigger paths to filter the events, which means that we will target events that pass any of these trigger paths. The trigger paths with the condition "AK8" use large-radius jets, i.e. jets with a radius of 0.8. The first trigger path thus targets events with large-radius jets and with a H_T^{LR} of over 800 GeV, and at least one large-radius jet with mass of over 50 GeV. The second trigger sets conditions for the p_T value, and demands that the event must contain at least one large-radius jet with a p_T value of over 400 GeV and a mass value of over 30 GeV. The third trigger just demands that the event contains a jet with a p_T value of over 500 GeV, and the fourth trigger that the event has an H_T value of over 1050 GeV.

The last three trigger paths have conditions for the H_T, p_T^{miss} and missing H_T, which is the same as p_T^{miss}, but calculated using small-radius jets instead of individual particles. The first trigger path demands an H_T value of over 500 GeV, and p_T^{miss} and missing H_T values of over 100, the second an H_T value of over 700 GeV, and p_T^{miss} and missing H_T over 85 GeV, and the last one demands an H_T value of over 800 GeV and p_T^{miss} and missing H_T of over 75 GeV. These triggers also require tight identification criteria for the reconstructed jets in the event, as the "IDTight" condition indicates.

6.1.2 Charged Higgs boson analysis

In the charged Higgs boson analysis we have 11 variables which we use as inputs for our algorithm. The variable names and explanations are presented in Table II. The distributions of all events and events passing the signal trigger as a function of these variables in the simulated samples are again presented in Figure 15. From these distributions we can observe that the TTbar and M200 samples have more similar distributions across several variables than the TTbar and M1000 samples.

The distributions of events in the data sample for the charged Higgs boson case are presented in Figure 16. Again, we can notice that the distributions are similar to the distributions of the TTbar sample.

The signal trigger we use in this analysis is HLT_MediumChargedIsoPFTau50_Trk30_eta2p1_1pr_MET100. The signal trigger sets conditions for the τ leptons in the event, and demands that the event has at least one τ lepton, which has a p_T value of over 50 GeV ("Tau50"), and there cannot be many charged hadrons near the τ lepton, as set by the "MediumChargedIso" -condition. The trigger also sets conditions for the decay of the τ lepton, and demands that it is a one-prong decay ("1pr") and the p_T value of the charged pion, which is the leading track in this case, must be at least 30 GeV ("Trk30"). Finally,

Table II: The names and explanations of the variables used in the charged Higgs boson analysis case.

Name	Explanation
$p_T(\tau_h)$	The p_T of the hadronic τ lepton having the highest p_T in the event (τ_h)
$\eta(\tau_h)$	The eta direction of τ_h
$N(\tau_h)$	Number of selected hadronic τ leptons in the event
$p_T(\pi_{\tau h})$	The p_T of the leading track pion associated with the τ_h
R_τ	The ratio between $p_T(\pi_{\tau h})$ and $p_T(\tau_h)$
p_T^0	The p_T of the jet having the highest p_T in the event
p_T^1	The p_T of the jet having the 2nd highest p_T in the event
p_T^2	The p_T of the jet having the 3rd highest p_T in the event
$N(jet)$	The number of selected jets in the event
m_T	Transverse mass of the τ and p_T^{miss} system
p_T^{miss}	Missing transverse momentum

the trigger also demands that the τ lepton is found within $\eta < 2.1$ ("eta2p1") and the p_T^{miss} is at least 100 GeV.

6.2 Developing the method using simulated samples

First, we develop the method using simulated samples. We use the simulated samples to validate the method, since with them we are able to compare the results given by this new method to the true trigger efficiency calculated using all the events in the sample. The main goal of the method is to measure the trigger efficiencies in data. For the boosted HH case, we have three simulated samples which we use: QCD, which simulates the multijet production due to quantum chromodynamics as the background sample, VBF, which simulates the process of vector boson fusion as the signal sample and ggF, which simulates the gluon fusion as another signal

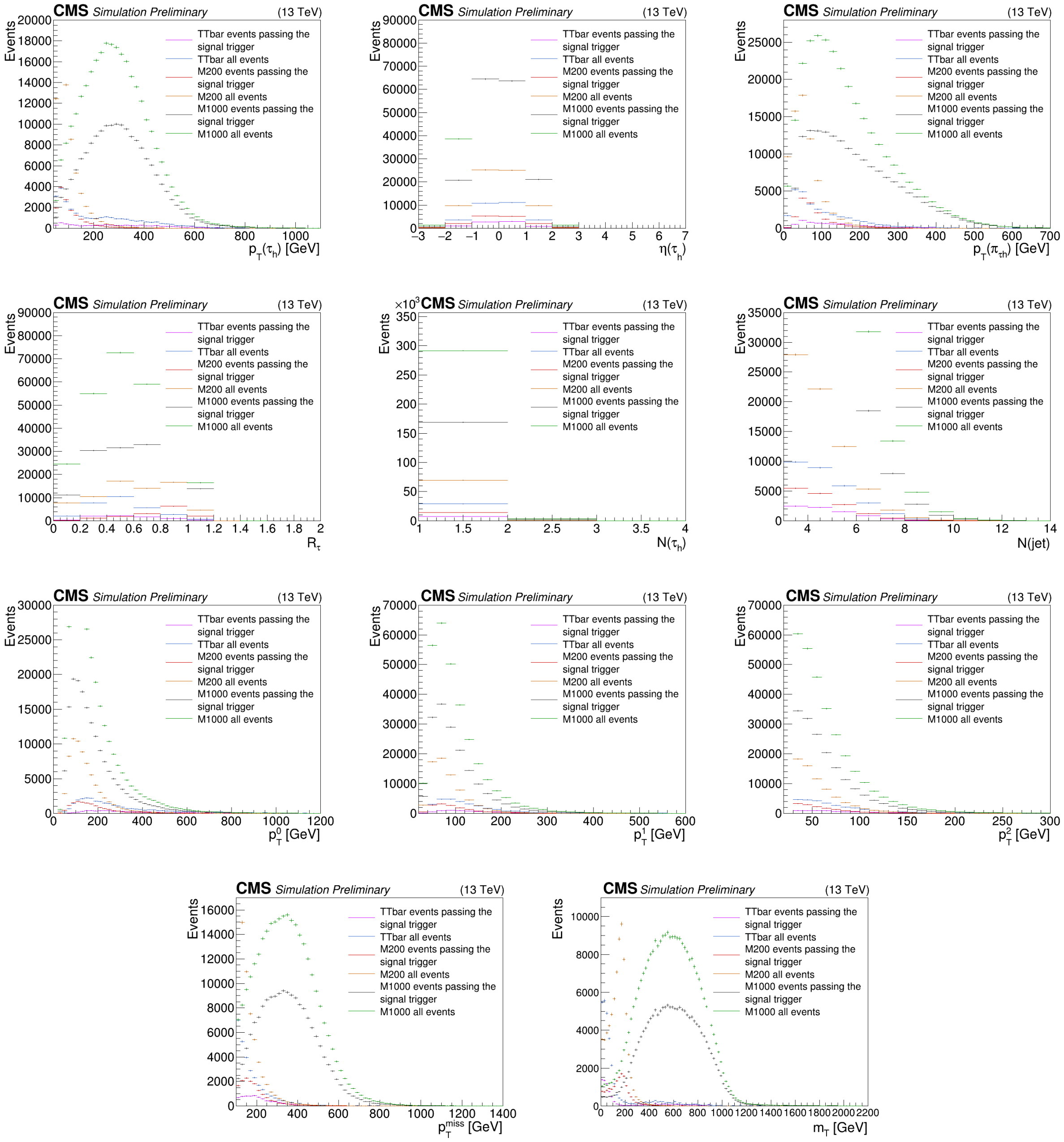


Figure 15: The distribution of all events and events passing the signal trigger in simulation samples TTbar, M200 and M1000 for all of the variables of interest in the charged Higgs boson analysis.

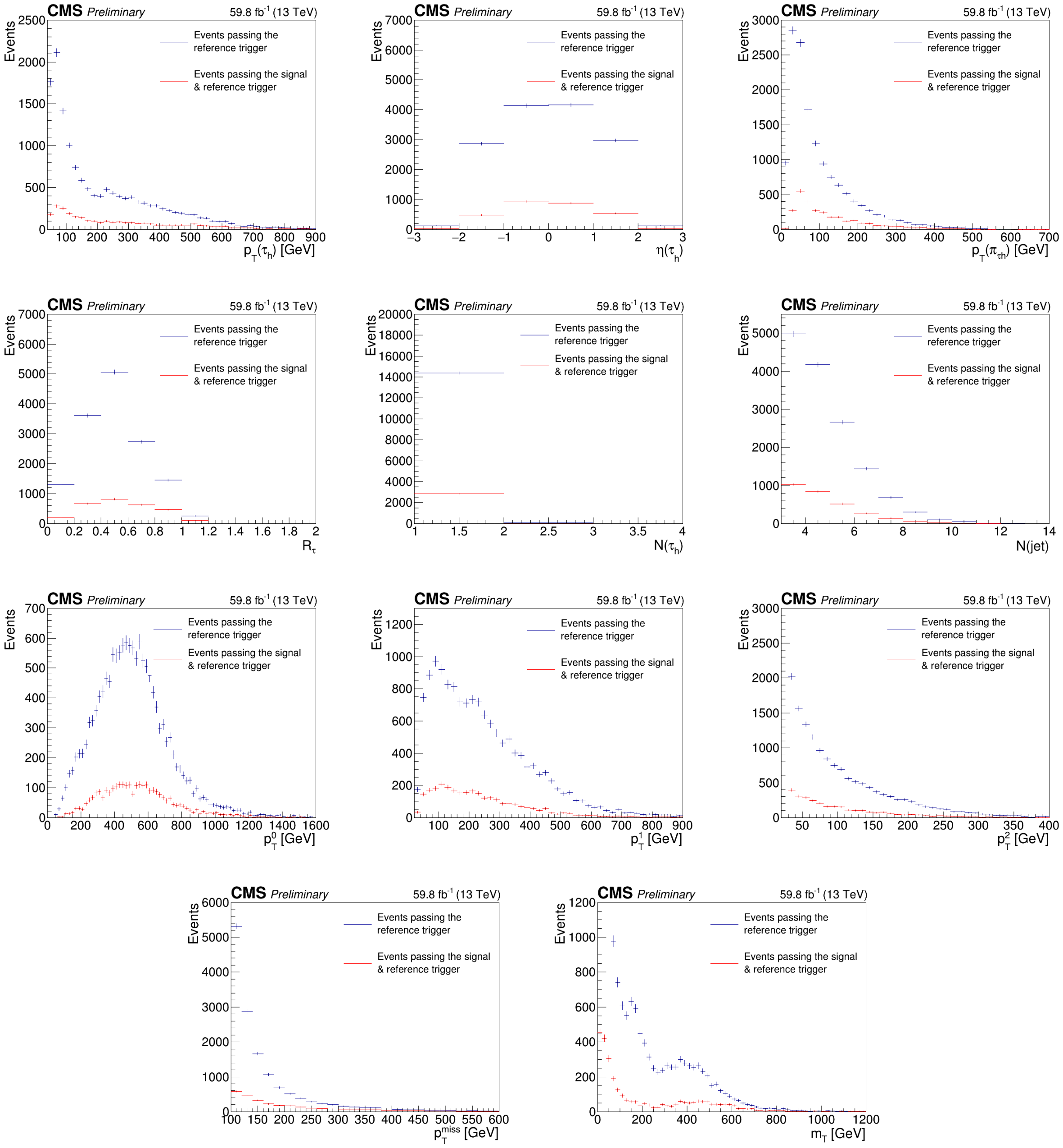


Figure 16: The distribution of events passing the reference trigger and events passing both the signal and reference trigger in the data sample for all of the variables of interest in the charged Higgs boson analysis.

sample. We use both the background sample and the signal samples in developing the method, since the data used in the actual boosted HH analysis consists of mostly QCD processes, but there might also be a small portion of signal processes, which the analysis aims to extract. With simulated samples, we are able to estimate the efficiencies in the cases where we would have only QCD events, or only signal events either produced via ggF or VBF. If we are able to confirm that the new method performs well in all of these cases, we can conclude it to perform well in also the case of the data being a mixture of all of these cases.

For the charged Higgs boson analysis case, we have a background sample TTbar, which simulates the $\mathrm{t\bar{t}}$ production, and signal samples M200 and M1000, which both simulate the production of a heavy charged Higgs boson. The M200 simulates the case where the H^+ mass is 200 GeV, and the M1000 simulates the case where the H^+ mass is 1000 GeV. The idea is the same as explained for the boosted HH analysis, since in this case we also validate the method using simulation samples for both the background and signal processes to make sure the method would be able to perform well in the analysis performed using data.

6.2.1 Choosing the reference trigger

As the first step, we have to select the right reference trigger with the help of simulated samples, as explained in Section 4. We have to select a reference trigger which gives good results in all of the three simulated samples and with all of the variables we are using as inputs for the gradient boosting model. For the boosted HH analysis case, we tried several triggers and trigger combinations to compare the results with the true efficiency produced when taking into account all the events in the simulated samples. The trigger which was chosen to be the reference trigger in this case was HLT_AK8PFJet260, meaning that the reference trigger in question targets events which have at least one large-radius jet with a transverse momentum of over 260 GeV.

Some examples of the trigger efficiencies in all of the HH samples are presented in Figure 17. The trigger efficiency is presented as a function of p_{T0}, H_T and p_T^{miss}. In the plots, the efficiency calculated using all the events in the simulated sample ("true efficiency") is compared to the efficiency measured using the reference trigger ("measured efficiency"). As seen in the plots, the efficiency from all the events and the measured efficiency using the reference trigger match quite accurately across the variables of interest.

