GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Fakultät für
Physik

Master's Thesis

Multi-Class Classification of $t\bar{t}\gamma$ Final States in Proton-Proton Collisions at $\sqrt{s} = 13\,\text{TeV}$ in ATLAS Using a Deep Neural Network

Mehr-Klassen-Klassifizierung von $t\bar{t}\gamma$ Endzuständen in Proton-Proton-Kollisionen bei $\sqrt{s} = 13\,\text{TeV}$ in ATLAS mit Hilfe tiefer Neuronaler Netze

prepared by

Steffen Korn

from Kiel

at the II. Institute of physics

Thesis number:	II.Physik-UniGö-MSc-2019/05
Thesis period:	15th April 2019 until 15th October 2019
First referee:	Prof. Dr. Arnulf Quadt
Second referee:	Prof. Dr. Stan Lai

Abstract

In recent years, machine learning approaches such as deep learning have shown remarkable results and have been implemented throughout different fields of science. Their applications in the field of high-energy physics have shown high potential for a vast range of different analyses.
In this thesis, two multi-class neural network approaches are presented in the scope of the $t\bar{t}\gamma$ process where a photon is radiated from charged particles in the initial, intermediate or final state. Two deep feed-forward neural networks are developed to separate events into three distinguished categories representing signal-like events, events with fake photons and events with photons from other prompt sources. The performance of these classifiers and their implementation in the scope of a measurement of the $t\bar{t}\gamma$ cross section is described.

Zusammenfassung

Machine-Learning-Ansätze wie zum Beispiel das sogenannte „Deep Learning“ haben in den letzten Jahren bemerkenswerte Resultate in vielen unterschiedlichen Bereichen der Wissenschaft hervorgebracht. Ihre Anwendung im Bereich der Hochenergiephysik zeigt großes Potential für eine große Bandbreite an verschiedenen Analysen.
In dieser Arbeit werden zwei mögliche Mehrklassenansatz im Rahmen des $t\bar{t}\gamma$-Prozesses präsentiert, wobei ein Photon von einem geladenen Teilchen im Anfangs-, Zwischen oder Endzustand abgestrahlt wird. Zwei „Deep-Feed-Forward-Netzwerke“ werden entwickelt um Ereignisse in drei unterschiedliche Kategorien, Signalereignisse, Ereignisse mit „Fake-Photonen“ und Ereignisse mit anderen prompten Photonen einzuteilen. Die Trennkraft dieser beiden Modelle und ihre mögliche Einbindung im Rahmen einer Messung des $t\bar{t}\gamma$-Wirkungsquerschnittes werden beschrieben.

Contents

1. Introduction

The purpose of experimental physics is to systematically study nature and to expand the human understanding of the universe. In the field of experimental particle physics the fundamental building blocks of our universe, elementary particles, are studied. Throughout the last century, these studies have widened our understanding of matter, atoms and the fundamental forces that govern the interaction of particles.
Two of these elementary particles are of particular interest for this analysis, the top quark and the photon. Processes involving these two elementary particles, as well as other background processes with similar signatures, will be the basis of a deep multi-class neural network that is developed in this thesis.
To motivate this approach, the standard model of elementary particle physics (SM) is introduced in Section 1.1. The top quark is introduced in Section 1.2 and its coupling to the photon is explained. Furthermore, background processes are presented which contribute to a measurement of the $t\bar{t}\gamma$ cross section with the ATLAS detector. The detector and the LHC are explained in Chapter 2. These background processes and their contribution are based on Monte Carlo simulations and on the applied event selection presented in Chapters 3 and 4. The resulting event selection yields a signal to background ratio of approximately $1:1$ with two major background categories. Based on the event selection the concepts, motivation and results for two different multi class approaches are given in Chapters 5 and 6. At the end of the thesis a summary, conclusions and an outlook are provided in Chapter 7.

1.1. The Standard Model

The Standard Model (SM) of elementary particle physics describes the behaviour of three out of the four known fundamental forces of nature. It combines electromagnetism, the weak force and the strong force in one model and describes the elementary particles as well as their interactions.
The SM is the combination of three field theories, two of which, electromagnetism and the weak force, are unified in the electroweak theory [1–4]. Its building blocks, such as

quantum electrodynamics (QED) [5–8], were created well before its formulation in the 1970's.

It is formulated as a local gauge and Lorentz invariant Lagrangian density function with the underlying gauge group $U(1)_Y \times SU(2)_L \times SU(3)_C$. The gauge fields of the underlying symmetry give rise to the three fundamental interactions mentioned above. Whilst $U(1)_Y \times SU(2)_L$ describe the electroweak interaction, $SU(3)_C$ describes the strong interaction in the framework of Quantum Chromodynamics (QCD) [9, 10].

The interactions between the twelve spin-$\frac{1}{2}$ fermions are mediated by spin-1 gauge bosons. While photons (γ) are the mediators of electromagnetism, the $W^{\pm}$ and the Z^0 boson mediate the weak force. The strong force is mediated by gluons (g).

The twelve fermions are grouped into three generations. Of these twelve fermions, six are called leptons and six are called quarks. The leptons consist of electrically neutral neutrinos and the electrically charged leptons: the electron (e), muon (μ) and tau (τ). While the neutrinos have a third component of the weak isospin of $+\frac{1}{2}$, the charged leptons have a third component of -$\frac{1}{2}$. The charged leptons interact via the electromagnetic and weak force and neutrinos exclusively via the weak force.

The quarks are also further grouped into up- and down-type quarks (up (u), charm (c), top (t), down (d), strange (s), bottom (b)). Each up-type quark carries a fractional charge of $\frac{2}{3}e$ and each of the down-type quarks carries a fractional charge of $-\frac{1}{3}e$. All quarks also carry a colour charge which is defined as red, green or blue. While the up-type quarks have a weak isospin of $\frac{1}{2}$, the down-type quarks have a weak isospin of $-\frac{1}{2}$.

The quarks interact via the strong, weak and electromagnetic force. Due to colour confinement, quarks and gluons cannot be observed as free particles. They hadronise and form composite particles, the colour neutral hadrons. These hadrons consist of either a quark and an antiquark pair (mesons), three quarks or three antiquarks (baryons).

Through the spontaneous breaking of the $SU(2)_L \times U(1)_Y$ gauge symmetry, the SM predicts the existence of a heavy scalar boson, the so called Higgs boson. Furthermore, the masses of the $W^{\pm}$ and Z^0 boson are also predicted by the SM. The fermions in the SM acquire mass through the interaction with the field of the heavy scalar boson. This mechanism is described by the Brout-Englert-Higgs mechanism [11, 12] which was predicted in 1964.

The particles of the SM were completed with the discovery of the Higgs boson by the Atlas and Cms collaborations in 2012 [13, 14].

Even though the predictions of the SM have been tested to high degrees of accuracy, it is limited to the low-energy sector and contains gaps in its formalism of which some are highlighted below.

The SM does not include gravity. Attempts to introduce gravity into the SM have not yet been successful and no experimental evidence for gauge bosons mediating gravity has yet been found in experiments. While gravity plays a major role in the cosmos its effects are negligible in high-energy physics.
The masses of all fermions and bosons within the SM are not predicted by the SM but have to be determined experimentally giving rise to a number of free parameters within the SM.
The discovery of neutrino oscillations [15] provides experimental evidence for a non-vanishing neutrino mass. This is not explained by the SM which predicts massless neutrinos. Their masses are yet to be determined experimentally.
Dark matter and dark energy make up about 96 % of the observable universe [16–19]. Their existence is suggested through indirect experimental evidence [20–22]. As of now, there is no experimental evidence for particles explaining the observations of dark matter and dark energy.
It is evident, that the SM is not complete and that we need to fill the gaps within it. Manifestations of these gaps could be new particles or interactions that can not be explained by the SM alone. By carefully interpreting the results of the measurements within the framework of the SM, these particles and interactions may be uncovered in the future.

1.2. The top quark

The top quark was discovered in 1995 by the CDF and DØ collaborations [23, 24] at the Tevatron. It is the heaviest particle in the SM. With its mass of $m_t = 173.0 \pm 0.4\,\text{GeV}$ [25, 26] it is significantly heavier than all the other particles in the SM. Its measured properties are in agreement with the SM. The high mass of the top quark leads to a life time of the order of $10^{-25}\,\text{s}$ assuming a top quark mass of $m_t \approx 173\,\text{GeV}$, a $W^{\pm}$ boson mass of $m_W \approx 80.4\,\text{GeV}$ and $\alpha_s(m_Z) \approx 0.118$ for the strong coupling constant. Since the hadronisation timescale is $\mathcal{O}(10^{-24})\,\text{s}$, the top quark decays before it hadronises therefore providing an opportunity to study bare quarks.

Top quark production and decay

The dominant production mode of top quark pairs at the Large Hadron Collider, Lhc, is gluon-gluon fusion in collisions of protons ($gg \to t\bar{t}$). In addition, quark antiquark annihilation ($q\bar{q} \to t\bar{t}$) contributes. At the current centre-of-mass-energy of the Lhc of 13 TeV, approximately 90 % of all top quark pairs are created through the process of gluon-gluon

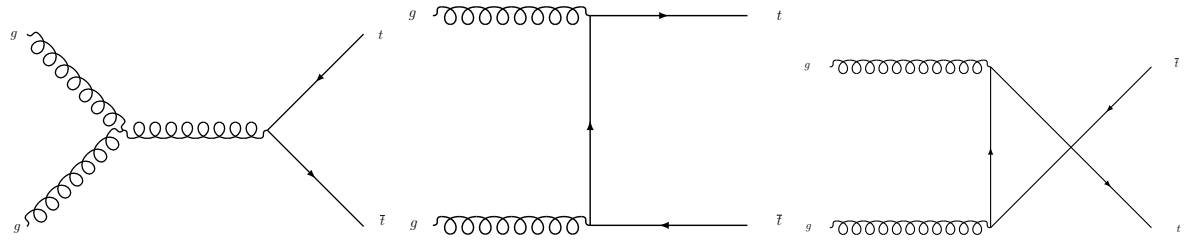

Figure 1.1.: Feynman diagrams depicting the process ($gg \to t\bar{t}$).

fusion. Feynman diagrams depicting the process of gluon-gluon fusion are shown in Figure 1.1.

Single top quarks are also produced at the LHC via the electroweak interaction. With a production cross section that is a approximately a factor of three times lower than the production cross section for top quark pairs, single top quarks are dominantly produced in the t-channel production mode $ub \to dt$ or $\bar{d}b \to \bar{u}t$ via the exchange of a virtual $W^{\pm}$ boson.

Furthermore, single top quarks are produced in the s-channel production mode, $u\bar{d} \to t\bar{b}$, also via the exchange of a virtual $W^{\pm}$ boson and with a real $W^{\pm}$ boson via $gb \to Wt$. The corresponding Feynman diagrams for these three processes are shown in Figure 1.2.