As we can also see when we compare the trigger efficiencies between different samples, the trigger efficiencies as a function of the variables of interest differ between the background and signal samples, and also between the two signal samples. Therefore it is important to ensure that the reference trigger performs well across all of these samples. From these plots we can see that the reference trigger is indeed able to perform well in both the background and signal samples. The reference trigger we choose in this stage of developing the method with simulated samples is later used in the data sample, and the example trigger efficiency curves are presented also for the data sample in Figure 17.

In the charged Higgs boson case, we have a combination of reference trigger paths to maximize the amount of events we are able to use in the data sample, and to get a good correspondence between the efficiency calculated using all the events and the efficiency calculated using these reference triggers in the simulated samples. We will again use an or-condition of these triggers to acquire the events. As reference triggers we will use triggers which target the H_T values of the event: HLT_PFHT180, HLT_PFHT250, HLT_PFHT350, HLT_PFHT370, HLT_PFHT430, HLT_PFHT680, HLT_PFHT780, HLT_PFHT890 and HLT_PFHT1050. Again, examples of the trigger efficiencies in the charged Higgs boson analysis are presented in Figure 18. The trigger efficiency is presented as a

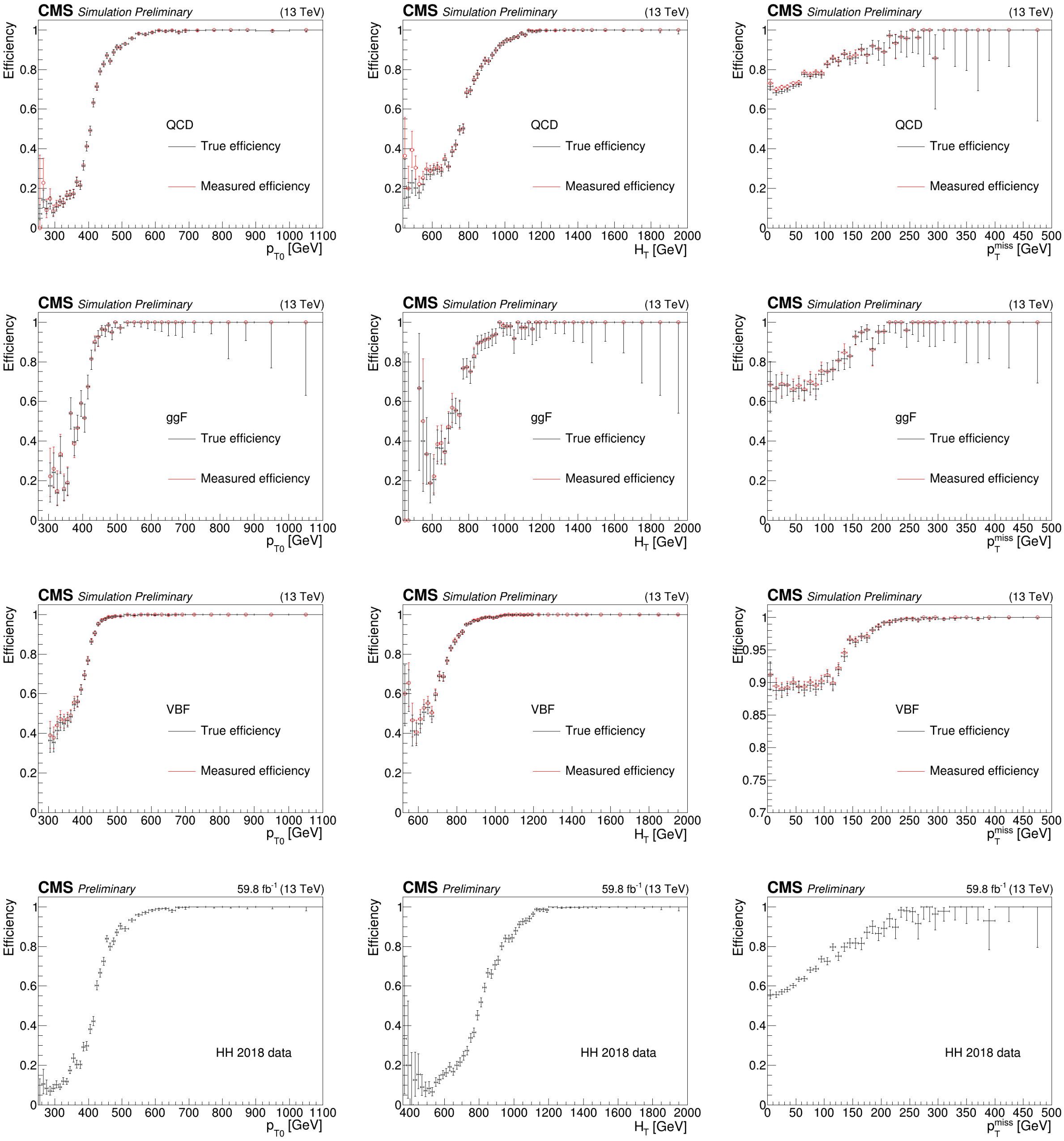


Figure 17: Trigger efficiency as a function of p_{T0}, H_T and p_T^{miss} in the QCD, ggF, VBF and data samples in the boosted HH analysis.

function of p_T^0, p_T^{miss} and m_T in the TTbar, M200, M1000 simulation samples and in the data sample.

6.2.2 Training the algorithm

When we have chosen the reference trigger, we can start to train the gradient boosting algorithm. For training the model, we use only the events which pass the reference trigger to simulate the situation with actual data, where we don't have access to all of the events.

The model is trained using only the background sample, which is the QCD sample in the case of the boosted HH analysis. This also simulates the situation with actual data. We resampled the original QCD sample to contain the same amount of events as the data sample, which in this case corresponds to 18324 events. This is done to make sure the model training process using the QCD sample is comparable with the training process when we use the data sample. The events are chosen as a function of the H_T distribution of the data sample to ensure that the resulting QCD sample represents the data sample well. We use 80% of the events in the sample to train the model, and the model is tested with 10% and validated with the remaining 10%. This means that in the training sample we have 14659 events, in the testing sample 1833 events, and in the validation sample 1832 events. The completed model is used to the whole QCD sample as well as to the signal samples VBF and ggF, so we can determine how well the model is able to generalize into different processes. Since the VBF and ggF samples simulate processes where a Higgs boson pair can be produced, it is important to make sure that the machine learning method is valid when estimating the trigger efficiency in these processes as well.

Similarly, the TTbar sample is used to train the model in the charged Higgs boson analysis case, and the model is used to the signal samples M200 and M1000 to test how well it is able to generalize into these processes. The distribution of

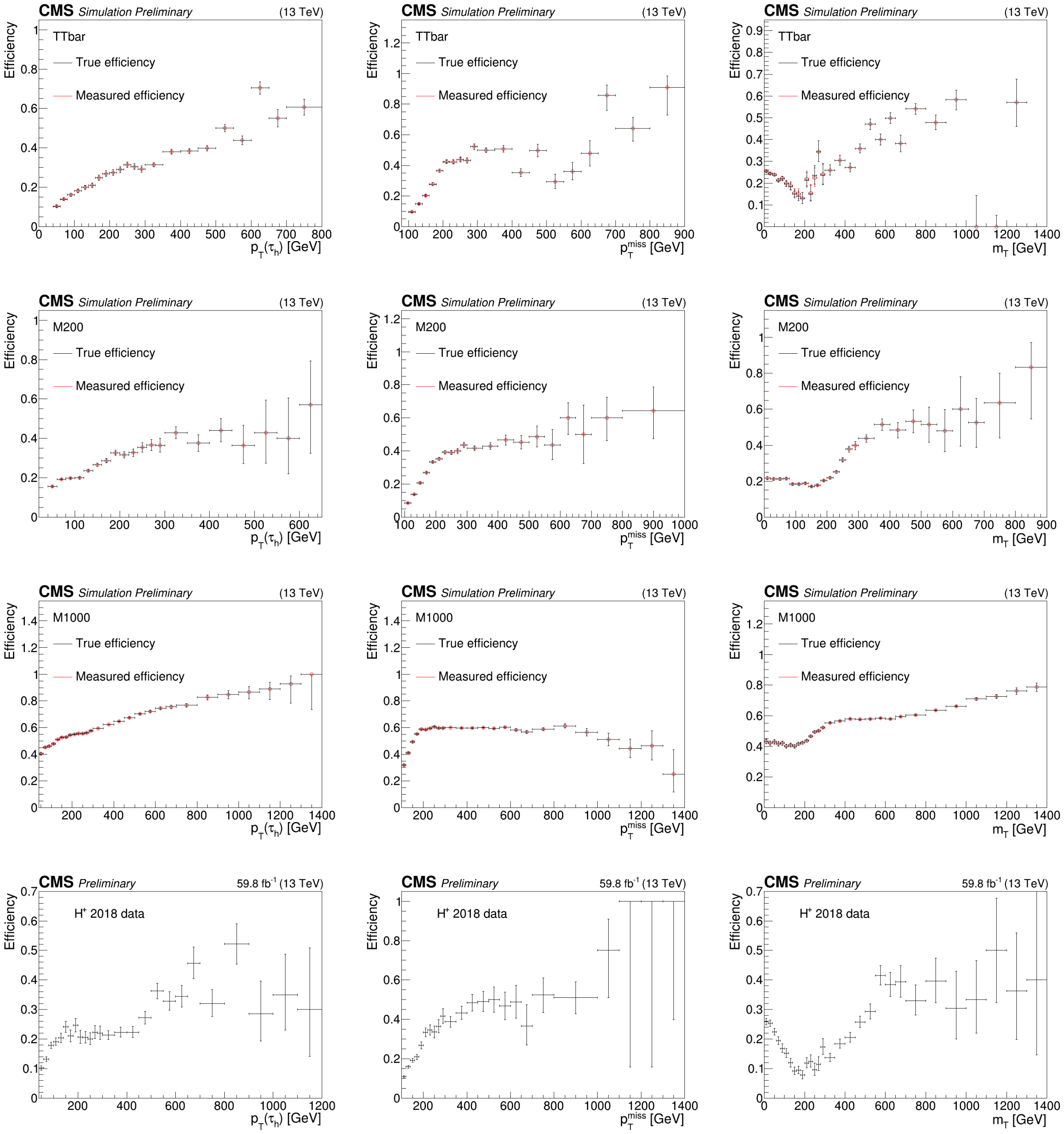


Figure 18: Trigger efficiency as a function of p_{T0}, p_T^{miss} and m_T in the TTbar, M200, M1000 and data samples in the charged Higgs boson analysis.

events used for training, testing and validating is the same as in the boosted HH case: 80% is used for training, 10% for testing and 10% for validating. In this case we have in total 14431 events, and they are chosen from the original TTbar sample as a function of the m_T distribution of the data sample. In the training sample we have 11544 events, in the testing sample 1444 events, and in the validation sample 1443 events.

6.3 Developing the method using data

After the model has been initially trained and tested with the simulated data, we can proceed to training the model using actual data. The training, testing and validation steps of the method are the same as with the simulated samples: in the boosted HH analysis we have a data sample consisting of 18324 events passing the reference trigger, and we train the model using a training set consisting of 80% of the events, test with 10% and validate the results with the remaining 10%. The trigger efficiency as a function of p_{T0}, H_T and p_T^{miss} for the boosted HH data sample is presented in Figure 17, where it can be compared to the efficiencies in the simulated samples.

In the charged Higgs boson analysis we have 14431 events passing the reference trigger, and the training set is again 80% of the events, the testing set 10% of the events and the validation set consists of 10% of the events. The trigger efficiency as a function of p_T^0, p_T^{miss} and m_T for the data sample in the charged Higgs boson analysis is presented in Figure 18.

6.4 Aleatoric uncertainties

Aleatoric uncertainty, also known as statistical uncertainty refers to the uncertainty that is caused by inherently random effects in the experiment. In our case, this would refer to the limited training sample, meaning that the events chosen for the training sample are picked randomly, but the choice might affect the outcome of the model. To estimate these aleatoric uncertainties emerging from the limited training sample we will use a method called bootstrapping. Bootstrapping is a method where we resample the original data set with replacement, meaning that the same data point can be selected into multiple new samples. Each of these new samples are used to train the model consequently, and we can then compare the predictions each of these models produces for each data point by calculating the mean and the standard deviation between the different predictions. [47] We can use these metrics to estimate the aleatoric uncertainties related to this method, when we compare the mean of the results obtained from all of these models to the traditional trigger efficiency measurement methods, and the standard deviation corresponds to the uncertainty.

In both the boosted HH analysis case and the charged Higgs boson analysis case, the number of models trained with sub-samples from the original data was 20. Since the sub-samples are sampled from the original sample with replacement, all the sub-samples have the same amount of events as the original sample. The amount of sub-samples being 20 ensures that we get enough results to construct a good estimate of the standard deviation and mean depending on the different combinations of training data, but the bootstrapping isn't too time-consuming in this stage of developing the method.

6.5 Epistemic uncertainties

Besides the aleatoric uncertainties related to the limited training sample, we have to take into account the epistemic uncertainties, which in principle mean uncertainties that are caused by a lack of knowledge. In this method, this type of uncertainty is related to the choice of the model architecture, meaning that we have to determine whether the choice of which architecture to use in this method causes differences in the outcomes that are larger than the uncertainties defined with the bootstrapping method. To test this, we will train another model using a DNN architecture presented in Section 5.3, and compare the results this model produces to the previous results.

7 Results

In this chapter, we will present the results for both example cases: the boosted HH analysis case and the charged Higgs boson analysis case. We will include the results for both cases obtained from both the model trained using the simulated background sample and the model trained with the data sample. Curves containing the gradient boosting predicted efficiencies and the aleatoric uncertainties are presented for both analyses. The uncertainties related to the model architecture are studied only with the boosted HH analysis case, since we can expect them to be similar in the charged Higgs boson analysis case. We will also present results when analyzing the model's performance as a function of two variables simultaneously for both analyses. Lastly, we will present the results which focus on the model's performance using the metrics presented in Section 5.4.

7.1 Boosted HH results

In this section, we will present the results obtained when developing the machine learning method using simulation and data samples presented in Section 1.3.1 and Section 6.2, related to the boosted HH analysis.

7.1.1 Trigger efficiency curves with aleatoric uncertainties

We will present the gradient boosting model predicted efficiencies as a function of all of the variables of interest presented in Section 6.1.1. The results obtained from the QCD sample are shown in Figure 19. The gradient boosting predicted efficiencies and their corresponding aleatoric uncertainties obtained with the bootstrapping method are laid on top of the efficiencies calculated with the reference trigger method as well as the true efficiencies. This is done to compare the gradient boosting predicted efficiencies to the efficiencies measured using a traditional reference trigger method. As seen from the plots, the gradient boosting predicted efficiencies agree quite well with the efficiencies measured using the reference trigger method as well as with the true efficiency in this sample. This is reasonable, since the model is trained using the QCD sample, so it is able to recognize the efficiencies related to this process well.

Next, we will present the results obtained from using the model trained with the QCD sample to predict the trigger efficiency in the signal samples. It's important to see how well the gradient boosting model is able to generalize into processes that it has not been trained with, and as explained before, the data used in the actual boosted HH analysis targets these signal events. We start with the ggF sample. The gradient boosting predicted trigger efficiencies and their aleatoric uncertainties as a function of the variables of interest for the ggF sample are presented in Figure 20, laid on top of the efficiencies obtained using the reference trigger method and the true efficiencies. When we look at these results obtained with the ggF sample, it can be seen that the model is able to predict the efficiencies in this sample reasonably well, and the bootstrapping uncertainties cover the differences between the efficiencies obtained with the reference trigger method, the true efficiencies and the gradient boosting predicted efficiencies.

The results obtained using the QCD trained model in the VBF sample are presented in Figure 21. We can see here that the model is able to generalize also into the vector boson fusion process simulated by the VBF sample, and the differences between the efficiencies calculated using the traditional method, the true efficiencies and the efficiencies from the gradient boosting method are covered by the uncertainties obtained by bootstrapping.

Finally, we will use the method for the data sample. The efficiencies obtained using the model trained with the data sample are presented in Figure 22. The gradient boosting predicted efficiencies are now compared to the efficiencies measured directly from data. From these results we can observe that the gradient boosting

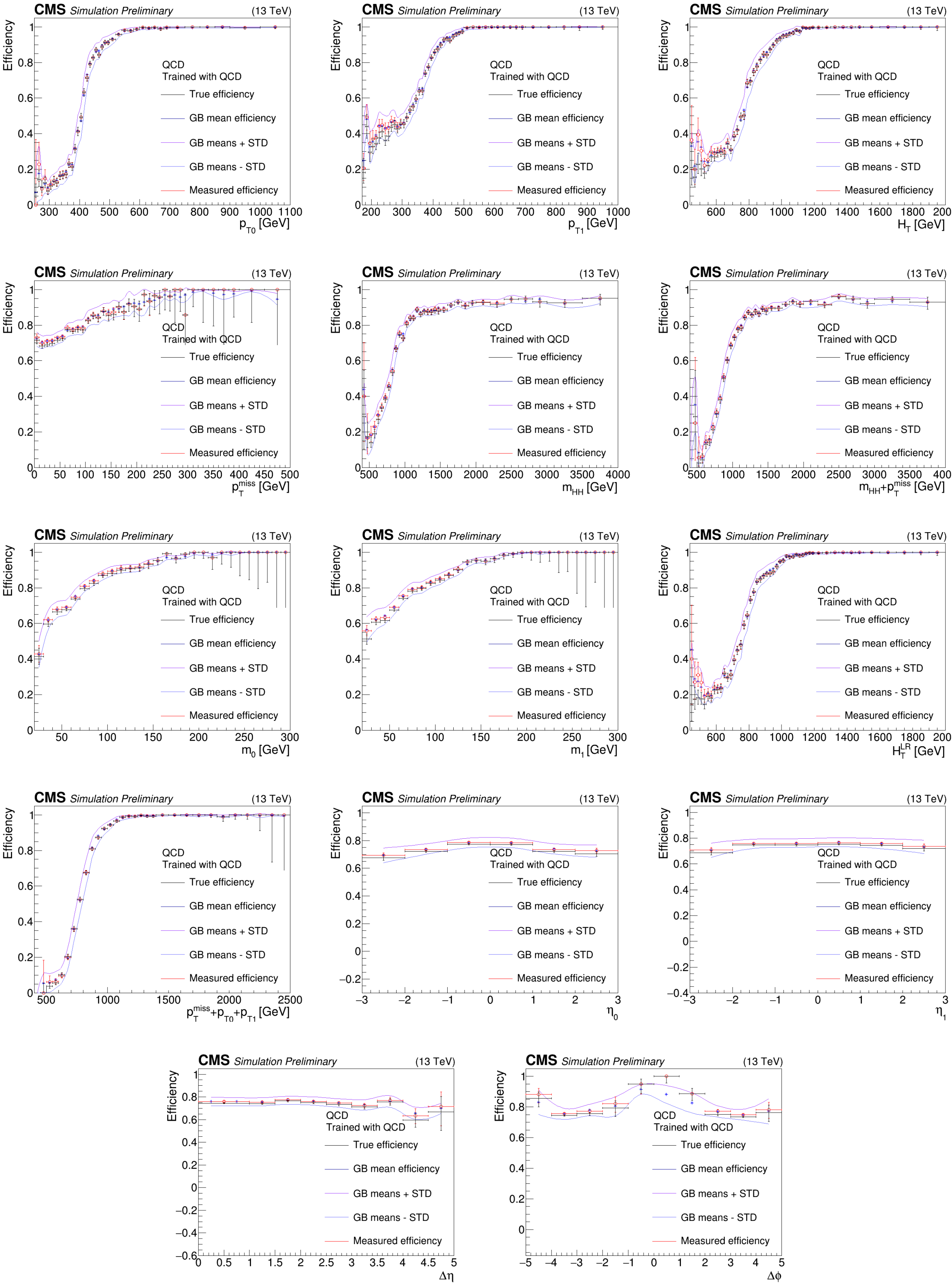


Figure 19: Trigger efficiency in the boosted HH analysis as a function of variables of interest in the QCD sample, containing the true efficiencies, the reference trigger measured efficiencies, the efficiencies predicted using the QCD trained gradient boosting (GB) model, and the corresponding aleatoric uncertainties, which are the standard deviations (STD) obtained with the bootstrapping method.

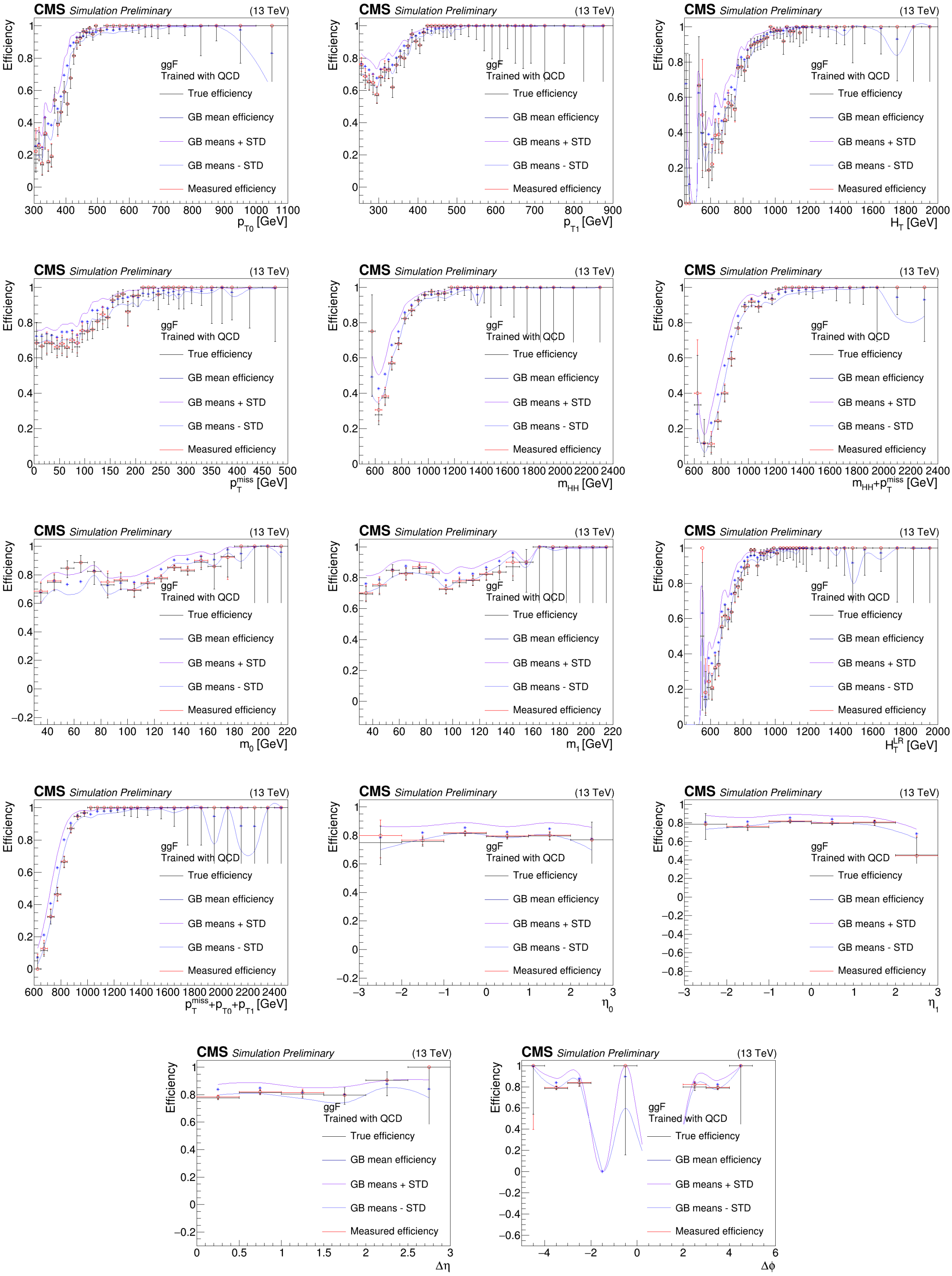


Figure 20: Trigger efficiency in the boosted HH analysis as a function of variables of interest in the ggF sample, containing the true efficiencies, the reference trigger measured efficiencies, the efficiencies predicted using the QCD trained gradient boosting (GB) model, and the corresponding aleatoric uncertainties, which are the standard deviations (STD) obtained with the bootstrapping method.

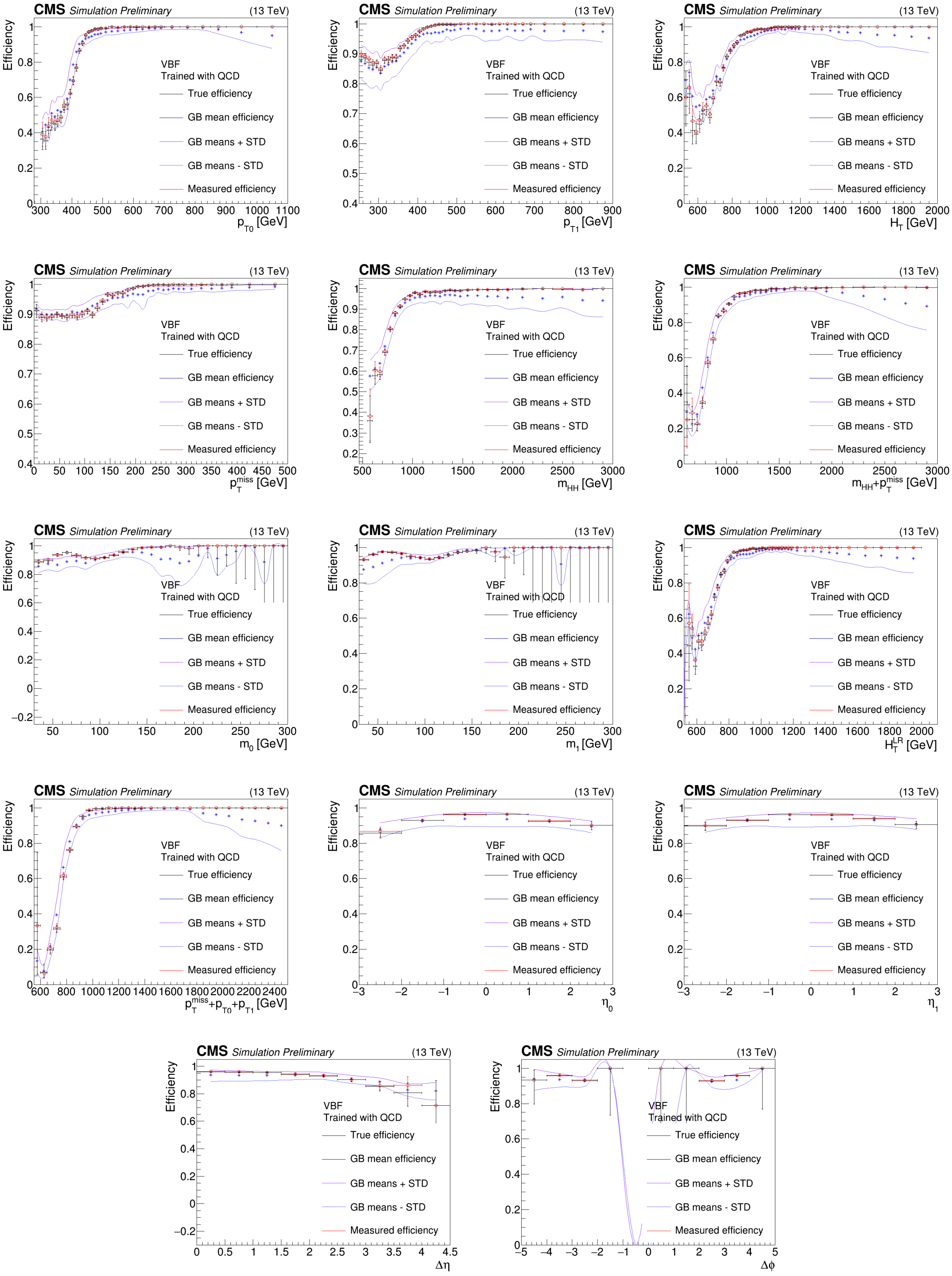


Figure 21: Trigger efficiency in the boosted HH analysis as a function of variables of interest in the VBF sample, containing the true efficiencies, the reference trigger measured efficiencies, the efficiencies predicted using the QCD trained gradient boosting (GB) model, and the corresponding aleatoric uncertainties, which are the standard deviations (STD) obtained with the bootstrapping method.