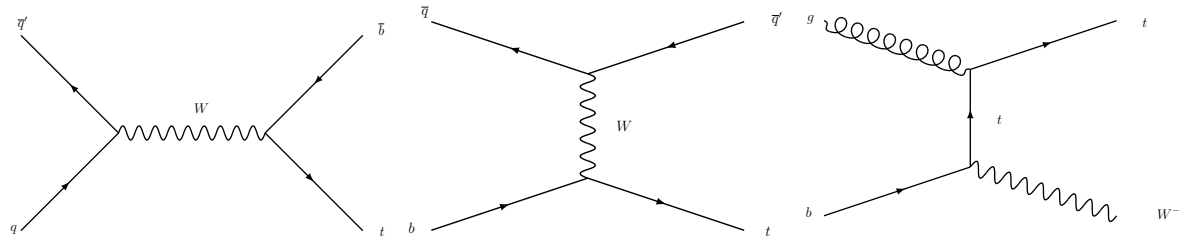

Figure 1.2.: Feynman diagrams depicting the process $\bar{q}'q \to \bar{b}t$, $\bar{q}'b \to \bar{q}t$ and $gb \to Wt$.

In the SM, the top quark decays into a $W^{\pm}$ boson and a b quark almost 100 % of the time [27] due to the suppressed off-diagonal elements of the CKM matrix [28, 29]. The $W^{\pm}$ boson subsequently decays into an up-type quark and a down-type antiquark or a lepton and its corresponding antineutrino.

While the leptonic decay happens 33 % of the time, the hadronic decay happens 67% of the time [30]. The decay channel of a pair of top quarks is characterised by the decay of the two $W^{\pm}$ bosons in the final state. If both decay hadronically, the channel is labelled as "all-hadronic". If one of the $W^{\pm}$ bosons decays hadronically and one decays leptonically, the channel makes up one of the three single-lepton channels (e+ jets, μ+ jets, τ+ jets). From these three channels the e+ jets and μ+ jets channel are used in this analysis. The full leptonic decay, $t\bar{t} \to W^+bW^-\bar{b} \to l^+\nu_l l^-\bar{\nu}_l b\bar{b}$ can be split into decay channels with two

opposite charged leptons. These dileptonic decay channels are not used in this analysis. Dileptonic events where one lepton is not reconstructed or dileptonic events containing τ-leptons however contribute. The fully hadronic decay channel is not used. The branching ratios for the decay of a pair of top quarks are shown in Figure 1.3.

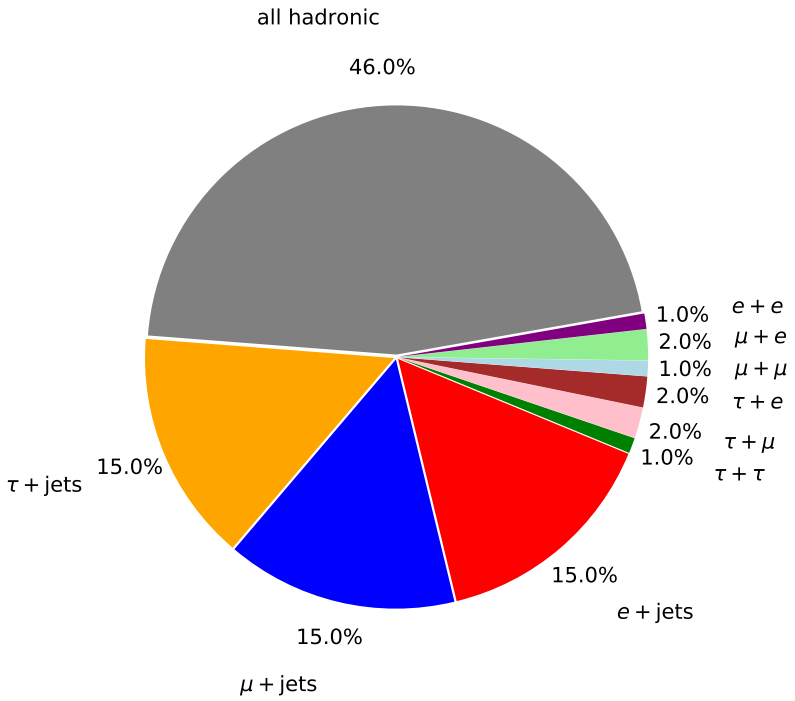

Figure 1.3.: Branching ratios of the $t\bar{t}$-decay [30]. Only the single-lepton channels composed of μ+jets and e+jets are used in this thesis.

The main irreducible background is composed of $W^{\pm}$+jets or Z^0+jets and a single lepton in the final states. In addition, electroweak processes involving single top quarks contribute to the background. A strong discriminant for distinguishing signal processes from background processes is the flavour of the jets since the $t\bar{t}$ final state is expected to contain two jets originating from b quarks.
Due to the high centre-of-mass-energy of the LHC and the rise of the $t\bar{t}$ production cross section with energy, the LHC is a factory for top quarks.

The top-photon coupling

In the following section, the measurement of the $t\bar{t}\gamma$ cross section is motivated.
Observation of the $t\bar{t}\gamma$ process with a significance of 5.3σ was seen at the LHC by ATLAS at $\sqrt{s} = 7\,\text{TeV}$ with a total dataset size of $4.59\,\text{fb}^{-1}$ [31].
The coupling of the photon to the top quark in the SM can be described by the QED

Lagrangian as shown in Equation (1.1)

$$\mathcal{L}_{QED} = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - q\bar{\psi}\gamma^\mu A_\mu \psi. \tag{1.1}$$

While the first term describes the kinematics of the spin-$\frac{1}{2}$ fermion field, ψ, the second term describes the kinematics of the photon field, A_μ. The electromagnetic field tensor $F_{\mu\nu}$ is defined as $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. The third term describes the interaction between the photon and the fermion field. It is directly proportional to the charge q of the fermion providing a handle on the charge of the involved fermion.

The top quark charge is an important parameter of the SM. It was previously measured by ATLAS [32] at a centre-of-mass energy of $\sqrt{s} = 7\,\text{TeV}$. The measurement of the cross section of a top-antitop pair in association with a prompt photon can be used as a probe of the electromagnetic coupling between the top quark and the photon as described above in Equation (1.1). The results of the direct measurement of the top quark coupling can be used to test anomalous top quark couplings that could manifest themselves in cross section measurements or as shape discrepancies in different kinematic distributions [33]. Furthermore, results can be interpreted in the framework of an effective field theory describing the top-photon coupling in the search for physics beyond the SM [34]. In addition, the $t\bar{t}\gamma$ contributes as a background process in other electroweak measurements [35, 36], thereby motivating a thorough understanding of the process and a measurement of its cross section.

While photons can be directly radiated from the top quark to form the process of interest, they can also be radiated from any charged particle in the $t\bar{t}\gamma$ final state. Furthermore, they can be radiated from any incoming quark in proton-proton collisions.

Effective measures to precisely discriminate the different origins of the photon are yet to be developed [37]. Therefore, in this analysis, $t\bar{t}\gamma$ characterises all final state configurations with two b-tagged jets coming from top quark decays, a lepton coming from the leptonic decay of a $W^\pm$ boson, two additional jets from the hadronic decay of the $W^\pm$ boson, and a prompt photon. Feynman diagrams depicting the radiation of a photon from charged particles in the initial, intermediate or final state are shown in Figure 1.4, with the photon marked blue. All shown diagrams are considered as signal.

Different abbreviations, categories and labels are used throughout this thesis to describe the different background processes as well as the signal process. They are introduced in the following paragraphs.

The backgrounds $W\gamma$ and $Z\gamma$ describe final states with a $W^\pm$ or Z^0 boson, a prompt photon and extra jets. These final states differ from the signal process by the flavour of the jets involved as described above. The background category "*Prompt*γ " characterises

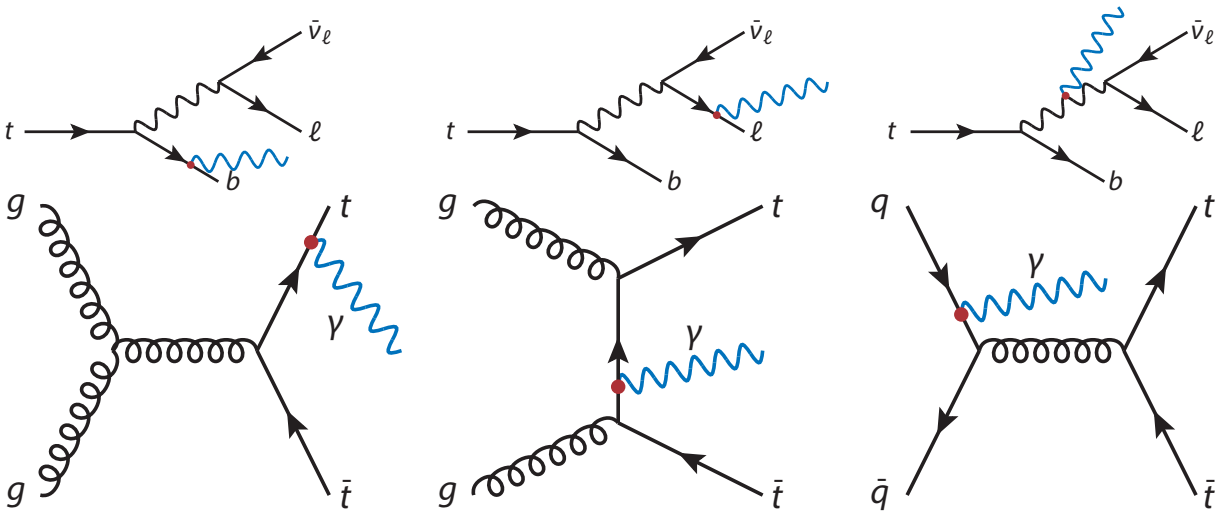

Figure 1.4.: Example diagrams of top quark production (top) and decays (bottom) where a photon is radiated from charged particles in either the initial or final state. All diagrams are considered as signal.

background events with a prompt photon coming from $t\bar{t}V$ processes (where V can be a $W^{\pm}$ or Z^0 boson) or diboson events with an associated photon.

Events with photons originating from jets or jets faking photons are characterised as hadronic fakes. In this analysis, hadronic fakes from processes involving a $t\bar{t}$-decay, defined as $t\bar{t}$ *h-fakes*, are treated separately from hadronic fakes from all remaining sources, defined as *h-fakes*. Electrons that are misidentified as photons are characterised as electronic fakes. Dominant sources for electronic fake photons are the dileptonic $t\bar{t}$-decay and the leptonic decay of a Z^0 boson into two electrons where one electron is misidentified as a photon. Electronic fakes from $t\bar{t}$-decays are defined as $t\bar{t}$ *e-fakes* and electronic fakes from all other sources are defined as *e-fakes*.

Single top processes with an additional photon from the s- and t-channel production modes via a virtual or real $W^{\pm}$ boson are referred to as $Wt\gamma$.

Due to the chosen event selection of this analysis which is explained in detail in Chapter 4 below, non-$t\bar{t}$ processes involving prompt photons define a significant contribution to the background. Processes involving fake photons also contribute significantly to the $t\bar{t}\gamma$ final state background. The different types of backgrounds motivate a separation into three different groups, each of which contains similar processes. The definition of these groups provides the basis of the multi-class approach pursued in this analysis.