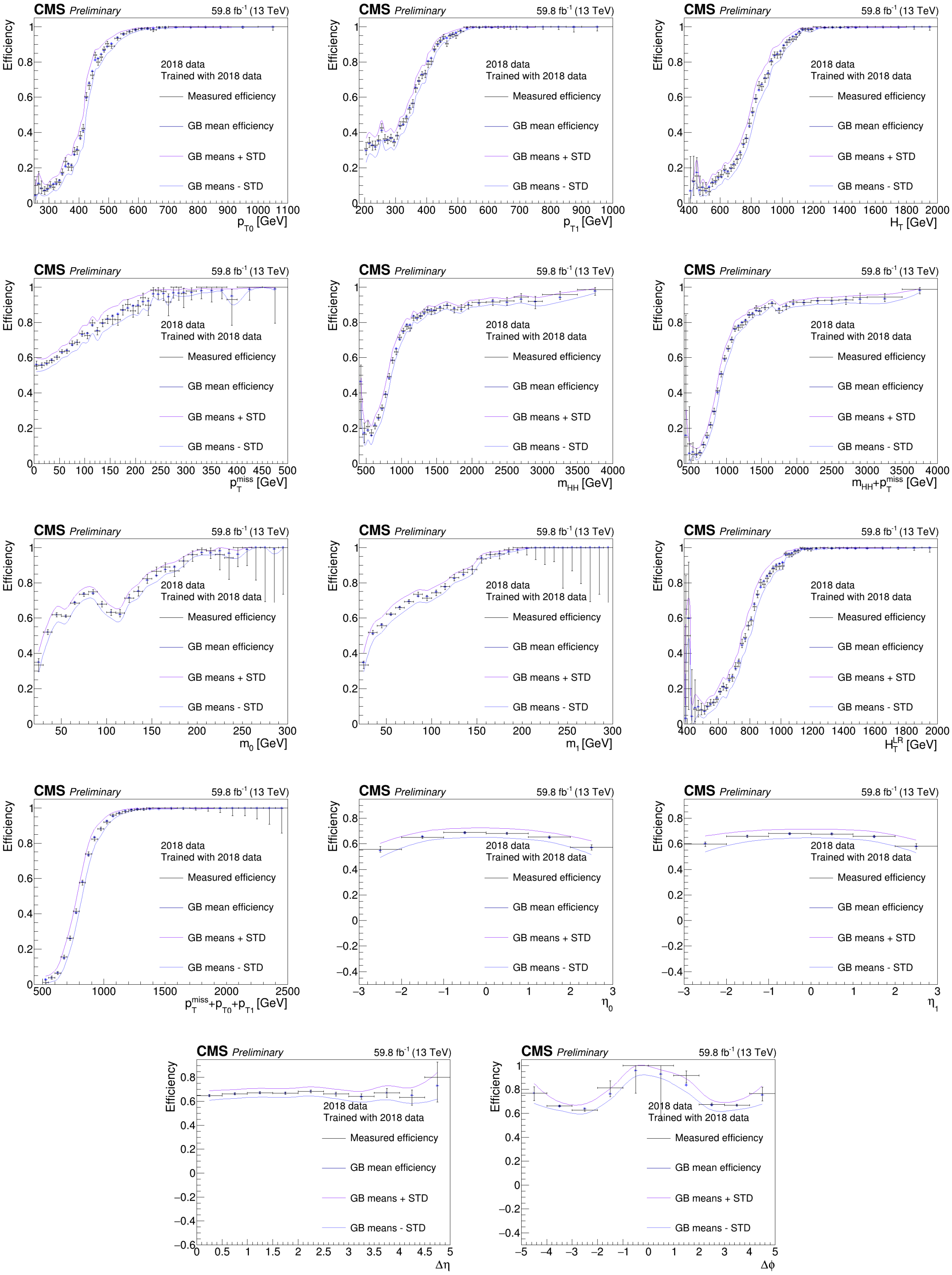


Figure 22: Trigger efficiency in the boosted HH analysis as a function of variables of interest in the data sample, containing the reference trigger measured efficiencies, the efficiencies predicted using the gradient boosting (GB) model trained with the data sample, and the corresponding aleatoric uncertainties, which are the standard deviations (STD) obtained with the bootstrapping method.

predicted efficiencies agree well with the efficiencies measured from data, meaning that the model is able to perform well in estimating the efficiencies in the data sample, which is the primary objective of the new method.

7.1.2 Epistemic uncertainties

The results containing the uncertainties related to the model architecture are presented and compared to the aleatoric uncertainties. First, the trigger efficiencies as a function of all of the variables of interest, containing the gradient boosting predicted efficiencies and their uncertainties, as well as the deep neural network predicted efficiencies for the QCD sample are presented in Figure 23. As seen in the plots, the DNN predicted efficiencies are compatible with the gradient boosting based prediction, i.e. lie within the aleatoric uncertainties defined previously. We can see the same effect in the plots containing the efficiencies measured in the data sample, which are presented in Figure 24. We can conclude that the choice of a model architecture does not cause uncertainties larger than the aleatoric uncertainties related to the limited training sample.

7.1.3 Trigger efficiency as a function of two variables simultaneously

To estimate the model's performance across multiple variables simultaneously, we can analyze the trigger efficiency as a function of one variable while varying another. In the boosted HH analysis, this is done by choosing events in certain H_T range and plotting the trigger efficiency as a function of these events' p_{T0}, and vice versa. The results from the data sample are presented in Figure 25, where there are three H_T ranges and the efficiency plotted as a function of p_{T0} in these ranges, as well as three p_{T0} ranges and the efficiency plotted as a function of H_T in these ranges. As we can conclude from these results, the gradient boosting predicted efficiencies align well

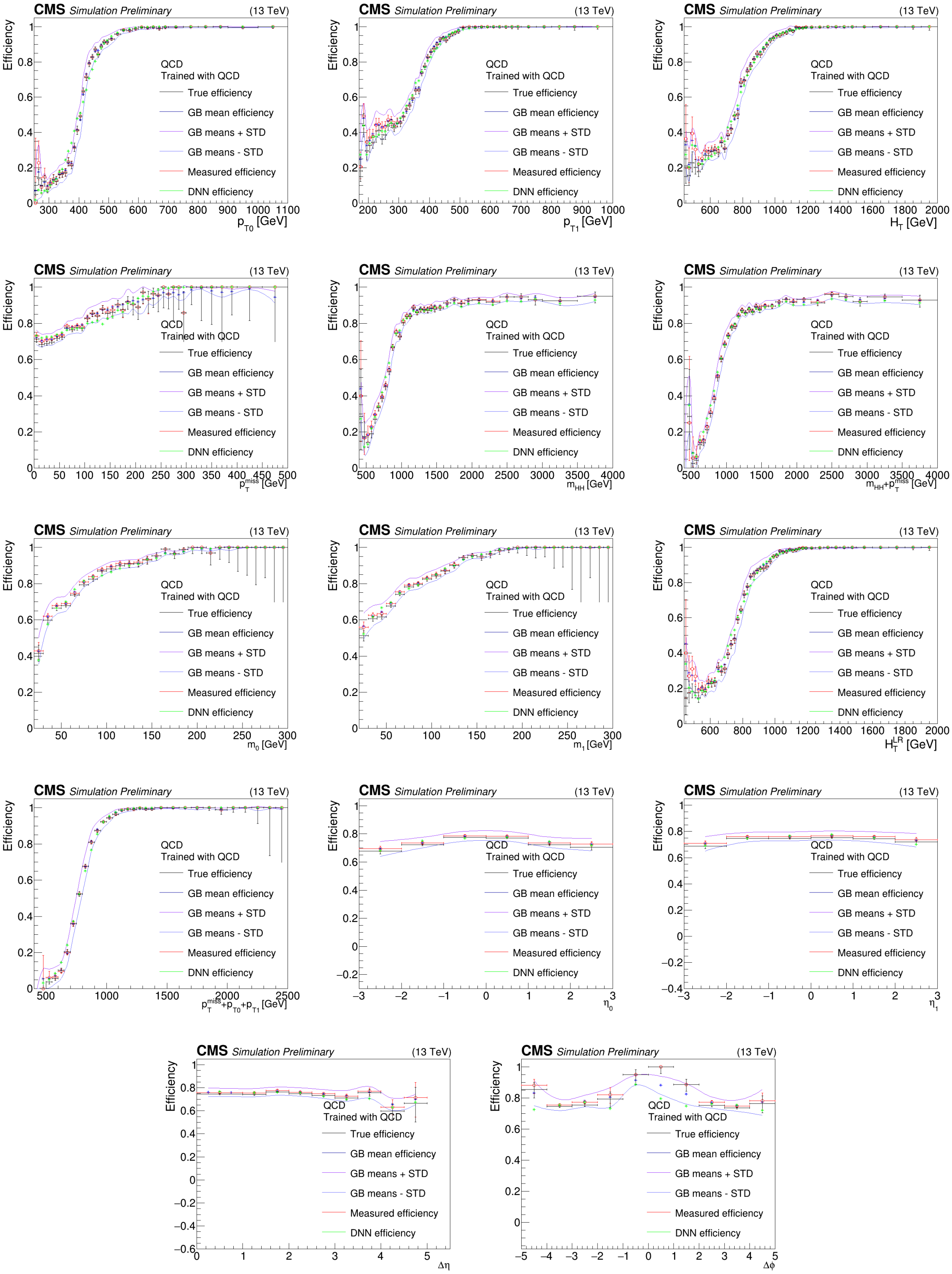


Figure 23: Trigger efficiency in the boosted HH analysis as a function of variables of interest in the QCD sample, containing the true efficiencies, the measured efficiencies, the efficiencies predicted using the QCD trained gradient boosting (GB) model, the corresponding aleatoric uncertainties, which are the standard deviations (STD) from the bootstrapping method, and the DNN predicted efficiencies.

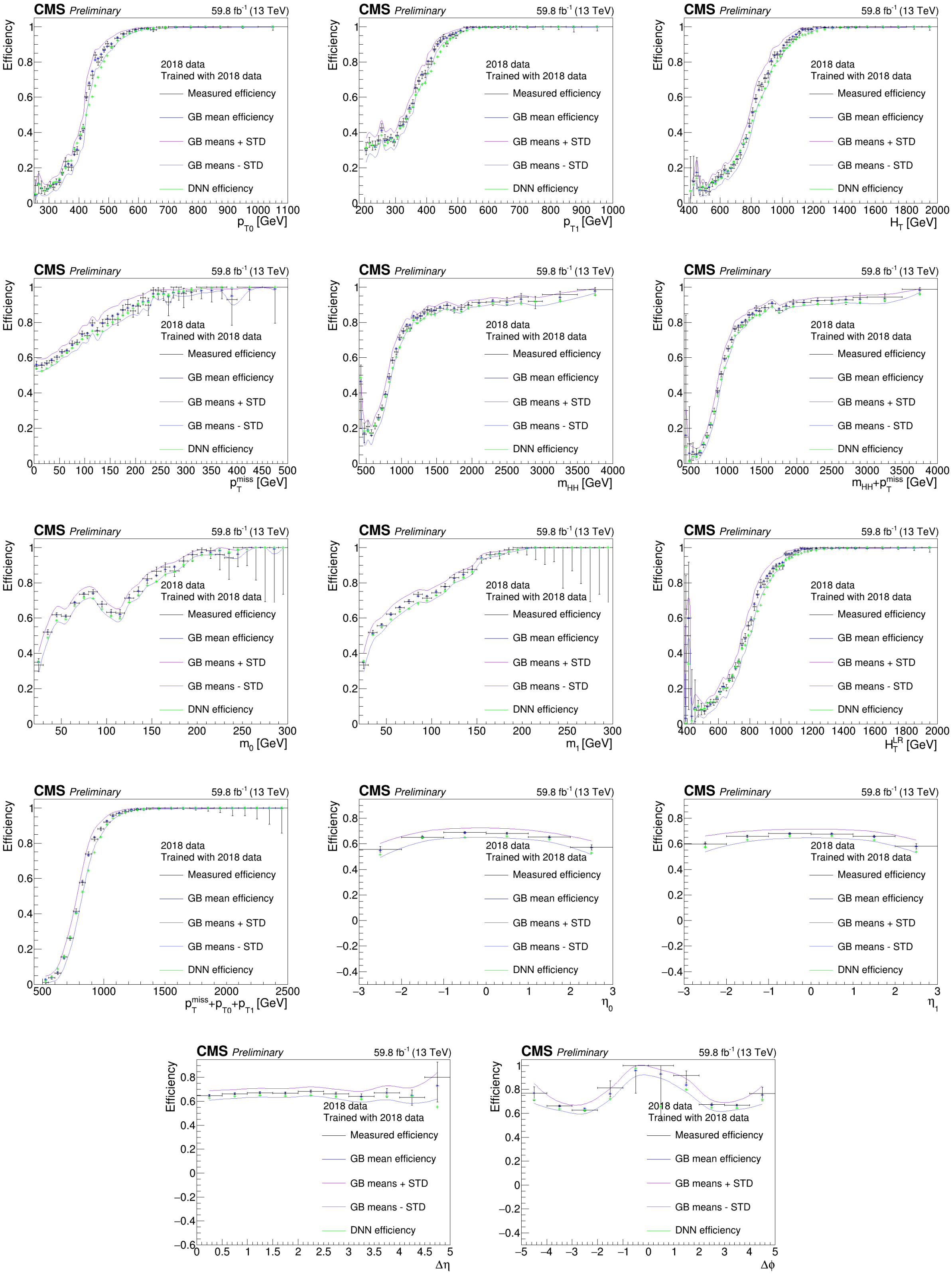


Figure 24: Trigger efficiency in the boosted HH analysis as a function of variables of interest in the data sample, containing the measured efficiencies, the efficiencies predicted using the gradient boosting (GB) model trained with data, the corresponding aleatoric uncertainties, which are the standard deviations (STD) from the bootstrapping method, and the DNN predicted efficiencies.