All non-$t\bar{t}\gamma$ final states involving prompt photons such as $W\gamma$, $Z\gamma$ and Promptγ are grouped together into a category named *Other Prompt*. Events involving hadronic fake photons and electronic fake photons are grouped together into a category named *Fakes*. The last category is composed of the remaining $t\bar{t}\gamma$ signal process and is named $t\bar{t}\gamma$.

2. The LHC and the ATLAS experiment

To study elementary particles, dedicated particle detectors and complex accelerators are needed. The European Organization for Nuclear Research (CERN) houses experiments for several collaborations and thousands of scientists, making it the largest scientific research cluster in the world.
Over the past decades several different accelerators such as the Super Proton Synchrotron (SPS), the Large Electron-Positron Collider (LEP) and the Large Hadron Collider (LHC) have been commissioned. These colliders, and advanced high-energy physics experiments such as UA1 and UA2 or ATLAS and CMS have made ground-breaking discoveries in the field of high-energy physics such as the discovery of the $W^{\pm}$ [38], Z^0 [39] and Higgs boson possible. In this chapter, the LHC and the ATLAS detector are described in further detail.

2.1. The Large Hadron Collider

The LHC is currently the most powerful accelerator in the world. It collides protons at a centre-of-mass-energy of 13 TeV. All protons originate from a single bottle of hydrogen. The electrons are stripped off from the hydrogen nuclei and the protons are injected into the Linear accelerator 2 (Linac 2) where they are accelerated to an energy of 50 MeV. Then they enter the Booster and are further accelerated up to 1.4 GeV. Next, they enter the circular Proton Synchrotron (PS) where their energy is increased up to 25 GeV. Afterwards, they enter the Super Proton Synchrotron where they are accelerated up to 450 GeV before they are injected into the LHC, the final accelerator.
The protons are then accelerated within its ring of 27 km circumference up to the centre-of-mass-energy of 13 TeV. The acceleration is performed within eight straight segments. The proton beam is bent using 1232 dipole magnets and focussed using 392 quadrupole magnets. Within the LHC, the beam is brought to a collision at four interaction points (IP).
At these interaction points the main experiments ATLAS (A large Toroidal LHC Appara-

tuS) [40], ALICE (A Large Ion Collider Experiment) [41], CMS (Compact Muon Solenoid) [42] and LHCb (LHC beauty) [43] are located. ATLAS and CMS are multipurpose detectors searching for any kind of new physics and probing the many aspects of the SM. ALICE and LHCb are special-purpose detectors. While ALICE is focusing on studying the quark-gluon plasma forming at high temperatures and densities, LHCb focuses on physics with hadrons containing b quarks (b-physics), thereby studying CP-violation.

2.2. The ATLAS detector

In Figure 2.1, an overview of the ATLAS detector is shown. On the left side of the detector humans are shown for scale to illustrate the size of the detector. The detector has a width

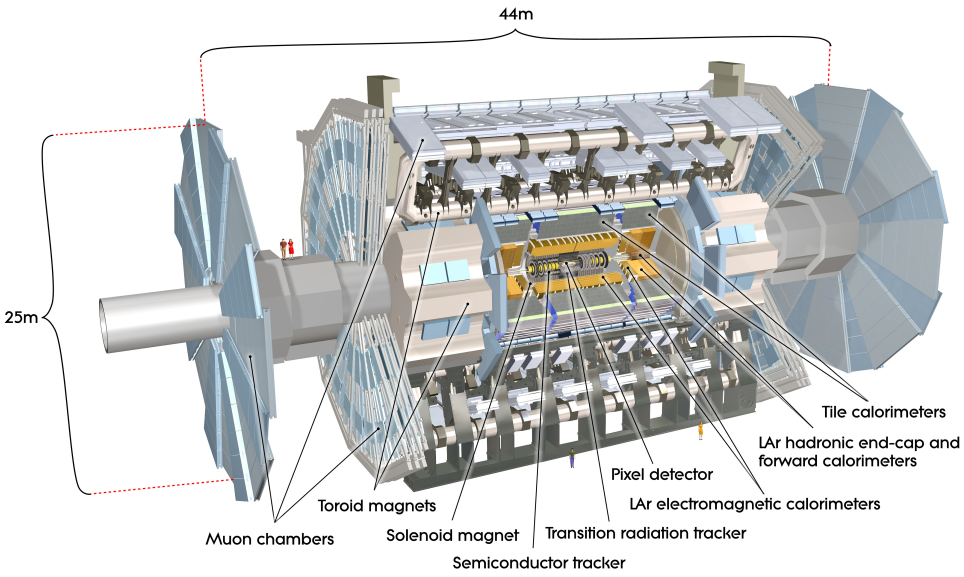

Figure 2.1.: An overview of the ATLAS detector. Several subdetectors and a human for scale are also shown ©CERN.

of 44 m, is 25 m tall and weighs about 7000 tons. It is composed of several different subdetectors that are arranged in an onion-like shape around the interaction point (IP). The IP defines the origin of the coordinate system which is used to describe the geometry within the detector. The z-axis is defined by the beam-axis with the x-y plane being transverse to the beam axis. The azimuthal angle ϕ is measured around the beam-axis and the polar angle θ is measured from the beam axis. The pseudorapidity $\eta = -\ln\tan(\theta/2)$ is often used to express the polar angle since differences in pseudorapidity are invariant under Lorentz transformation along the z-axis.

The onion-like structure of the ATLAS detector and examples of signatures caused by

traversing particles are further illustrated in the cross section view of the ATLAS detector in Figure 2.2. The innermost part of the detector is composed of the Inner Detector (ID).

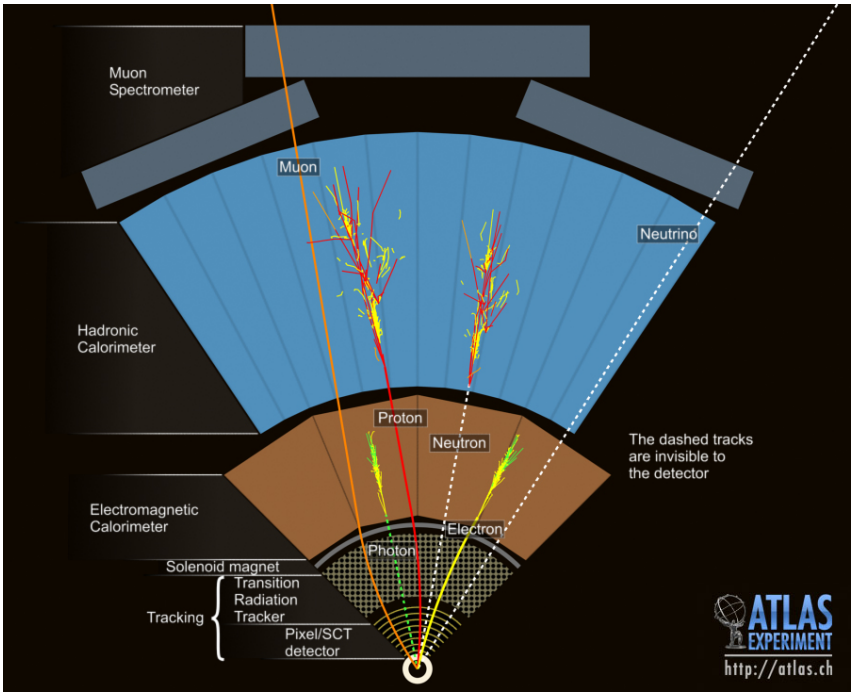

Figure 2.2.: Cross section view of the ATLAS detector with examples of signatures of different particles as measured by the detector ©CERN.

It is responsible for tracking charged particles. The ID is enclosed by one of two high field magnets within ATLAS. The solenoid magnet produces a homogeneous field of 2 T causing the tracks of charged particles to bend, allowing a measurement of their momenta. The largest magnet in ATLAS is a toroidal magnet which produces an inhomogeneous field between 0.5 and 4 T outside the ID. It surrounds the electromagnetic and the hadronic calorimeter. While the electromagnetic calorimeter mainly measures energy deposits from electrons and photons interacting with matter, the hadronic calorimeter measures the energy deposition of hadrons through the same process. Due to the different interactions of photons, electrons and hadrons with matter it is necessary to construct two specialised calorimeters to stop the particles and measure their energy deposition. Muons only interact weakly with matter. They traverse the Inner Detector and the calorimeters and are detected in the muon chambers which form the outermost detector layer in ATLAS.
Neutrinos are not detected within the ATLAS detector due to their negligible interaction cross section with matter. Their presence can be inferred by the missing of transverse energy (E_T^{miss}) due to the conservation of energy. The different components of the detector are described in detail in the next sections.

2.2.1. The Inner Detector

The Inner Detector is composed of three different subdetectors which are all enclosed by the solenoid, providing a homogeneous 2 T magnetic field. Tracks of charged particles are bent in this magnetic field allowing for the measurement of their momentum.
Closest to the beampipe, the high resolution Pixel Detector is placed, followed by the Semiconductor Tracker (SCT) and the Transition Radiation Tracker (TRT). At larger values of $|\eta|$, several disk-like semiconductor trackers are positioned around the beampipe to allow for better $|\eta|$-coverage. A detailed overview and a cross section of the Inner Detector is shown in Figure 2.3. These subdetectors are split into a barrel and end-

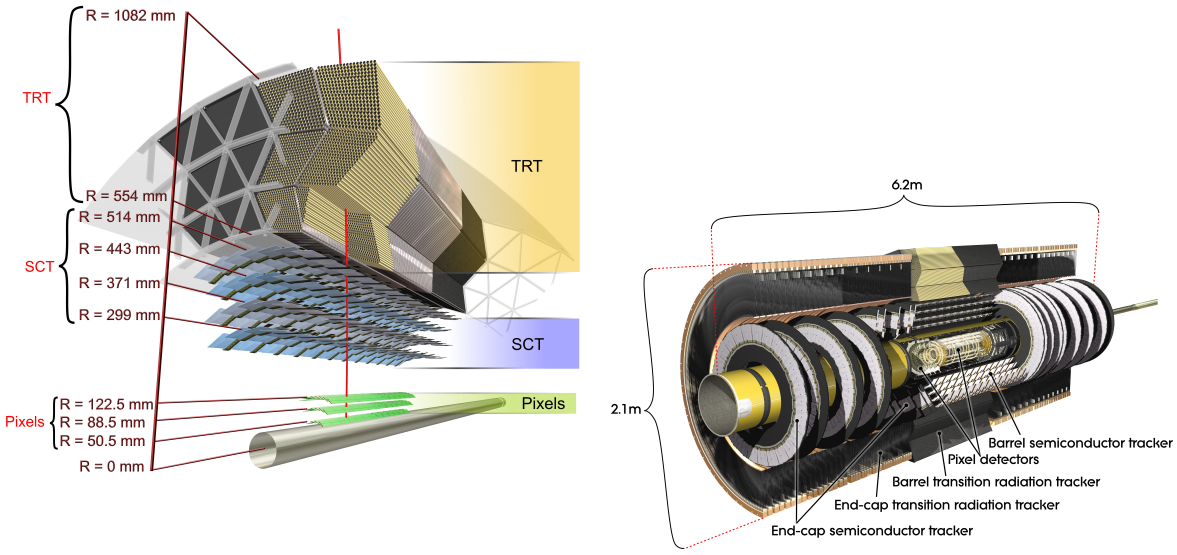

Figure 2.3.: Detailed overview of the Inner Detector (right) and a cross section of all parts within the Inner Detector (left) ©CERN.

cap region. Their main purpose is providing tracks for charged particles and reconstruct primary and secondary vertices.