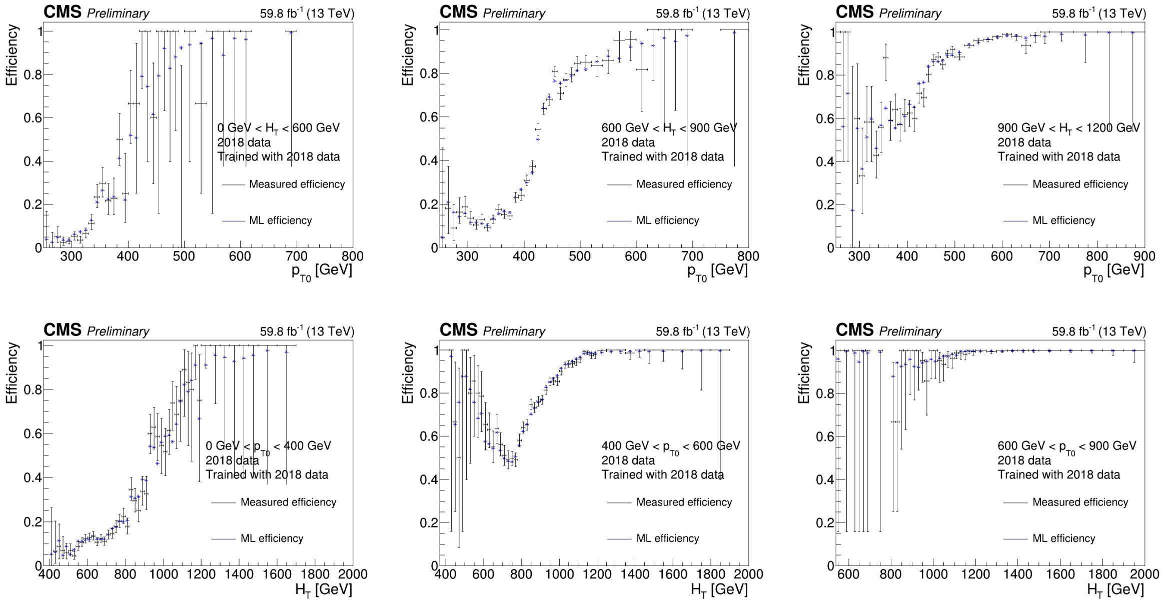


Figure 25: The trigger efficiency as a function of p_{T0} in three different H_T ranges (top row), and the trigger efficiency as a function of H_T in three different p_{T0} ranges (bottom row) in the boosted HH data sample.

with the measured efficiencies also in this case of examining the trigger efficiency as a function of several variables simultaneously.

7.1.4 Performance

The AUC score and accuracy for the model trained using the QCD sample and the model trained using the data sample are presented in Table III. As seen in the table, the accuracies are around 0.9 for both of the models, the accuracy of the model trained using QCD reaching a value of 0.910 in the QCD sample and the model trained using data having a slightly lower value of 0.896 in the data sample. The AUC scores are excellent for both models, the AUC score of the model trained with QCD being 0.959 and the AUC score trained with data being 0.958. The table also includes accuracies and AUC scores for the QCD trained model applied to the two signal samples ggF and VBF. The accuracy in the ggF sample is 0.854 and

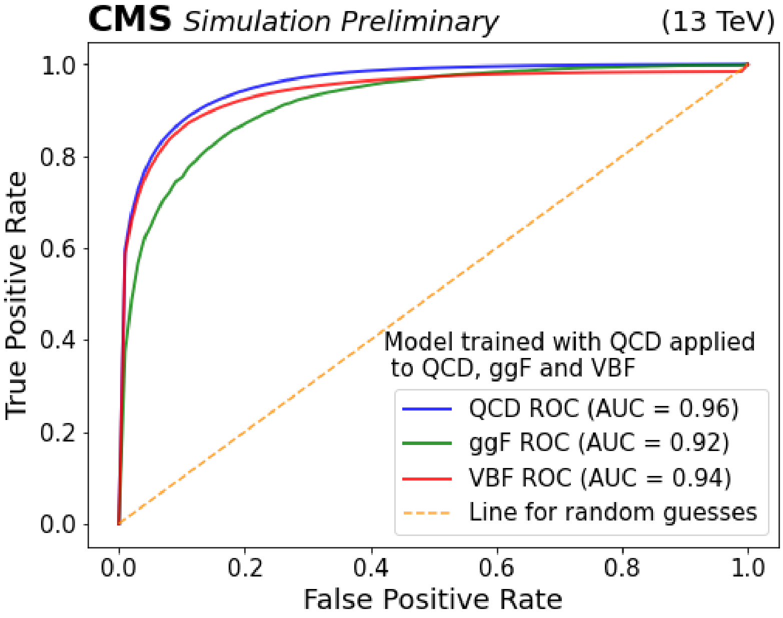


Figure 26: The ROC curves for the model trained using the QCD sample, applied to the QCD, ggF and VBF samples.

the AUC score is 0.919. For the VBF sample, the accuracy is 0.928 and the AUC score is 0.941. The model performs well in both of these samples, which means that the model is able to generalize well into unseen processes. We can also notice that the model is performing better with the VBF sample than with the ggF sample, and this indicates that the VBF sample has more similar properties with the QCD sample than the ggF has with the QCD sample, which we could also observe when comparing the distributions of these variables in the simulation samples.

The ROC curves for the model trained using the QCD sample and applied to all simulated samples are also illustrated in Figure 26, and for the model trained using data in Figure 27. From the ROC curves we can also observe that both models perform significantly better than a model which makes random guesses for all the samples.

The feature importance ranking introduced in Section 5.2 for the model trained using the QCD sample is presented in Figure 28. As seen in the figure, the p_{T0}

Table III: The accuracies and AUC scores in the boosted HH analysis for the model trained using the simulated QCD sample applied to the QCD, ggF and VBF samples, and the model trained with and applied to the data sample

	Accuracy	AUC score
Data	0.896	0.958
QCD	0.910	0.959
ggF	0.854	0.919
VBF	0.928	0.941

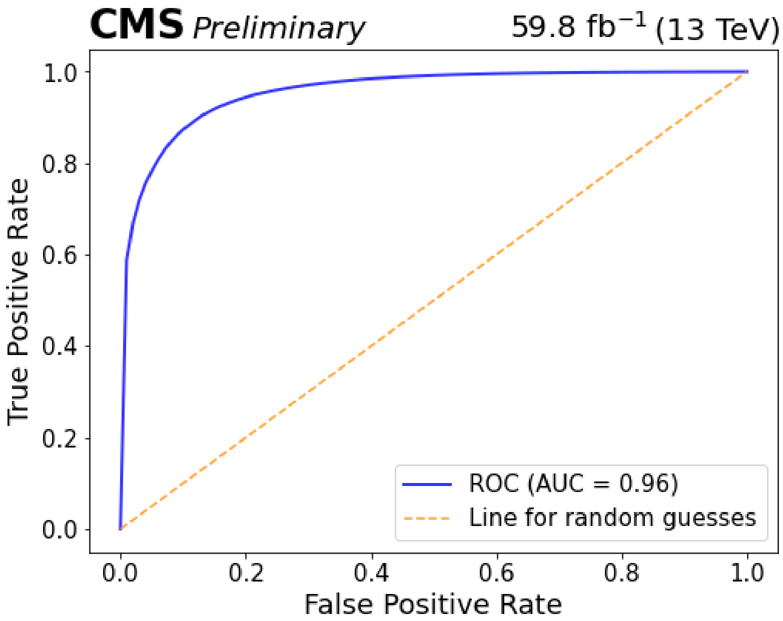


Figure 27: The ROC curve for the model trained using the data sample of the boosted HH analysis.

variable is clearly the most important feature that the model uses for making predictions. The importance value is calculated with respect to the importance values of all other features, meaning that the importance values of all features will sum up to 1. The importance value of the p_{T0} reaches over 0.6, meaning that compared for example to the second most important feature H_T with a feature importance value of around 0.1, it is more than six times more important in making the decisions than the H_T. This is reasonable, since the trigger paths we use for this analysis as the signal trigger combination largely target the large-radius p_T values of the event. The same phenomena is also observed for the model trained using the data sample. In this case, the p_{T0} variable is also the most important feature, with its importance being just around 0.6. The other variables which have some importance in the decision making in both models are the H_T, H_T^{LR}, p_T^{miss} and $p_T^{miss} + p_{T0} + p_{T1}$, which also make sense, since these are also variables targeted by the signal trigger combination.

A large number of variables do not seem to contribute to the decision significantly, and we tried also training the model using only the most important features p_{T0}, H_T, H_T^{LR}, p_T^{miss} and $p_T^{miss} + p_{T0} + p_{T1}$ and leaving out the other features to see if the model would perform as well with only these features. The results were similar for the QCD sample, but for the ggF and VBF samples the model's performance decreased. Thus, it seems that features which have a lower feature importance still have an effect in helping the model generalize better into unseen data.

7.2 Charged Higgs boson results

In this section, we will present the results obtained when developing the model using simulation and data samples presented in Section 1.4.1 and Section 6.2, related to the charged Higgs boson analysis. The variables used in this analysis are presented in Section 6.1.2.

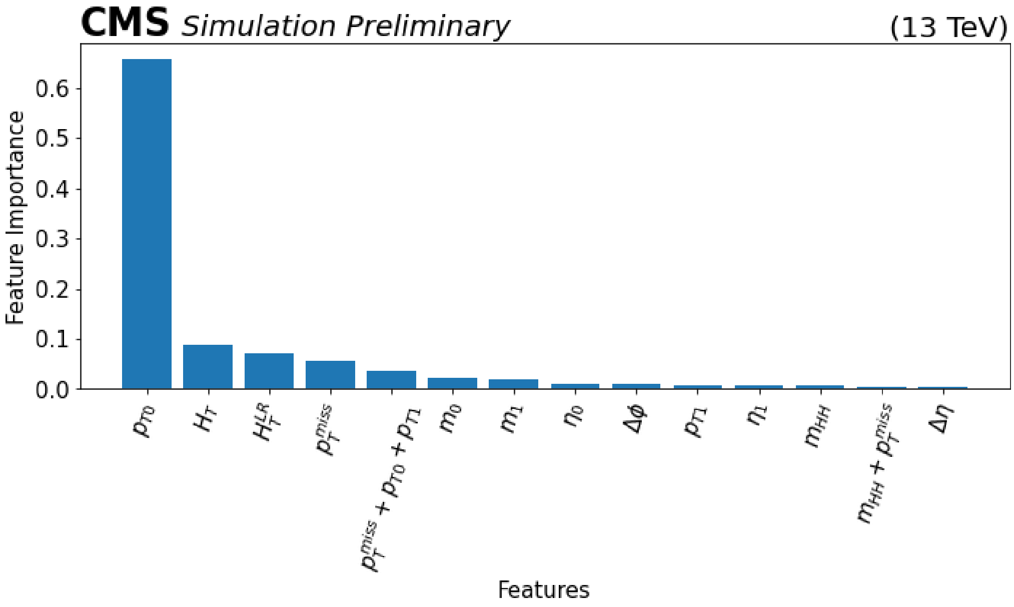


Figure 28: Feature importance ranking plot for the model trained using the QCD sample.

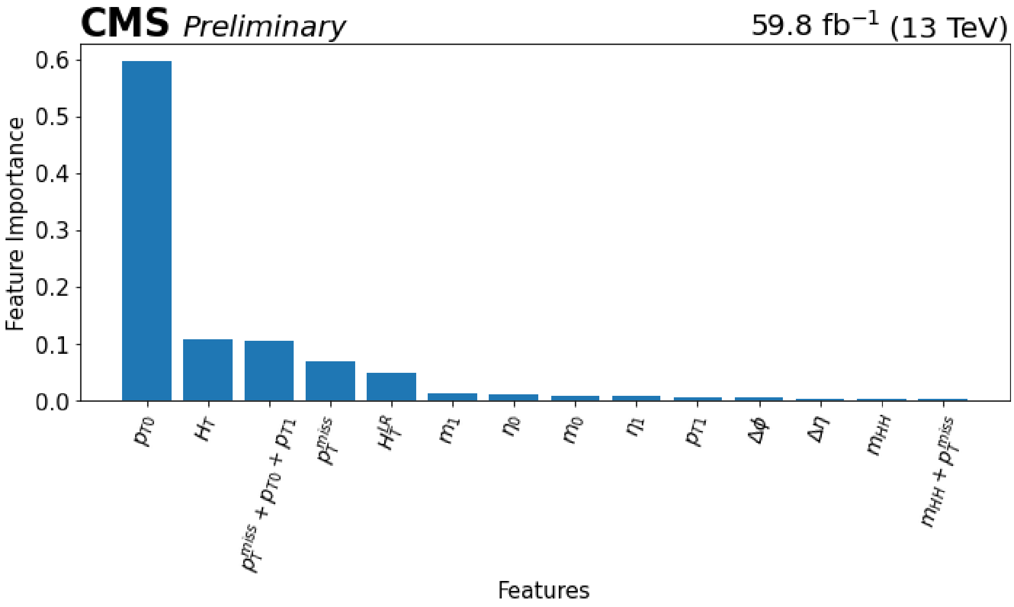


Figure 29: Feature importance ranking plot for the model trained using the data sample of the boosted HH analysis.

7.2.1 Trigger efficiency curves with aleatoric uncertainties

The trigger efficiency curves containing the gradient boosting predicted efficiencies, the true efficiencies and the traditionally measured efficiencies as a function of the variables of interest in the TTbar sample are presented in Figure 30. As in the boosted HH analysis case, the method seems to work well when applied to this background sample the model was also initially trained with.

The results for the TTbar trained model used in the M200 sample are presented in Figure 31. As can be observed from the plots, the model seems to be able to generalize into this process well. The signal process simulated in this M200 sample is closer to the background process than the process simulated in the M1000 sample, so it is reasonable that the model is able to generalize into this process.

The trigger efficiency plots comparing the gradient boosting predicted efficiencies to the efficiencies measured using traditional methods for the M1000 sample are presented in Figure 32. As seen in the plots, the model isn't able to generalize into this sample, and the efficiencies predicted by the model trained using the TTbar sample have systematically lower values than the true efficiencies in this sample. There are some qualities in the M1000 sample that improve the efficiency in a way that the TTbar model doesn't recognize, and this can be caused, for example, by the training data not having enough events resembling the events in the M1000 sample.

Lastly, the results for the model trained using the data sample are presented in Figure 33. The method seems to work well in the data sample also in this analysis case, which is promising regarding the possibility of using this method in the charged Higgs boson analysis as well, with the caveat that the efficiency estimate does not generalize to signal scenarios with very heavy (TeV-scale) H^+ masses.

The uncertainties in this analysis acquired using the bootstrapping method are larger than the bootstrap uncertainties in the boosted HH analysis across all the simulated samples and the data sample. Especially with higher p_T values, there is a

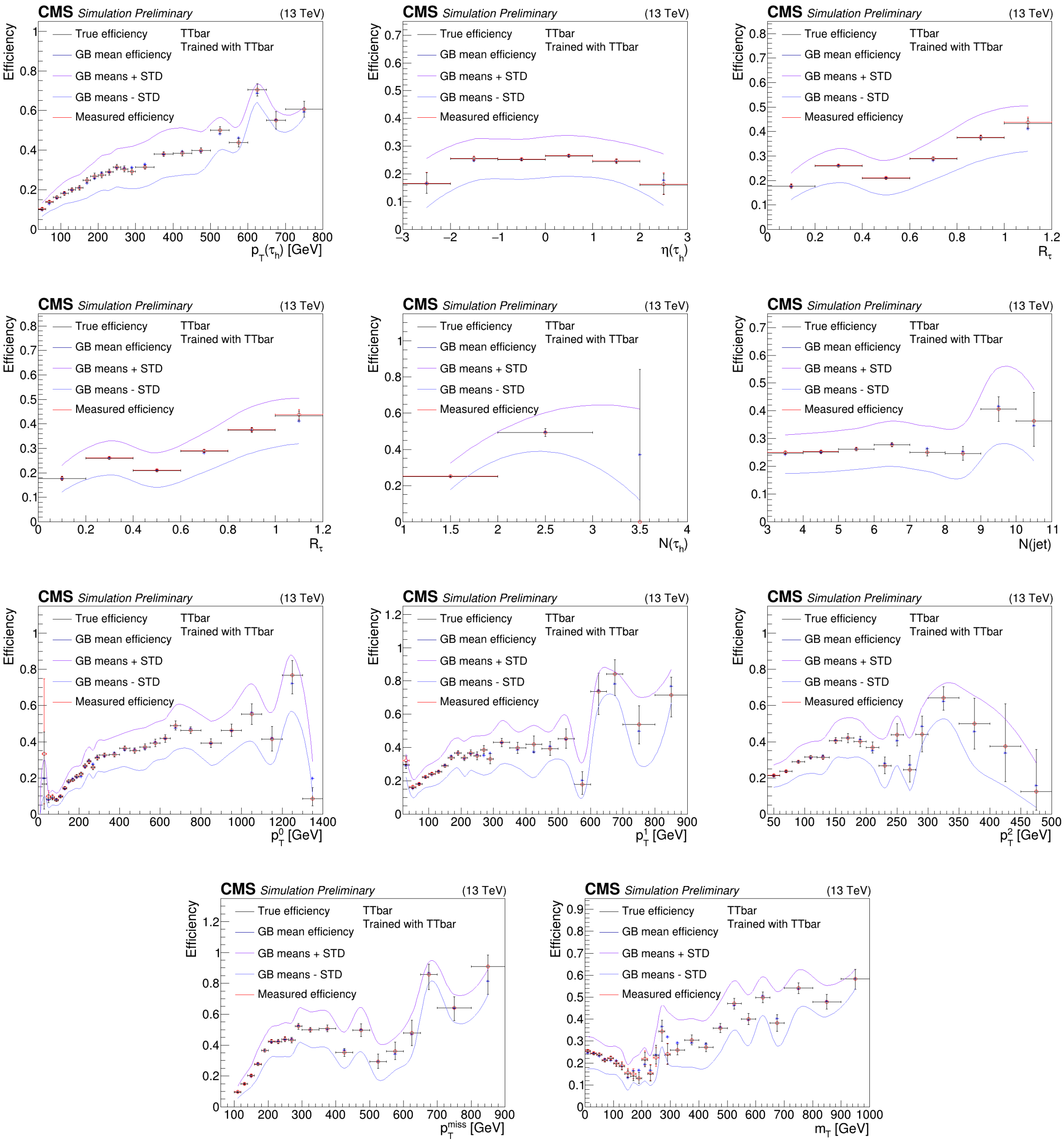


Figure 30: Trigger efficiency in the charged Higgs boson analysis as a function of variables of interest in the TTbar sample, containing the true efficiencies, the reference trigger measured efficiencies, the efficiencies predicted using the TTbar trained gradient boosting (GB) model, and the corresponding aleatoric uncertainties, which are the standard deviations (STD) obtained with the bootstrapping method.

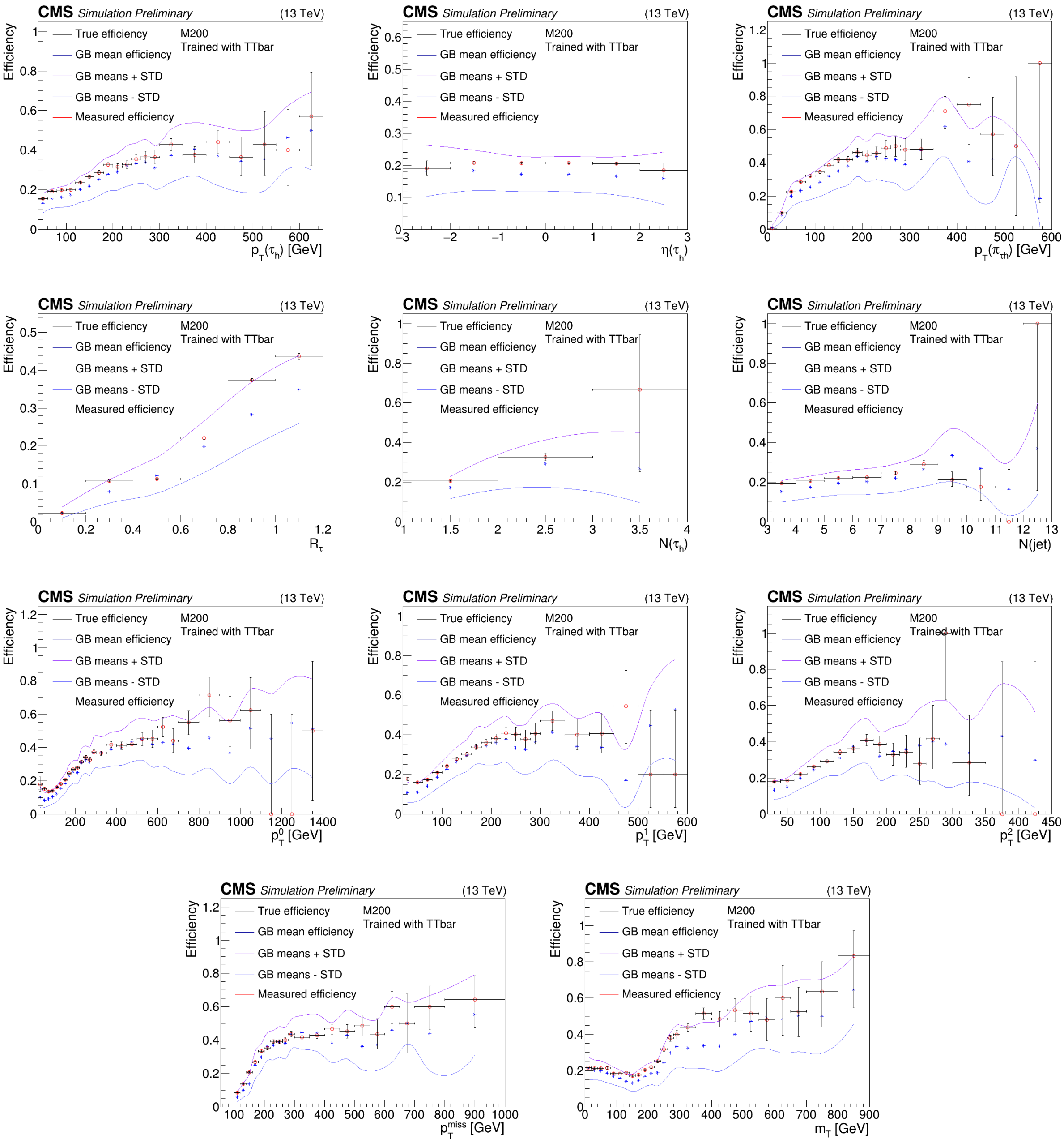


Figure 31: Trigger efficiency in the charged Higgs boson analysis as a function of variables of interest in the M200 sample, containing the true efficiencies, the reference trigger measured efficiencies, the efficiencies predicted using the TTbar trained gradient boosting (GB) model, and the corresponding aleatoric uncertainties, which are the standard deviations (STD) obtained with the bootstrapping method.

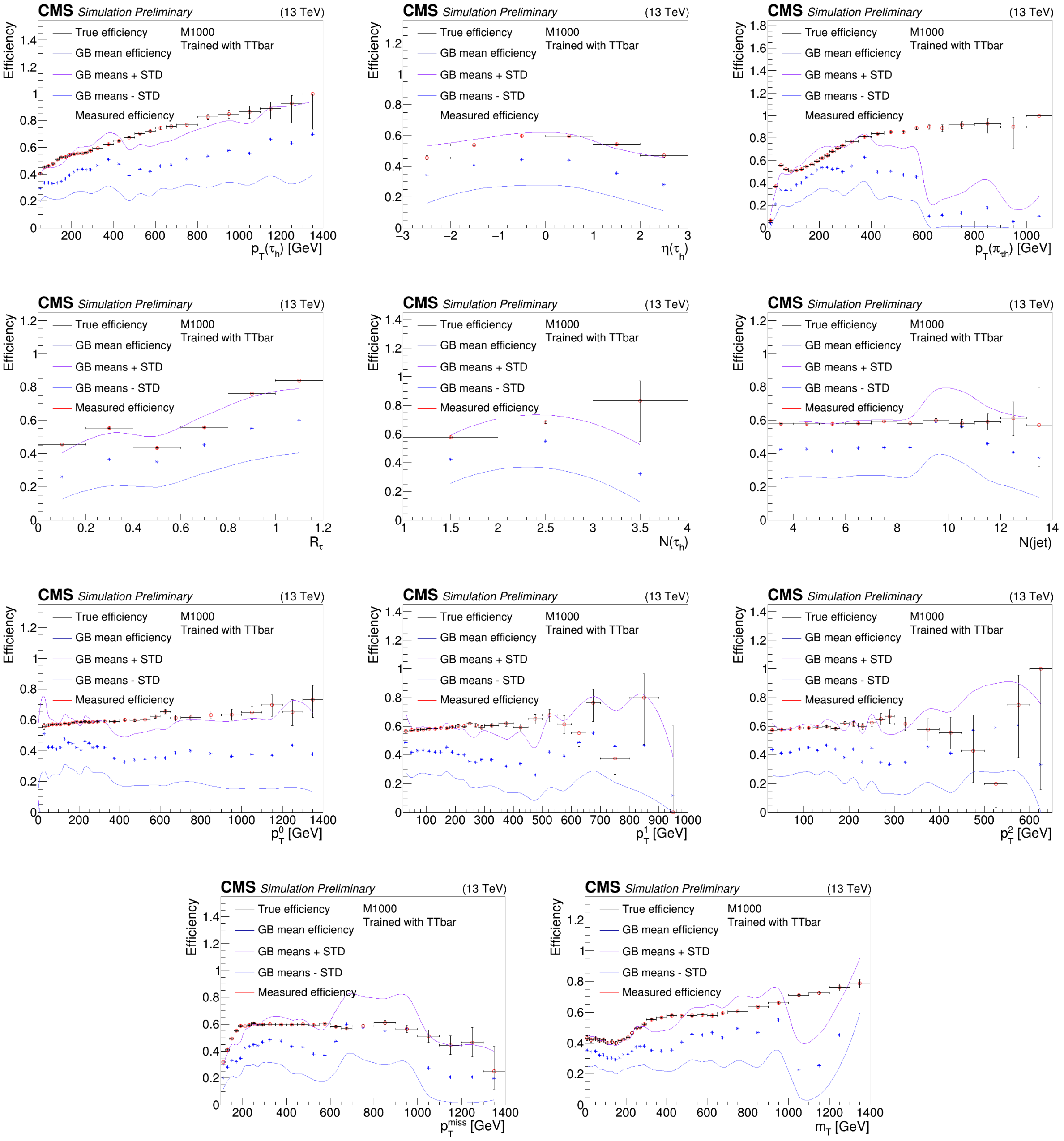


Figure 32: Trigger efficiency in the charged Higgs boson analysis as a function of variables of interest in the M1000 sample, containing the true efficiencies, the reference trigger measured efficiencies, the efficiencies predicted using the TTbar trained gradient boosting (GB) model, and the corresponding aleatoric uncertainties, which are the standard deviations (STD) obtained with the bootstrapping method.

major difference in the predictions made by the models with different subsamples of training data in the bootstrapping phase. This is probably because the samples do not have many events with these high p_T values, so depending on the events chosen for the training subsamples there might be only a very low amount of events with high p_T values in certain subsamples.