2.2.2. Calorimeter

ATLAS has two calorimeters, the electromagnetic calorimeter and the hadronic calorimeter. Both calorimeters provide a full coverage in ϕ. The accordion-type electromagnetic calorimeter is a lead-liquid argon (LAr) sampling calorimeter covering an $|\eta|$-range of $|\eta| < 3.2$ using lead as an absorber. It consists of a barrel and an end-cap (EMEC) structure. The calorimeter is used for measuring the energy deposition of photons and electrons.

The non-compensating hadronic calorimeter is a sampling calorimeter that sits behind the electromagnetic calorimeter. It consists of an iron-scintillating tile calorimeter in the range $|\eta| < 1.6$. For rapidity values larger than 1.6, the hadronic calorimeter is a LAr calorimeter. The calorimeter is well suited for measurements of $E_{\mathrm{T}}^{\mathrm{miss}}$ and jet reconstruction using dedicated reconstruction algorithms.

2.2.3. The muon spectrometers

The muon spectrometers form the outermost detector layer of the ATLAS detector since muons traverse the inner subdetectors due to their minimum ionising nature. The momenta of muons are measured in these outermost subdetectors using magnetic field contributions from both magnets. Since muons leave clear signals in the muon spectrometer, they are good trigger candidates.

2.2.4. Triggers

The maximum bunch crossing rate at the ATLAS IP is 40 MHz. To reduce the overwhelming amount of data caused by the proton-proton collisions within ATLAS, a machinery which only saves interesting events is needed.
This task is performed by a two level trigger system which reduces the number of events saved to the order of a few hundred Hertz. The first trigger is a custom made electronic hardware trigger which uses information from subdetectors such as the muon spectrometers and the calorimeters. Candidate events are events containing for example muons, tau particles, jets, missing transverse momentum, electromagnetic clusters or a certain total energy. This trigger reduces the event rate to the order of 100 kHz. These events are then piped to the second trigger level, the High Level Trigger (HLT) [44]. It reduces the event rate further to a manageable event rate of about 1 kHz. Events that pass these triggers are subsequently written to disk.

3. Nominal signal and background Monte Carlo samples

To compare measured data with theory predictions, Monte Carlo (MC) samples are needed. The modelling of signal and background contributions is based on simulated MC events of proton-proton collisions at a centre-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$, produced with MC event generators. The samples that are used in this analysis are presented in detail in this chapter.

After the events are generated using MC event generators, GEANT4 [45] is used to simulate the response of the ATLAS detector [46]. For all samples, except for those generated using the SHERPA [47, 48] MC programme, EVTGEN is used to model the decay of heavy-flavour hadrons [49]. Additional proton-proton interactions are generated with PYTHIA 8 [50, 51]. The MSTW2008LO parton distribution functions (PDF) set [52] and a set of tuned parameters called *A2* is used. To improve the agreement of MC simulation with data, corrections derived from dedicated data samples are applied to the MC simulation. Both inclusive and dedicated samples are used in which the process is either generated at matrix-element (ME) level without explicitly including a photon or by explicitly including photons in the ME level generation step. Dedicated samples are used for the $t\bar{t}\gamma$ process as well as for $W\gamma$ and $Z\gamma$ processes with additional jets. In the inclusive samples, radiation of photons is accounted for by the showering algorithm. The overlap between inclusive and dedicated samples due to possible double-counting of photons is removed.

The dedicated $t\bar{t}\gamma$ signal process is produced at leading order (LO) in QCD using the MADGRAPH5_aMC@NLO generator [53] and the NNPDF3.4LO PDF set [54]. Diagrams where the photon is radiated by charged particles in the initial or final state and by intermediate top quarks or $W^{\pm}$ bosons are included. The sample is normalised to next-to-leading (NLO) accuracy in QCD using the dedicated k-factor used in the analysis. Feynman diagrams of these processes were previously shown in Figure 1.4. These processes also include a small contribution of events that are labelled as $t\bar{t}$ h-fake or $t\bar{t}$ e-fake events in this analysis.

To model parton shower, hadronisation, fragmentation and the underlying event, the

event generation is interfaced to Pythia 8 using the *A14* tune.
Events with $W\gamma$ or $Z\gamma$ final states with additional jets are simulated as dedicated samples with Sherpa [47, 48], where all steps from the hard process to the observable particles are performed by the generator itself. Furthermore, the Sherpa-internal parton showering matches and merges all samples. While $W\gamma$ events are simulated at NLO accuracy in QCD, $Z\gamma$ events are simulated at LO in QCD. Both times the NNPDF3.0NNLO PDF set is used and the samples are normalised to the cross section given by the corresponding MC simulation.
To simulate events with inclusive $W^{\pm}$ boson and Z^0 boson production in association with additional jets, Sherpa [47, 48] is used. These samples are simulated at NLO in QCD using the NNPDF3.0NLO PDF set in association with a dedicated tune provided by the Sherpa authors. These inclusive samples are normalised to next-to-next-to-leading order (NNLO) in QCD [36].
Single top processes ($Wt\gamma$) are simulated separately for the three production channels introduced in Chapter 1. In addition, each channel is generated for top quark and anti-top-quark separately at NLO level using POWHEG and the NNPDF2.3LO PDF set which is interfaced to Pythia using the *A14* tune. The calculated cross section of the sample is normalized to NNLO in QCD [55].
Inclusive $t\bar{t}$ production processes are simulated at ME level at NLO accuracy in QCD. For simulating this sample, POWHEG-BOX-v2 [56–59] is used. The calculation uses the NNPDF3.0NLO PDF set [60]. The parton shower for this sample is also generated with Pythia 8 using the *A14* tune. The normalisation of the $t\bar{t}$ samples is based on the cross section calculated with the Top++2.0 programme at NNLO in perturbative QCD (see [61] and references therein). The inclusive $t\bar{t}$ production processes are the dominant source for $t\bar{t}$ h-fake and $t\bar{t}$ e-fake events in this analysis.
Diboson events, events with two vector bosons such as WW, WZ or ZZ, are generated with Sherpa at LO in QCD. The simulation uses the NNPDF3.0NNLO PDF set and a dedicated tune provided by the Sherpa authors. The diboson samples are normalised to NLO accuracy cross sections in QCD [62].
The $t\bar{t}$ production process in association with a vector boson, the $t\bar{t}V$ process as introduced in Chapter 1, is simulated at NLO ME level with Madgraph5_aMC@NLO. The NNPDF3.0NLO PDF set is used and the generator is interfaced to Pythia 8. The *A14* tune is used in association with the NNPDF2.3NLO PDF set. The $t\bar{t}V$ samples are normalised to NLO in electroweak theory and QCD [63].
Samples that are produced at NLO or NNLO accuracy in QCD as presented in this chapter can contain a number of events with negative event weights. While these events

originate from physics effects, their treatment in the training process of a neural network is non-trivial. The treatment of these events will be revisited in Chapter 5.

4. Event selection and object definition

In this chapter the definition of photons, electrons, muons and jets is discussed. Based on these reconstructed objects selections are placed, called the event selection. The event selection suppresses background contributions and thus enriches the purity of $t\bar{t}\gamma$ within the region defined by it. The strategy of the event selection is loosely based on the event selection required for a typical $t\bar{t}$ analysis but with additional requirements on a photon. All reconstructed photons [64] pass the pseudo-rapidity cut of $|\eta| < 2.37$ which is motivated by the efficiency drop of the electromagnetic calorimeter for higher values of $|\eta|$. Due to the crack region of the electromagnetic calorimeter photons with $1.37 < |\eta| < 1.52$ are excluded. The photon candidates must have a transverse energy of $E_{\mathrm{T}} > 20\,\mathrm{GeV}$. They must pass both, the tight identification criteria [65, 66] and isolation criteria defined as $E_{\mathrm{T}}^{\mathrm{topocone40}} < 0.022\,p_{\mathrm{T}} + 2.45\,\mathrm{GeV}$ as well as $p_{\mathrm{T}}^{\mathrm{cone20}} < 0.05\,p_{\mathrm{T}}$. The isolation variable $E_T^{\mathrm{topocone40}}$ describes the amount of energy measured by the calorimeter in a cone of radius $\Delta R = 0.4$ around the particle. The region of $\Delta R < 0.1$ is excluded. The second isolation variable, p_T^{cone20}, describes the amount of transverse momentum associated with charged particle tracks in a cone of radius $\Delta R = 0.2$ around the particle with $p_{\mathrm{T}} > 1\,\mathrm{GeV}$. Requiring the photons to be isolated reduces hadronic background processes where soft photons are emitted (h-fakes, $t\bar{t}$ h-fakes).

Photons can be either converted or unconverted. While a converted photon is reconstructed from a calorimeter cluster matched to a conversion vertex or several vertices, an unconverted photon is made from a cluster matched to neither an electron track nor a conversion vertex.

To distinguish prompt photons in $t\bar{t}$ final states from fake photons and other prompt photons being radiated in final states involving $W^{\pm}$ or Z^0 bosons, Monte-Carlo (MC) truth information is used.

Muon identification and reconstruction in ATLAS [67] is based on information obtained with the muon spectrometers and the calorimeters. Muons are identified using information related to track measurements and measured distances from the IP. Muon candidates

have to pass a medium identification and *FCTight_FixedRad* isolation criteria [67]. The transverse momentum of the muon must be $> 25\,\text{GeV}$ for data taken in 2015, 2016, 2017 and 2018. Muon candidates must also fulfil $|\eta| < 2.5$.
Electron candidates must be within $|\eta| < 2.47$ excluding the calorimeter crack region at $1.37 < |\eta| < 1.52$. They must have a calibrated transverse momentum of $p_{\text{T}} > 25\,\text{GeV}$.
Jet candidates are reconstructed by using the anti-k_{T} algorithm [68]. A radius parameter of $R = 0.4$ is used. All jets are required to have a transverse momentum of $p_{\text{T}} > 25\,\text{GeV}$ and to be in the pseudo-rapidity range of $|\eta| < 2.5$. A calibrated b-tag working point of 85 % is used using the MV2C10 algorithm [69–71].
This analysis uses single-lepton triggers for electrons and muons. The selection targets final states with at least four jets, exactly one isolated photon and either an electron or a muon. Using the invariant mass of the reconstructed photon and the lepton, $m(\ell, \gamma)$, a Z^0-window cut $|m_Z - m(\ell, \gamma)| > 5\,\text{GeV}$ is applied in the $e+$ jets channel. The mass of the Z^0 boson is assumed to be $m_Z = 91.23\,\text{GeV}$. Applying the Z^0-window cut reduces the contamination of electronic fakes coming from $Z \to \ell\ell$ final states where one lepton is misidentified as a photon. To reduce the contribution of photons that are radiated by leptons, events with $\Delta R(l, \gamma) > 1.0$ are selected, where $\Delta R(l, \gamma)$ measures the distance between the lepton and the photon. The event selection is summarised in Table 4.1 below. For this analysis conservative normalisation uncertainties of 50 % are used for all

Table 4.1.: Table summarising the event selection of the analysis.

Channel	$e+$ jets	$\mu+$ jets
Photon	1 γ with $p_{\text{T}} > 20\,\text{GeV}$	
Jets	≥ 4	
b-jets	≥ 1 using the 85 % working point	
$m(\gamma, \ell)$	$\lvert m(\gamma, \ell) - m_Z \rvert > 5\,\text{GeV}$	
$\Delta R(\gamma, \ell)$	> 1.0	

backgrounds because a dedicated determination of systematic uncertainties was beyond the scope of this thesis. A 20 % normalisation uncertainty is assigned to the $t\bar{t}\gamma$ signal based on the theoretical uncertainty on the NLO k-factor used in a previous ATLAS measurement of the $t\bar{t}\gamma$ cross section [72]. A luminosity uncertainty of 1.7% is used [73]. Furthermore, no background scale factors are applied.
Applying the event selection described above results in a signal-to-background ratio of approximately 1 : 1. The details of the resulting event composition in the signal region are discussed in the following paragraphs.

The largest background in the $e+$ jets and $\mu+$ jets channel is the contribution from $t\bar{t}$ e-fakes which makes up around 12.5% of the combined signal and background MC. In the $\mu+$ jets channel, $t\bar{t}$ e-fakes are created by the dileptonic $t\bar{t}$-decay in the $e\mu$ channel with extra jets where the electron is misidentified as a photon. The contribution of $W\gamma$ is 12.0% in the $\mu+$ jets channel and 11.7% in the $e+$ jets channel. It ranks second behind the $t\bar{t}$ e-fake contribution. The third highest contribution in both channels comes from $t\bar{t}$ h-fake events. The contributions are 6.0% in the $e+$ jets channel and 6.5% in the $\mu+$ jets channel. $Z\gamma$ events contribute with 4.75% in the $e+$ jets channel and with 3.45% in the $\mu+$ jets channel. Single top events, $Wt\gamma$, make up 3.9% in the $e+$ jets channel and 4.3% in the $\mu+$ jets channel. Electronic fake events make only 0.7% in the $\mu+$ jets channel but 5.2% in the $e+$ jets channel where they originate from a Z^0 decay with two electrons in the final state, one of which is misidentified as a photon. The smallest contributions come from h-fake and Promptγ backgrounds that make up 0.9% and 0.5% in both channels, respectively.

Table 4.2 shows the event yields of the signal and the background in both channels as well as in the combined channel. The total uncertainty as a combination of the systematic uncertainties previously mentioned and statistical uncertainties is also shown. Figure 4.1

Table 4.2.: Yields for all background processes and the signal for each channel. Scale factors for the different backgrounds are not applied.

	$\mu+$ jets	$e+$ jets	combined
$t\bar{t}\gamma$	$13\,700 \pm 2700$	$11\,800 \pm 2300$	$26\,000 \pm 5000$
$t\bar{t}$ e-fake	2900 ± 1400	2900 ± 1400	5700 ± 2900
$W\gamma$	2800 ± 1400	2700 ± 1300	5500 ± 2700
$t\bar{t}$ h-fake	1500 ± 700	1400 ± 700	2900 ± 1500
$Z\gamma$	800 ± 400	1100 ± 600	2000 ± 1000
$Wt\gamma$	1000 ± 500	900 ± 500	1900 ± 900
e-fake	160 ± 80	1200 ± 600	1400 ± 700
h-fake	200 ± 100	200 ± 100	400 ± 200
Promptγ	120 ± 60	120 ± 60	240 ± 120
Total	$23\,200 \pm 3500$	$22\,400 \pm 3300$	$46\,000 \pm 7000$
Data ($139\,\mathrm{fb}^{-1}$)	24 430	25 993	50 423

shows the event composition of the $e+$ jets channel and the $\mu+$ jets channel in detail. Since the composition of backgrounds for both, the $e+$ jets and the $\mu+$ jets channels is similar, these channels are merged and hereafter referred to as the (combined) single-lepton channels. No further differentiation between electrons and muons is performed. The total signal contribution within the lepton+jets channel is 56 %. Backgrounds with

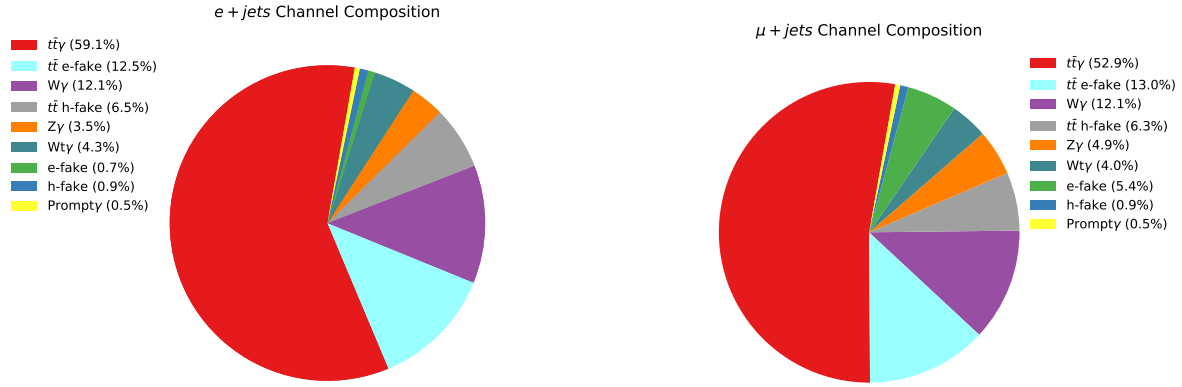

Figure 4.1.: Event composition of the $e+$ jets channel (left) and the $\mu+$ jets channel (right).

fake photons (e-fake, h-fake, $t\bar{t}$ e-fake and $t\bar{t}$ h-fake) make up 22.9 % of all events and backgrounds with prompt photons ($W\gamma$, $Z\gamma$, $Wt\gamma$ and Promptγ) make up 21.1 % off all events. Overall the background contribution of $t\bar{t}$ e-fake events ranks highest with 12.5 % followed by the $W\gamma$ contribution with 12.1 %.

5. Machine learning in the scope of a $t\bar{t}\gamma$ cross section measurement in ATLAS

In recent years, machine learning (ML) approaches and new developments became popular throughout ATLAS, CMS and the high-energy-physics community. Several different analyses have improved their precision by utilizing machine learning tools like boosted decision trees or neural networks (NNs) to discriminate signal processes from background processes [74–76]. The discriminating power of NNs has significantly increased due to the development of new learning techniques which use computational resources more efficiently.

A neural network is one of several different approaches to machine learning. Even though its design is motivated by the biological function of the human brain, its mathematical description differs from the working principles of neurons and synapses within the human brain.

In this analysis, the concept of a uni-directional neural network in which information is fed from an input towards an output, a so called feed-forward NN, is used. All feed-forward neural networks are composed of an input layer and an output layer with several intermediate, hidden layers in between. Each of these layers consists of a variety of nodes. The depth of the network is defined by the number of hidden layers and the width is defined by the number of nodes in these layers. Combining the width and depth of all layers defines the size of the network. Deep neural networks are neural networks that consist of multiple hidden layers between the input and the output layer.

The goal of any ML approach is to find a function that transforms an n-dimensional input describing a complex parameter space or set of variables A, into simpler space or set B[1]. In classification problems, the aim is to transform the complex n-dimensional input defined by the input variables, thereby separating the events of each sub-class within the dataset. In case of an input vector of low dimensionality, the relationships between the

[1] In this analysis: $\dim(B) = 3$

components of the vector are simple.
This is not necessarily the case for higher dimensions and for the output distributions of the hidden layers due to the complexity and multidimensionality of the problem at hand. Each node within each hidden layer is fully connected to each node in the previous and next layer. Each node then performs a transformation of its input value, $\tilde{x}$, using a specific weight, w_i, describing the connection to the previous node i as a scalar and its own bias b as an offset, in a transformation $T(x) = g(w_i \cdot \tilde{x} + b)$ using an activation function g. The activation function introduces complex non-linear mappings between input and output values, thereby providing the possibility to describe non-linear relations within the dataset. The transformation of an input vector composed of input variables, x_i, $\vec{x} \in \mathbb{R}^n$ into an output vector $\vec{y} \in \mathbb{R}^m$ is then described by a non-linear function combining all transformations $T(x)$ from all layers using k total weights and l total biases:

$$f : \mathbb{R}^n \to \mathbb{R}^m$$
$$\vec{x} \to f(\vec{x}, w_1, w_2, \dots w_{k-1}, w_k, b_1, b_2, \dots, b_{l-1}, b_l) = \vec{y}$$

The function $f(\vec{x}, w_1, w_2, \dots w_{k-1}, w_k, b_1, b_2, \dots, b_{l-1}, b_l)$ therefore assigns a label or a set of labels to each event. In the case of a binary classification problem $m = 1$ is usually sufficient. For a multi-class classification problem with three classes, $m = 3$ is used. This function effectively represents a multi-dimensional hyperplane of dimension $n - 1$, that defines a boundary hyperplane within the hyperspace defined by the n input parameters, thereby separating events. The properties of this function are difficult to grasp and need to be examined in detailed numerical studies.
To find the optimal function to separate the subclasses a loss or cost function is optimised that describes the performance of the current model. The weights and biases of each node are adapted using a method called *the backpropagation of errors* [77]. For binary classification the *binary-cross-entropy* is used (see Equation (A.2)) as the loss function in this analysis. For multi-class classification the *categorical-cross-entropy* is used as described in Equations (A.1) and (A.2) in the appendix.
When training a classifier, training data is needed. This data set is called *training set*. When the entire training dataset was passed forward through the network and the errors were passed backwards through it, one training step named *epoch* is passed. Each epoch is further divided into several *batches* that are passed to the NN one by one[2]. This set is repeatedly used during every epoch of training and directly influences the weights and

[2]The size of these batches is only limited by the available capacity of CPU memory. Larger batch sizes require less backpropagation steps since the errors are only propagated backwards at the end of each batch.