7.2.2 Trigger efficiency as a function of two variables simultaneously

We can again investigate the model's performance as a function of several variables simultaneously by varying one variable and plotting the trigger efficiency as a function of another. In this charged Higgs boson analysis, we analyze the model's performance as a function of $p_T(\tau_h)$ and p_T^{miss}, first by selecting events in certain p_T^{miss} ranges and plotting the trigger efficiency as a function of $p_T(\tau_h)$, and then varying the $p_T(\tau_h)$ and plotting the trigger efficiency as a function on p_T^{miss}. The results obtained from the data sample are presented in Figure 34, where we can again notice that the gradient boosting predicted efficiencies are aligning with the measured efficiencies well. The model is thus performing well as a function of multiple variables simultaneously.

7.2.3 Performance

The AUC score and accuracy for the model trained using the TTbar sample and the model trained using the data sample are presented in Table IV. The model trained using data has an accuracy of 0.928 and an AUC score of 0.859 when applied to the data sample, and the model trained using the TTbar sample has an accuracy of 0.911 and an AUC score of 0.920 when applied to the TTbar sample. The ROC curve for the model trained and used in the TTbar sample is presented in Figure 35, and for the model trained and used in the data sample in Figure 36. From these results we can conclude that the models perform well in these two samples,

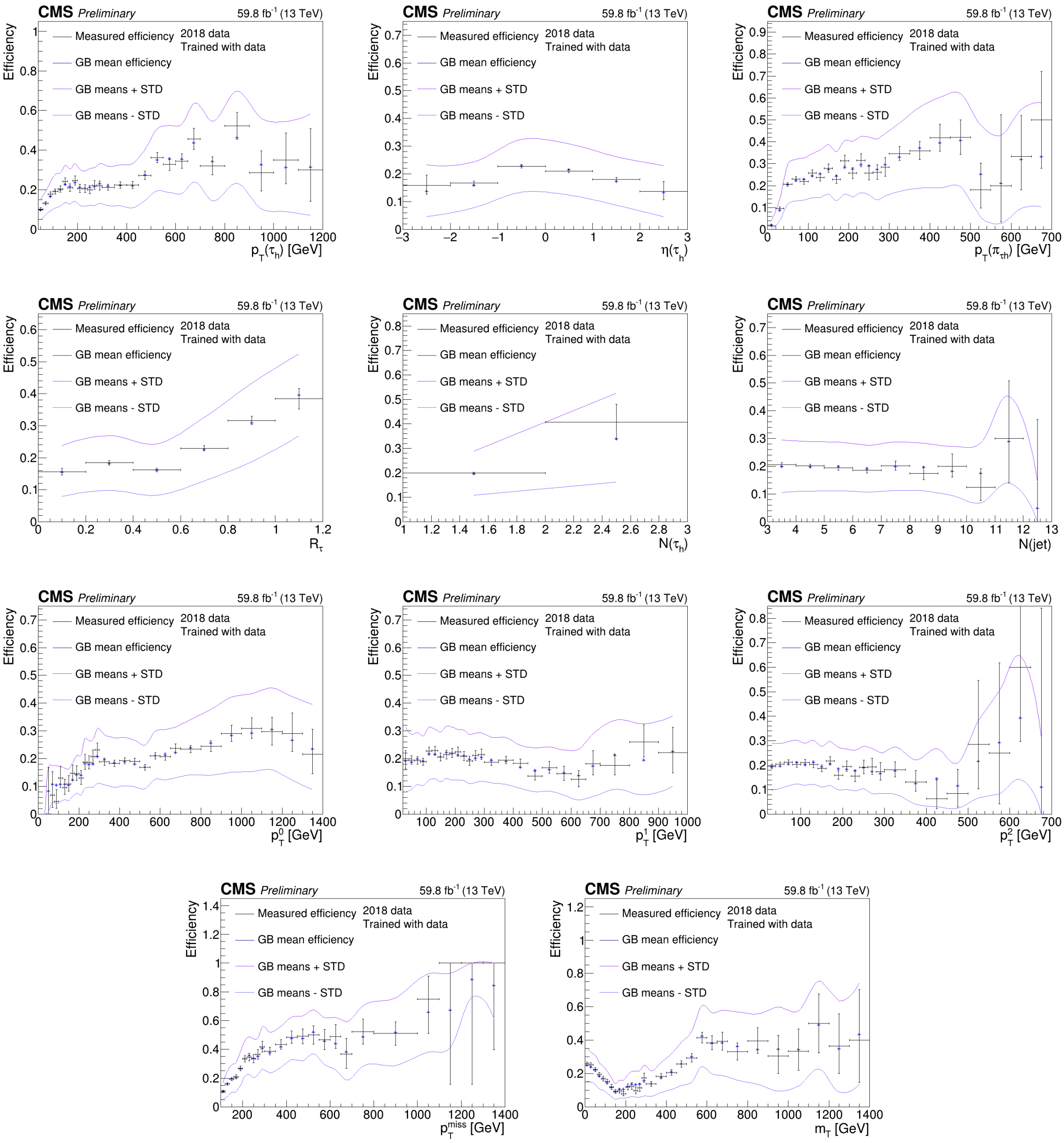


Figure 33: Trigger efficiency in the charged Higgs boson analysis as a function of variables of interest in the data sample, containing the reference trigger measured efficiencies, the efficiencies predicted using the gradient boosting (GB) model trained with the data sample, and the corresponding aleatoric uncertainties, which are the standard deviations (STD) obtained with the bootstrapping method.

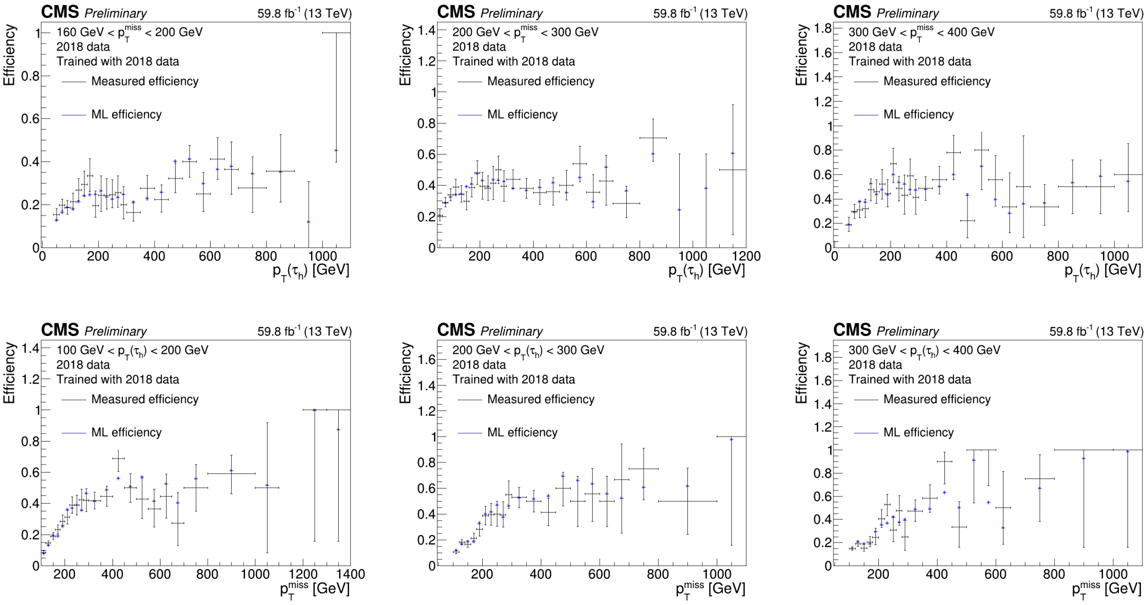


Figure 34: The trigger efficiency as a function of $p_T(\tau_h)$ in three different p_T^{miss} ranges (top row), and the trigger efficiency as a function of p_T^{miss} in three different $p_T(\tau_h)$ ranges (bottom row) in the charged Higgs boson data sample.

the model trained using the data sample having a slightly higher accuracy and the model trained using the TTbar sample having a slightly higher AUC score out of these two models.

The accuracies and AUC scores of the model trained with the TTbar sample and applied to samples M200 and M1000 are also included in Table IV. The model performs well in the M200 sample, as can be seen from the accuracy, which is 0.892, and the AUC score, which is 0.874. However, as was already evident in the trigger efficiency plots presented previously, the model isn't able to perform well in the M1000 sample. This can also be observed in the performance metric values for this sample, since the model's accuracy in this sample is 0.561 and the AUC score is 0.628. The model is thus performing only slightly better than a random classifier in this sample. The ROC curves for the model applied to both of these signal samples are also presented in Figure 35.

Table IV: The accuracies and AUC scores in the charged Higgs boson analysis for the model trained using the simulated TTbar sample applied to the TTbar, M200 and M1000 samples, and the model trained with and applied to the data sample

	Accuracy	AUC score
Data	0.928	0.859
TTbar	0.911	0.920
M200	0.892	0.874
M1000	0.561	0.628

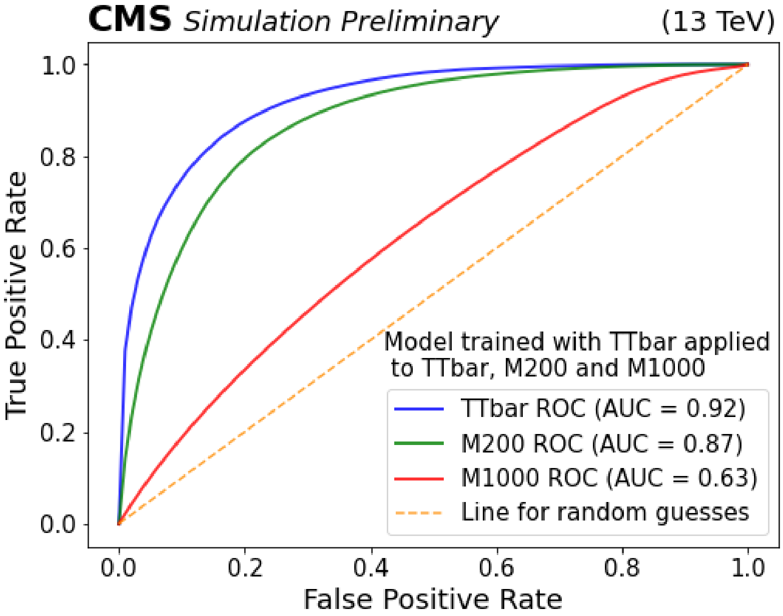


Figure 35: The ROC curve for the model trained using the TTbar sample applied to all simulated samples in the charged Higgs boson analysis.

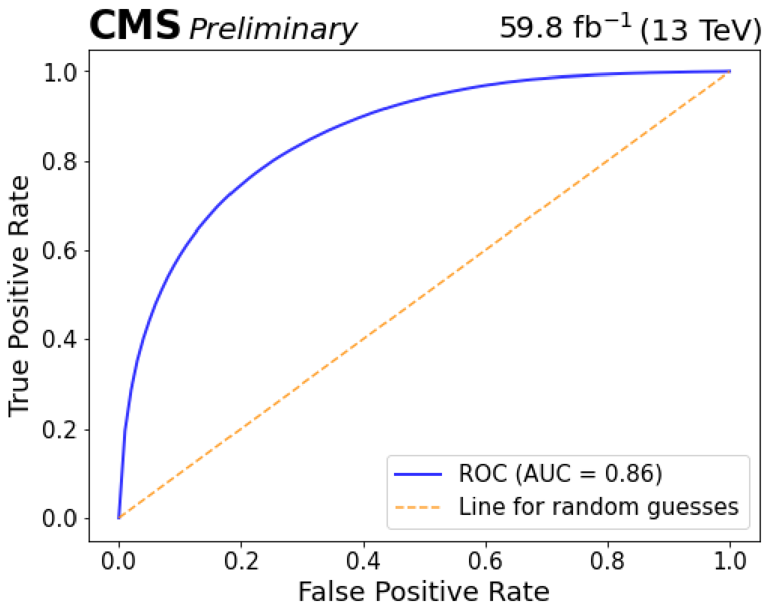


Figure 36: The ROC curve for the model trained using the data sample of the charged Higgs boson analysis.