biases of each node in each layer of the network.
To measure the performance of any classifier, metrics need to be defined. These metrics need to be calculated using events that the network was not previously trained on. This dataset is called the *testing set*. The testing of the classifier is performed after the training is completed. Comparing the performance of the classifier using both the training and the testing set provides a way of quantifying the generalisation capability of the classifier. To determine the evolution of the performance of the classifier during training, a third statistically independent set needs to be defined. This third set is the *validation set*. It is drawn from the initial training set and used to calculate the validation metrics. Events in the validation set are not used for training, effectively reducing the training set in size. The validation loss and the validation accuracy are used as measures for the goodness of the model on data points the classifier was not previously trained on during training. These metrics describe the value of the used loss function and the weighted accuracy of the classifier to correctly predict these events calculated only using the validations set. During the training process, the validation loss is monitored and training is aborted when the validation loss stops improving.
If training is stopped before reaching the minimum of the loss function, the classifier is defined as being undertrained. It "underfits" the training data. If training is continued over too many epochs and the training loss starts to diverge from the validation loss, generalisation power is lost. The classifier is said to be overtrained. It is "overfitting" the training data.
The size of the training, validation and testing set is chosen to be 60 %, 20 % and 20 % of the entire dataset, respectively.
The high discrimination power of deep multi-class neural networks motivates a development of a such a supervised NN to discriminate $t\bar{t}\gamma$ processes from background processes in this analysis.
In a previous measurement of the $t\bar{t}\gamma$ cross section in ATLAS using proton-proton collision data taken in 2015 and 2016 a binary approach was used to separate $t\bar{t}\gamma$ final states from the underlying background [72]. In this thesis, a deep multi-class neural network is developed to discriminate $t\bar{t}\gamma$ signal processes from processes involving hadronic fakes or electronic fakes and events that mimic the $t\bar{t}$ topology with a prompt photon. The names of these classes are defined as "$t\bar{t}\gamma$" for the $t\bar{t}\gamma$ signal process with a prompt photon, "*Fakes*" for all processes involving hadronic or electronic fakes and "*Other Prompt*" for all processes involving prompt photons that do originate from a top quark pair. Events that are separated by such a multi-class classifier are therefore grouped into sub-categories

such that

$$\vec{y} = \begin{cases} (1,0,0)^t & \text{for} \quad \text{Event} \in t\bar{t}\gamma \\ (0,1,0)^t & \text{for} \quad \text{Event} \in \textit{Fakes} \\ (0,0,1)^t & \text{for} \quad \text{Event} \in \textit{Other Prompt.} \end{cases} \tag{5.1}$$

Fake-lepton events, events where a lepton is faked by a jet or a photon, are not investigated within the scope of this analysis since their impact was found to be small in a previous $t\bar{t}\gamma$ cross section measurement by ATLAS [72].

6. Development of multi-class classifiers for the $t\bar{t}\gamma$ cross section analysis

For the development of all classifiers in this analysis, the open source machine learning libraries *Tensorflow* [78] and *Keras* [79] are used. Furthermore, several components of the analysis are based on additional functions provided by the *scikit-learn* machine learning library [80]. The Python extension module *root_numpy* is used to provide an efficient interface between *ROOT* [81] and *NumPy* [82] which is then used for training *Keras*-based neural networks. All classifiers are trained using events passing the event selection described in Chapter 4.

To utilize all events for training and testing, a 5-fold cross validation is used. For each fold the validation set is randomly drawn from all events within the training set of that fold. The size of the validation set is 25 % of the size of the folds training set. Its statistics are therefore comparable with the statistics of the testing set.

Due to the choice of cross validation, the number of events used for testing is equivalent to a fifth of all events passing the event selection for each tested classifier. To minimise the loss function, the *Nadam* optimizer is used, which incorporates Nesterov momentum into *Adam* [83–85].

To improve the generalisation performance, the validation is monitored and the training is stopped if the decrease in validation loss is smaller than 0.0001 for 300 epochs. The batch size is chosen to be 100000 training events. Additionally, *dropout* layers are used as a regularisation measure. These layers set a fraction (30 %) of inputs units to 0 at each update during the training time thereby preventing overfitting [86].

As a performance measure the area under the receiver operating characteristics curve (AUROC) is used, where in the case of a multi-class output the AUROC is calculated for each class separately. Furthermore, the shape of the receiver operating characteristics (ROC) curve is used to investigate the generalisation performance of the classifier. For these ROC curves the false-positive-rate (FPR) and the true-positive-rate (TPR) are used.

Events with negative event weights are not considered during training since their physical meaning is not reflected by their negative impact on the loss function. These event weights are subsequently set to zero for all training purposes. These events originate from samples that are calculated at NLO and NNLO previously presented in Chapter 3. The impact of this decision on the training is checked by comparing the shape of the input variables with and without negative event weights. The differences of the kinematic distributions were found to be below 0.1 % for most input variables and backgrounds and are therefore considered negligible.
The absence of all negative event weights is taken into account by scaling all event weights accordingly to preserve the ratios of expected events between all classes. To handle the imbalance between classes used during training, the MC training weights are rescaled to form the training weights $w_{\text{train},i}$ by applying scale factors c_i and c_j, for $i, j \in [1, 2, 3]$ in the multi-class and $i, j \in [1, 2]$ in the binary case, such that $c_i \cdot \sum w_{\text{train},i} / c_j \cdot \sum w_{\text{train},j} = 1$ for all permutations of i and j and $i \neq j$.
For the training, all input variables are scaled into a range between zero and one, thereby eliminating large spreads of values between different input variables, which may otherwise result in large error gradient values causing weight values to change significantly, leading to an unstable learning process. Using this scaling method the shape of all kinematic distributions is preserved.

6.1. Building a stacked model

One general purpose of the multi-class classifiers developed for the $t\bar{t}\gamma$ analysis is to separate $t\bar{t}\gamma$ events from underlying background events. By doing this, two different improvements of the analysis can be achieved. Firstly, the separation of $t\bar{t}\gamma$ signal events from underlying background events enables the definition of new cuts based on the output of the used multi-class classifier, thereby improving the signal over background ratio and the significance of a cut and count based analysis. Secondly, the non $t\bar{t}\gamma$ output nodes can be used to define control regions for $W\gamma$, $Z\gamma$, $t\bar{t}$ e-fake, e-fake, $t\bar{t}$ h-fakes and h-fake events which can be used to constrain uncertainties on these types of backgrounds. The kinematics of these control regions are very similar to the kinematics of the remaining orthogonal signal region and should therefore, especially in the case of $W\gamma$ events, be preferred over light flavour control regions.
Two different multi-class classifiers are developed by using stacked generalisation ensembles of neural network models.
These dedicated multi-class classifiers consist of several sub-classifiers (SCLs). They are

trained using a procedure consisting of two main steps. First, all SCLs are trained and afterwards combined in a stacked generalisation ensemble. A multi-class classifier is then connected to the output of all sub-classifiers. In a second step, the weights and biases of all hidden layers of all SCLs are fixed and the training of the entire stacked generalisation ensemble is repeated. Only the additional layers added by the multi-class classifier are affected during this final step.
Two different approaches on defining these SCLs are pursued. While the one-vs-one approach uses SCLs that directly discriminate the $t\bar{t}\gamma$ signal from each background process separately, the SCLs in the one-vs-all approach focus on determining probabilities for each background and the signal separately.

6.1.1. Variable selection

The two different multi-class classifiers and all their SCLs are trained with an identical set of input variables. A detailed list of all used input variables is given in Section 6.1.1. Kinematic distributions of all used variables are shown on Figures C.1 to C.4 in the appendix. The choice of input variables is predominantly motivated by the differences

Relation	Variables (31 in total)
Photon related	photon $\lvert\eta\rvert$, photon p_{T}, photon conversion type, $\Delta R(\gamma,\mathrm{jet})_{\mathrm{min},1}$, $\Delta R(\gamma,\mathrm{jet})_{\mathrm{min},2}$, $\Delta R(\gamma,\mathrm{jet})_{\mathrm{min},3}$, $\Delta R(\gamma,\mathrm{jet})_{\mathrm{min},4}$, $\Delta R(\gamma,\mathrm{jet})_{\mathrm{avg}}$
b-tagging related	leading b-tag score, subleading b-tag score, Number of $b-$jets
Jet related	Number of jets, leading jet p_{T}, sub-leading jet p_{T}, $\mathrm{Jet}_{\gamma,1}$ p_{T}, $\mathrm{Jet}_{\gamma,2}$ p_{T}, $\mathrm{Jet}_{\gamma,3}$ p_{T}, $\mathrm{Jet}_{\gamma,4}$ p_{T}
Mixed	$m(l,b)$, $m(\gamma,b)$, $\Delta R(\gamma,b)$, $m(l,\gamma)$, $\Delta R(l,\gamma)$, Leading b-jet p_{T}, Sub-leading b-jet p_{T}, Third-leading b-jet p_{T}, Fourth-leading b-jet p_{T}
Others	$p_{\mathrm{T}}(\mathrm{W})$, H_{T}, $m_T(\mathrm{W})$, $E_{\mathrm{T}}^{\mathrm{miss}}$

Table 6.1.: The selected variables that were used as input variables for the neural networks.

in topologies highlighted by them. It is based on an in-depth study of their separation power.
The resulting list of variables is based on comparisons taking $t\bar{t}\gamma$, $t\bar{t}$ e-fake, $t\bar{t}$ h-fake, e-fake, h-fake, $Wt\gamma$, $W\gamma$, $Z\gamma$ and Promptγ event topologies into account. Combinations of background-versus-background or signal-versus-background topologies are investigated

totalling 36 unique combinations. For each combination the separation power for all investigated input variables is calculated, yielding a dedicated ranking of input variables for every combination. To calculate the separation power S, the function

$$S = \frac{1}{2}\frac{\sum_i^n (S_i - B_i)^2}{S_i + B_i}. \tag{6.1}$$

is used, where S_i describes the number of signal events in bin i and B_i describes the number of background events in bin i.

As an example, the ranked separation power of the 15 highest ranking variables based on the study of the $t\bar{t}\gamma$ event topology and the $W\gamma$ event topology are shown in Figure 6.1. The background distributions are compared to the $t\bar{t}\gamma$ distributions. Additional summary

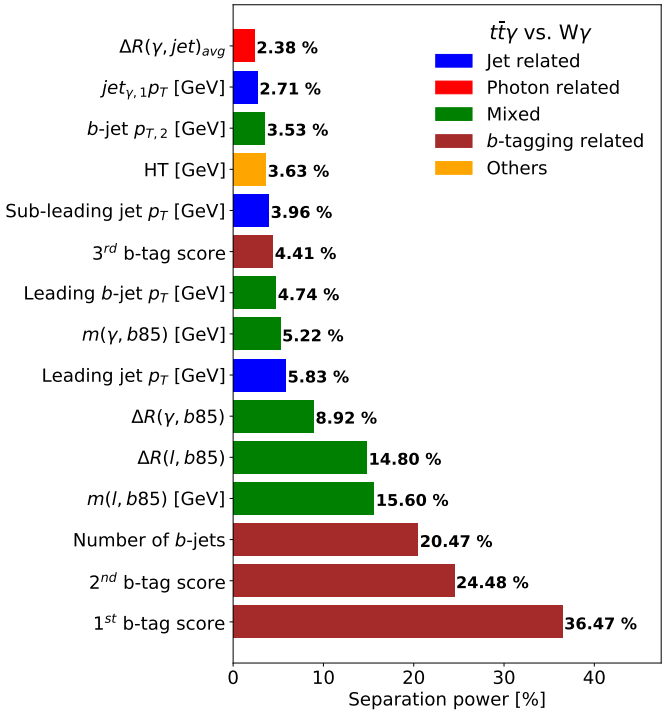

Figure 6.1.: Ranked separation power plots of the input variables based on the study of the $t\bar{t}\gamma$ topology and the $W\gamma$ event topology

plots are shown in Figures C.5 to C.7[1] in the appendix.

Due to the high separation power of b-tagging related variables for separating $W\gamma$ and $Z\gamma$ events from $t\bar{t}\gamma$ events and the similarity of the ranking of the input variables for these two types of events, $W\gamma$ and $Z\gamma$ events are combined into one class called $WZ\gamma$.

Furthermore, variables incorporating b-tagging information are included to utilize their

[1]These figures contain additional separation rankings on which the following results are based. In total 18 out of 36 additional rankings are shown.

separation power of prompt processes such as $W\gamma$ and $Z\gamma$ from the $t\bar{t}\gamma$ signal process. Both $W\gamma$ and $Z\gamma$ processes have a lower number of jets tagged as b-jets and are therefore easily separated from the $t\bar{t}\gamma$ signal process. The same statement can be made when comparing $W\gamma$ and $Z\gamma$ events with $t\bar{t}$ e-fake and $t\bar{t}$ h-fake events, since both $t\bar{t}$ h-fake and $t\bar{t}$ e-fake events also have two top quarks that decay into b quarks, leading to significant differences in the kinematic distribution of said variables. Especially the pseudo-continuous b-tagging scores provide high separation power in these cases and are therefore used. Distributions of the highest and second highest pseudo-continuous b-tagging scores are shown in Figures 6.2 and 6.3 for $Z\gamma$, $W\gamma$ and $t\bar{t}\gamma$ events as well as $t\bar{t}$ e-fake and $t\bar{t}$ h-fake events. Additionally, the distance between the lepton and the closest b-jet in the η-ϕ

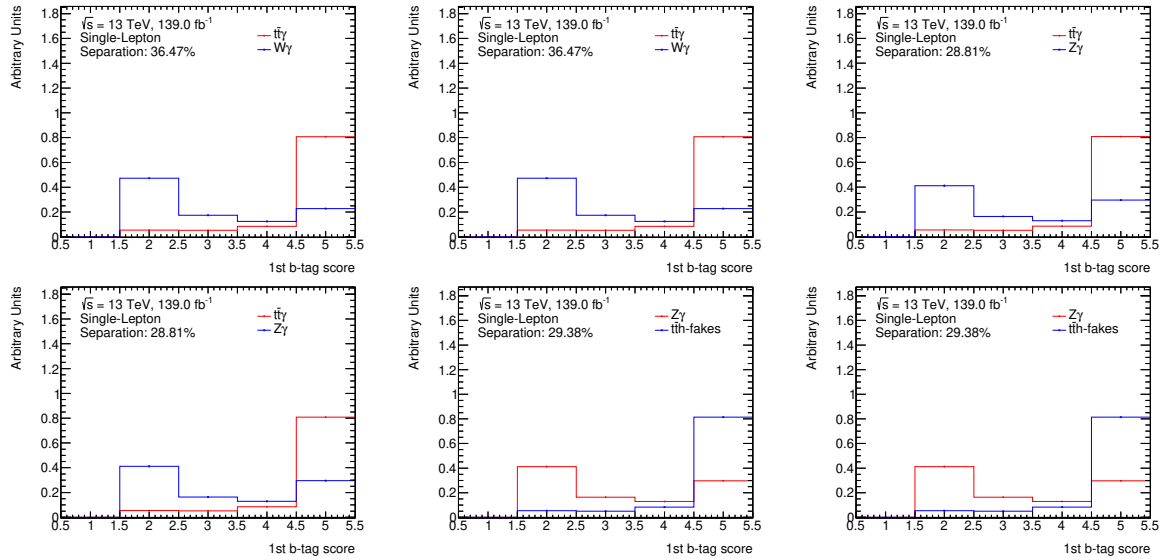

Figure 6.2.: Kinematic distributions of the leading b-tag score depicting the separation of $t\bar{t}\gamma$ events from $W\gamma$ and $Z\gamma$ events.

plane, $\Delta R(l, b)$, as well as the invariant mass of the lepton and the closest b-jet, $m(l, b)$, have high separation power and provide a good handle to further separate events. Distributions of these two variables are shown in Figures 6.4 and 6.5 for $t\bar{t}\gamma$, $Z\gamma$ and $W\gamma$ events as well as $t\bar{t}$ e-fake and $t\bar{t}$ h-fake events. Since shower shape variables obtained with the electromagnetic calorimeter show discrepancies between MC and data for the lateral profile development of the electromagnetic shower [66, 87], they are not used as input variables to discriminate h-fakes, $t\bar{t}$ h-fakes, e-fakes and $t\bar{t}$ e-fakes from other events. Therefore, several variables are defined instead, describing the proximity of the reconstructed photons and jets. While variables such as $\Delta R(\gamma, \text{jet})_{\text{min},i}$ describe the distance between the photon and a jet, where $i = 1$ describes the closest jet ($(i = 2) \hat{=}$ second closest jet, et cetera), $\text{jet}_{\gamma,i} p_{\text{T}}$ describes the transverse momentum of said jet. Distributions

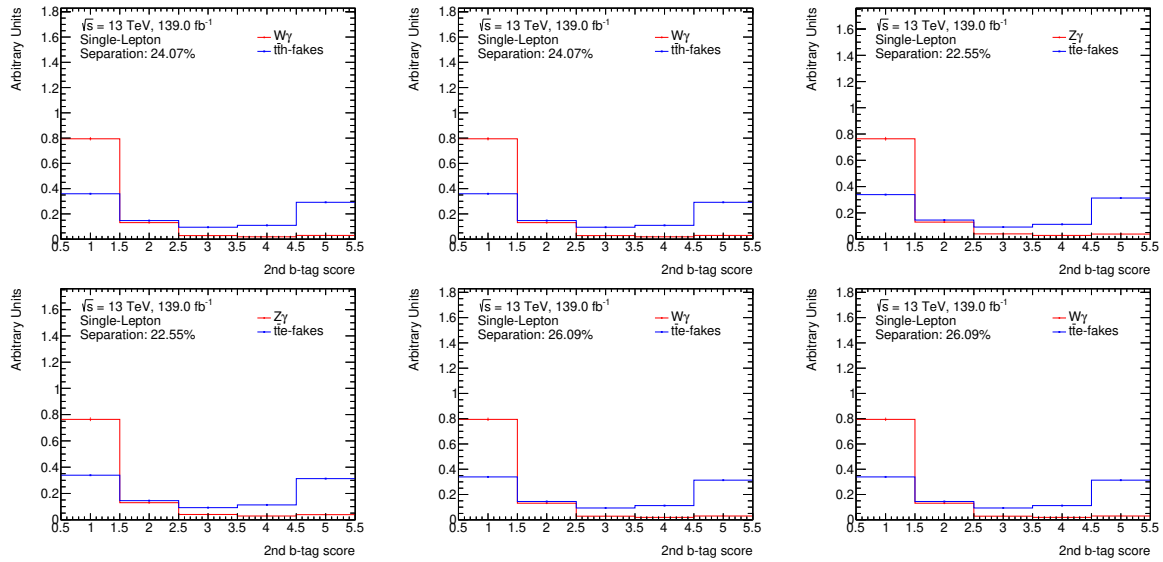

Figure 6.3.: Kinematic distributions of the sub leading b-tag score depicting the separation of $t\bar{t}\gamma$ events from $W\gamma$ and $Z\gamma$ events.

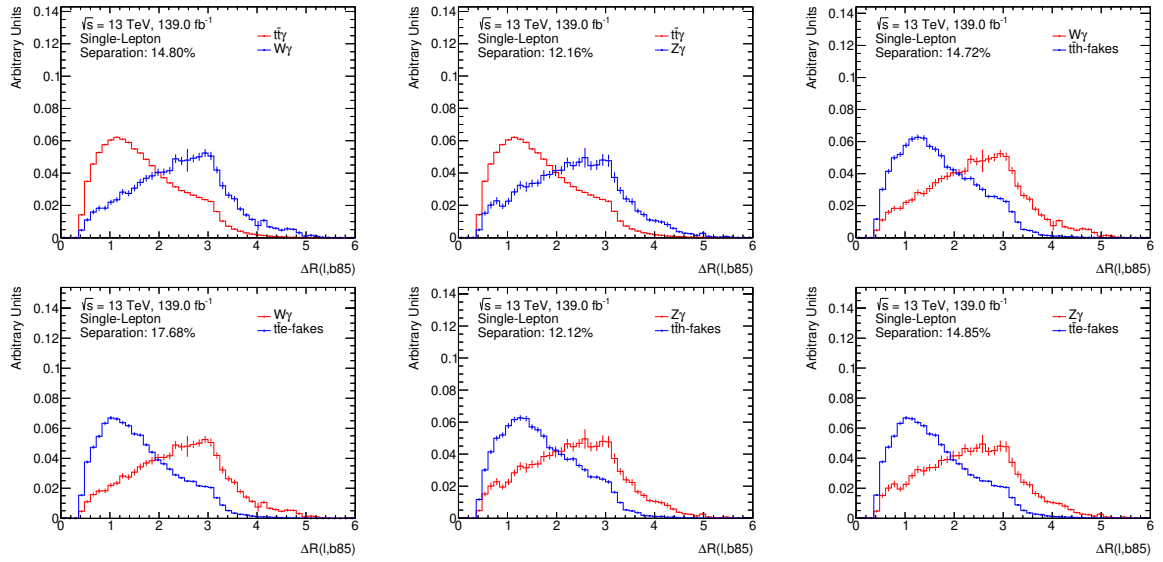

Figure 6.4.: Kinematic distributions of $\Delta R(l, b)$ depicting the separation of $t\bar{t}\gamma$ events from $W\gamma$ and $Z\gamma$ events as well as the separation of $W\gamma$ and $Z\gamma$ from $t\bar{t}$ e-fake and $t\bar{t}$ h-fake events.