The feature importance ranking plot for the model trained using the TTbar sample is presented in Figure 37. Compared to the boosted HH case, the importance of each feature is slightly more balanced in this charged Higgs boson case. The most important feature is the p_T^{miss}, which has an importance of over 0.4. This is again quite reasonable, since one condition in the signal trigger targets the p_T^{miss} value. The $p_T(\pi_{\tau h})$ and m_T have an importance value of around 0.1, and p_T^0 and $p_T(\tau_h)$ have an importance value of slightly less than 0.1. The signal trigger targets also the $p_T(\tau_h)$ and the $p_T(\pi_{\tau h})$, so it is understandable that these features contribute to the model's decisions. The $N(jet)$ and $N(\tau)$ variables have the lowest importance values out of all the variables.

In Figure 38, the feature importance ranking plot for the model trained using the data sample is presented. The importance ranking and values are quite similar than in the TTbar trained model, but in this case the m_T variable has a larger importance value than $p_T(\pi_{\tau h})$, and the $p_T(\tau_h)$ has also a slightly larger importance value than p_T^0.

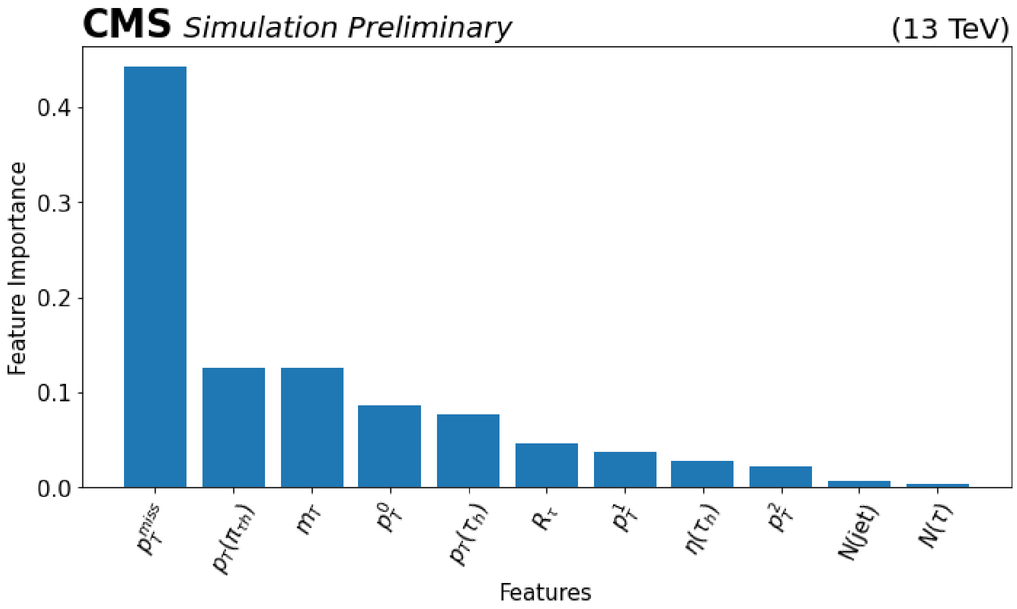


Figure 37: Feature importance ranking plot for the model trained using the TTbar sample.

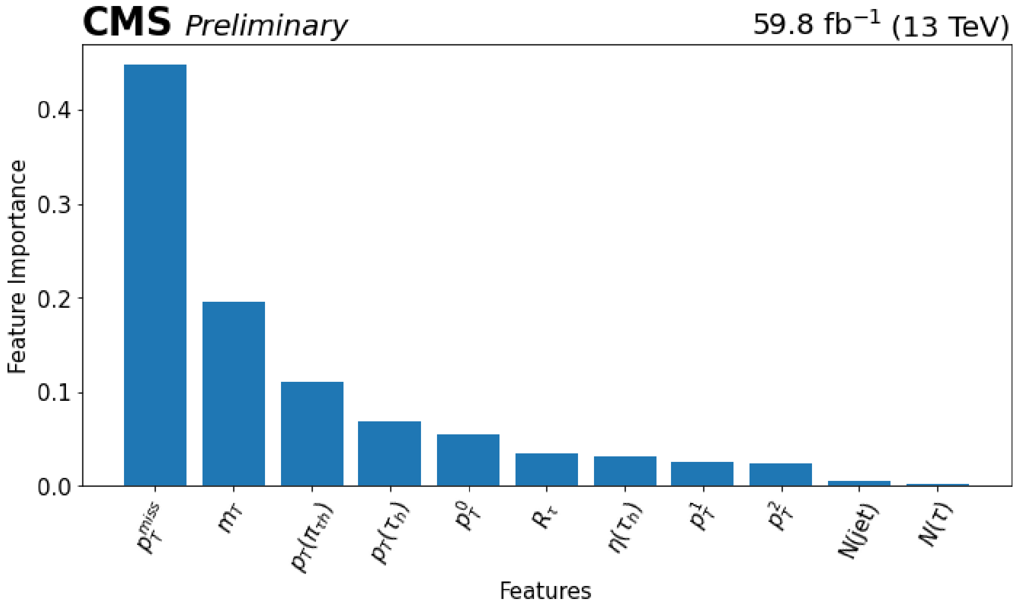


Figure 38: Feature importance ranking plot for the model trained using the data sample of the charged Higgs boson analysis.

8 Conclusion

In this thesis, a new method for estimating the trigger efficiency with the help of a machine learning algorithm was developed. The goal of the new method was to model the trigger efficiencies as a function of several variables simultaneously, since this is difficult to perform using traditional methods. The method is based on a gradient boosting algorithm, and the aleatoric uncertainties related to this method were estimated using a bootstrapping method. The epistemic uncertainties related to the model architecture were also studied by comparing the results given by a deep neural network model to the bootstrapped gradient boosting model. This new method was tested on two ongoing analyses: a boosted Higgs boson pair production analysis and a search for charged Higgs bosons analysis. The model was tested by using simulated samples and data in these both cases. Simulated samples were used to validate the method, since we are able to compare the results given by this new method to the true trigger efficiency calculated using all the events in the sample. After validating the method with the simulated samples, the method was then applied on a data sample, since the main goal of the method is to measure the trigger efficiency in data.

In the boosted Higgs boson pair production analysis, the model was trained using a background sample that simulates multijet production via quantum chromodynamics interactions, and the model was then used in the background sample and two signal samples, which simulate the Higgs boson self-coupling via gluon fusion and vector boson fusion. The new method seems to work well across all simulated samples. The model was then trained and tested using the data sample, and it was concluded that the method also works well in the data sample.

In the charged Higgs boson analysis case, the method seems to work well for the background sample which simulates top quark pair production, and for the data sample. It seems to also work well for the simulated sample that simulates the process where the mass of the charged Higgs boson is 200 GeV. However, in the sample that simulates the process where the charged Higgs boson has a mass of 1000 GeV, there seems to be some qualities that improve the efficiency in a way that the model is not able to recognize.

For the charged Higgs boson analysis case, the method would need improving in the case of modelling the process where the charged Higgs boson has a mass of 1000 GeV. One possibility is to use a different kind of training data, which could be, for example, a mixture of the samples simulating both the multijet production via quantum chromodynamics interactions and the top quark pair production. This could increase the events which have similar qualities than the events in the sample that simulates the process where the mass of the charged Higgs boson is 1000 GeV. However, this would only work when training the model with simulated samples. We could also study the differences between the sample simulating the process of 1000 GeV charged Higgs boson and other samples, and possibly add new variables to the gradient boosting model which might improve the performance. There is also a possibility of trying a model architecture which is not presented in this thesis to see if it would perform better in this case.

In conclusion, the method seems to be beneficial for the boosted Higgs boson pair production analysis, since the model performs well across all signal samples and in the data sample, where the model would be used in actual analysis. Testing of this new method in the boosted Higgs boson pair production analysis conducted at the CMS experiment is currently starting.

References

[1] The ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. Phys. Lett. Sect. B Nucl. Elem. Part. High-Energy Phys. 716, 1–29 (2012).

[2] Mariotti, C. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. 13th Marcel Grossmann Meet. Recent Dev. Theor. Exp. Gen. Relativ. Astrophys. Relativ. F. Theor. - Proc. MG13 Meet. Gen. Relativ. 2012 0, 352–372 (2015).

[3] The CMS Collaboration. Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s}$=7 and 8 TeV. J. High Energy Phys. 2013, 81 (2013).

[4] The CMS Collaboration. The CMS experiment at the CERN LHC. J. Instrum. 3 (2008) S08004 doi:10.1088/1748-0221/3/08/S08004.

[5] Morii, T., Lim, C. S. & Mukherjee, S. N. The Physics of the Standard Model and Beyond. World Scientific. (2004).

[6] Pauli, W. The Connection Between Spin and Statistics. Phys. Rev. 58, 716–722 (1940).

[7] T. D. Lee. A theory of spontaneous t violation. Phys. Rev. D, 8:1226–1239, (1973). 1226, doi: 10.1103/PhysRevD.8.1226

[8] MissMJ. Standard Model of Elementary Particles. PBS NOVA, Fermilab, Office of Science, United States Department of Energy, Particle Data Group (2006).

[9] Higgs, P. W. Broken symmetries and the masses of gauge bosons. Physical Review Letters vol. 13 508–509 (1964).

[10] Englert, F. & Brout, R. Broken symmetry and the mass of gauge vector mesons. Physical Review Letters vol. 13 321–323 (1964).

[11] Guralnik, G. S. & Hagen, C. R & Kibble, T. W. B. Global conservation laws and massless particles. Physical Review Letters vol. 13 585 (1964).

[12] The CMS Collaboration. Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons. J. High Energy Phys. 2014, 104 (2014).

[13] The CMS Collaboration. Search for nonresonant Higgs boson pair production in final states with two bottom quarks and two photons in proton-proton collisions at $\sqrt{s} = 13$ TeV. J. High Energy Phys. 2021, 257 (2021).

[14] The LHC Higgs Cross Section Working Group et al. Handbook of LHC Higgs Cross Sections: 3. Higgs Properties. (2013) doi:10.5170/CERN-2013-004.

[15] The CMS Collaboration. Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector. Phys. Lett. Sect. B Nucl. Elem. Part. High-Energy Phys. 779, 283-316 (2018).

[16] The CMS Collaboration. Search for charged Higgs bosons in the $\mathrm{H}^{\pm} \rightarrow \tau^{\pm}\nu_{\tau}$ decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV. J. High Energy Phys. 2019, 142 (2019)

[17] Degrande, C. et al. Accurate predictions for charged Higgs production: Closing the $m_{H}^{\pm} \sim m_{t}$ window. Phys. Lett. Sect. B Nucl. Elem. Part. High-Energy Phys. 772, 87–92 (2017).

[18] The CMS Collaboration. Search for a charged Higgs boson in pp collisions at $\sqrt{s} = 8$ TeV. J. High Energy Phys. 2015, 1–64 (2015).

[19] Evans, L. & Bryant, P. LHC Machine. J. Instrum. 3 (2008).

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

[21] ALICE collaboration. ALICE: Technical proposal for a Large Ion collider Experiment at the CERN LHC. LHC technical proposal. (1995).

[22] The LHCb Collaboration. The LHCb detector at the LHC. J. Instrum. 3, (2008).

[23] Adriani, O. et al. The LHCf detector at the CERN Large Hadron Collider. J. Instrum. 3, (2008).

[24] The TOTEM Collaboration. The TOTEM Experiment at the CERN Large Hadron Collider. J. Instrum. 3, (2008).

[25] The MoEDAL-MAPP Collaboration. MoEDAL-MAPP, an LHC Dedicated Detector Search Facility. 1–39 (2022).

[26] FASER Collaboration. Search for Dark Photons with the FASER detector at the LHC. 1–19 (2023).

[27] SND Collaboration. Scattering and Neutrino Detector Letter of Intent. (2020).

[28] CMS Detector website [online, accessed 2024-01-15]

[29] Barney, David. CMS Detector Slice. CMS Collection. (2016)

[30] The CMS Collaboration. The TriDAS Project Technical Design Report, Volume 1: The Trigger Systems. Design 599 (2000).

[31] The CMS Collaboration. Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV. J. Instrum. 15, (2020).

[32] The CMS Collaboration. The CMS trigger system. J. Instrum. 12, (2017).

[33] The CMS Collaboration. Particle-flow reconstruction and global event description with the CMS detector. J. Instrum. 12, (2017).

[34] The CMS Collaboration. Trigger efficiency studies of the CMS Run-3 Data Scouting of Jet, Electron and Photon objects and comparison with standard HLT paths during proton-proton collisions at $\sqrt{s} = 13.6$ TeV. (2023).

[35] Breiman, L. Random forests. Random Forests, 1–122. Mach. Learn. 5–32 (2001).

[36] Bengio, Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning vol. 2 (2009).

[37] Marius-Constantin, P., Balas, V. E., Perescu-Popescu, L. & Mastorakis, N. Multilayer perceptron and neural networks. WSEAS Trans. Circuits Syst. 8, 579–588 (2009).

[38] Breiman, L. Arcing the edge. (1997)

[39] Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. Annals of statistics pp. 1189–1232, (2001)

[40] Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, (2011).

[41] Scikit-learn: GradientBoostingClassifier [online, accessed 2024-07-24]

[42] Barredo, A. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58 82–115, (2020).

[43] Casalicchio, G., Molnar, C., Bischl, B. Visualizing the Feature Importance for Black Box Models. In: Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science(), vol 11051. Springer, Cham. https://doi.org/10.1007/978-3-030-10925-7_40 (2019).

[44] Scikit-learn: Metrics and scoring: quantifying the quality of predictions: Log loss [online, accessed 2024-04-18]

[45] Géron, A. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Incorporated, (2019). ISBN: 9781492032649.

[46] Fawcett, T. An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874 (2006).

[47] Efron, B., & Tibshirani, R.J. An Introduction to the Bootstrap (1st ed.). Chapman and Hall/CRC. (1994). https://doi.org/10.1201/9780429246593