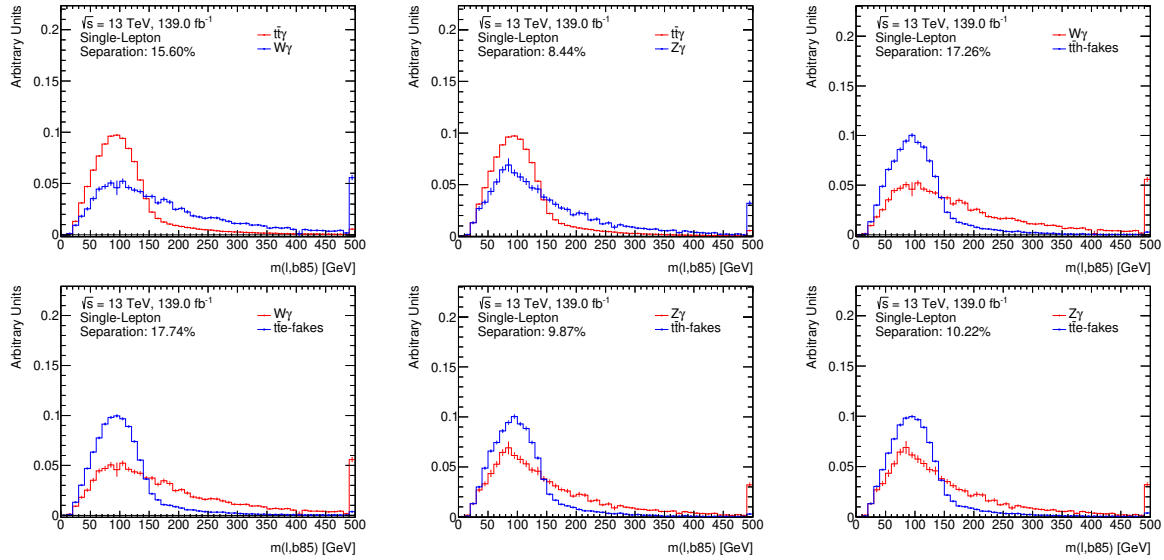

Figure 6.5.: Kinematic distributions of $m(l, b)$ depicting the separation of $t\bar{t}\gamma$ events from $W\gamma$ and $Z\gamma$ events as well as the separation of $W\gamma$ and $Z\gamma$ from $t\bar{t}$ e-fake and $t\bar{t}$ h-fake events.

for $\Delta R(\gamma, \text{jet})_{\text{min},1}$, $\Delta R(\gamma, \text{jet})_{\text{min},2}$ and $\Delta R(\gamma, \text{jet})_{\text{min},3}$ are shown in Figure 6.6.

To identify e-fake and $t\bar{t}$ e-fake events, variables such as the photon conversion type,

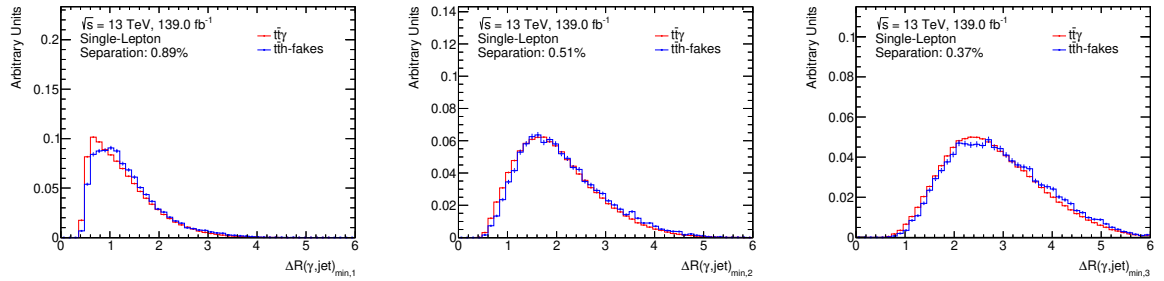

Figure 6.6.: Kinematic distributions of $\Delta R(\gamma, \text{jet})_{\text{min},i}$, for $i \in [1, 3]$, depicting the separation of $t\bar{t}\gamma$ events from $t\bar{t}$ h-fake events.

$E_{\text{T}}^{\text{miss}}$, and $m(l, \gamma)$ are used, highlighting kinematic differences caused by the electron misidentified as a photon as well as differences related to it coming from a dileptonic Z^0 boson decay.

Separation plots for the kinematic distributions of these variables for $t\bar{t}$ e-fake and e-fake events as well as for $t\bar{t}\gamma$ events are shown in Figures C.8 and C.9. Additionally, separation plots comparing $t\bar{t}$ h-fakes and $t\bar{t}$ e-fakes as well as e-fakes and h-fakes are also shown for reference.

6.1.2. Optimisation of sub-classifiers

As described in Section 6.1, both the one-vs-one and the one-vs-all multi-class classifier are based on a stacked approach. To optimise the performance of each multi-class classifier, its corresponding set of hyper-parameters is also optimised. For this purpose, an optimisation algorithm is developed and applied to all SCLs that are fed as inputs to the multi-class classifiers. A separate hyper-parameter optimisation of the multi-class classifiers themselves is not performed.

Genetic optimisation algorithm

Since the search for an optimal set of hyper-parameters is non-trivial and computationally expensive, several possible configurations are tested simultaneously using grid resources to effectively scan a wide space of different hyper-parameters. The method of testing each possible combination of hyper-parameters is not feasible for large sets of different hyper-parameter combinations. In such cases, it can be shown that random approaches are more efficient than standard search approaches [88]. Therefore, a genetic optimisation algorithm (GOA) is developed and applied to reduce the computation time by reducing the set of hyper-parameters that is actually tested. The GOA is then executed using the open-source high-throughput computing software framework *HTCondor* in a diamond shaped directed acyclic graph (DAG) application.
The set of hyper-parameters tested in the GOA focuses on the size of network, its depth, the used activation functions and the layers where dropout is applied because these parameters are assumed to have the highest impact on the performance. For the size of a layer 30, 50 and 100 nodes are considered with a maximum depth of 2,3 or 5 layers in total. The tested activation functions are *ReLU*, *ELU* and the $\tanh(x)$ function (see Equations (A.3) to (A.5) in the appendix), where each layer can have any of the three activation functions. Dropout layers are randomly added in between the hidden layers. The cardinality of the set of possible combinations, $\mathrm{card}(M)$, can be calculated using

$$\mathrm{card}(M) = \underbrace{\sum_{i \in \{2,3,5\}}}_{\text{Number of layers}} 3^i \cdot 3^i \cdot \underbrace{\sum_{j=1}^{5} \frac{i!}{j!(i-j)!}}_{\text{Dropout configurations}} = 1,835,865, \tag{6.2}$$

where $3^i \cdot 3^i$ describes the number of possible permutations of nodes and activation functions based on the three possible values for each. The algorithm is initialised by defining an arbitrary set of i SCLs by randomly picking a subset of hyper-parameters from the set M with $\mathrm{card}(M) = 1,835,865$ different configurations.

This initial set of configurations forms the first generation of sub-classifiers. All SCLs within the generation are then trained using identical splits for the training set and a 5-fold cross validation. In the GOA, the validation set is composed of events that are randomly drawn from the training set.
After all SCLs are trained, they are tested and compared using identical testing sets. All classifiers within one generation are then sorted based on the achieved AUC values calculated using both the training and testing set. The top forty percent of all tested classifiers are retained and serve as parent networks for the next generation. Furthermore, additional SCLs that are not within the top forty percent are randomly chosen and also retained for the next generation. From this set, pairs are randomly combined. Each pair is then used to breed *children networks* consisting of randomly chosen hyper-parameters based on the set of hyper-parameters defined by the parents. The final set of networks has the same size as the previous generation.
At the end of this process all SCLs have a chance of 10 % to be further mutated by randomly changing hyper-parameters of the classifier for new hyper-parameters from the initial set of parameters, M.
This resulting generation is then trained again and the process described above is repeated. The algorithm is stopped after 10 generations and the best configuration is kept. To determine an optimal set of hyper-parameters for each sub-classifier, the genetic optimisation algorithm is used on each sub-classifier configuration separately. After 10 generations all trained networks are ranked[2].
The choice of the optimal hyper-parameter configuration based on the results of the GOA is determined by ranking all configurations based on the achieved AUC score using the testing set and excluding configurations that show overfitting behaviour. These configurations are excluded by defining tight[3] constraints on the results of the genetic optimisation algorithm. To describe the generalisation capability, $c(\Delta\text{AUC})$ is defined, denoting the difference between the mean AUC score calculated using the testing set, $c(\overline{\text{AUC}_{\text{test}}})$, and the mean AUC score calculated using the training set, $c(\overline{\text{AUC}_{\text{train}}})$, for any configuration c. To constrain the difference between each fold within the cross validation further, the standard deviation of all folds in testing, $\sigma_{\text{AUC}_{\text{test}}}$, is used. The resulting network is asked to have a standard deviation for both testing and training of less than 0.01. Furthermore,

[2]It needs to be noted here, that a more thorough study of the usage of the GOA within the scope of a $t\bar{t}\gamma$ analysis was not performed. The usage of this concept should therefore be understood as an alternative way to draw possible hyper-parameter configurations from a greater subset. Proving its effectiveness compared to a more traditional grid search was beyond the scope of this thesis.
[3]These constraints are arbitrary.

to put more emphasis on generalisation capability than on overall performance [4], the difference between the mean AUC score in training and testing must be smaller than the standard deviation of the testing AUC score. From all possible hyper-parameter configurations $\tilde{c}$ the configuration $c \subset \tilde{c}$ is chosen that fulfils the logical condition defined by the constraints defined below:

$$\max\, c(\mathrm{AUC}_{\mathrm{test}}) \quad \wedge \quad c(\Delta\mathrm{AUC}) < 0.01 \tag{6.3}$$

$$\wedge \quad c(\sigma_{\mathrm{AUC}_{\mathrm{test}}}) < 0.01 \quad \wedge \quad c(\Delta\mathrm{AUC}) < c(\sigma_{\mathrm{AUC}_{\mathrm{test}}}) \tag{6.4}$$

$$\text{with} \quad c(\Delta\mathrm{AUC}) = c(\overline{\mathrm{AUC}_{\mathrm{train}}}) - c(\overline{\mathrm{AUC}_{\mathrm{test}}}) \quad . \tag{6.5}$$

If no resulting sub-classifier fulfils these constraints, the constraints are loosened by gradually increasing the allowed deviations starting from 0.01. The list of SCLs fulfilling these looser constraints is then checked again and a configuration is picked. Imposing the constraints and selecting SCLs accordingly serves as a measure to pick SCLs that perform well while not overfitting the data. The results of the GOA for all SCLs can be found in Table B.1 and Appendix B for the one-vs-all approach and in Appendix B for the one-vs-one approach in the appendix.

6.2. Sub-classifiers

In the following two sub-sections, the resulting SCLs obtained by using the GOA are separately presented for the one-vs-one and one-vs-all approach.

6.2.1. One-vs-All

For the one-vs-all approach, six SCLs were trained. These SCLs represent the probability of an event to be a $t\bar{t}\gamma$, $W\gamma$, $Z\gamma$, $Wt\gamma$, e-fake, $t\bar{t}$ e-fake or $t\bar{t}$ h-fake event. Due to the kinematic similarities among the separating variables, $W\gamma$ and $Z\gamma$ events are treated in a combined sub-classifier. No dedicated SCLs are used for h-fake, $t\bar{t}V$ and diboson events, because studies on these types of SCLs yielded unstable configurations. The obtained results for the six SCLs and a discussion of their individual performances is presented in the following sections.

[4] This decision is based on studies showing that configurations with higher AUC values tend to have greater differences between testing and training performance.

$WZ\gamma$-versus-all

The purpose of the $WZ\gamma$ sub-classifier is to separate $W\gamma$ and $Z\gamma$ events from all other events. Since the variables with the highest separation power are mostly identical for both $W\gamma$ and $Z\gamma$ events when comparing them to other event topologies, these two topologies are grouped into one sub-classifier. The obtained output distributions for the sub-classifier and the corresponding separation power are shown in Figure 6.7. The sub-

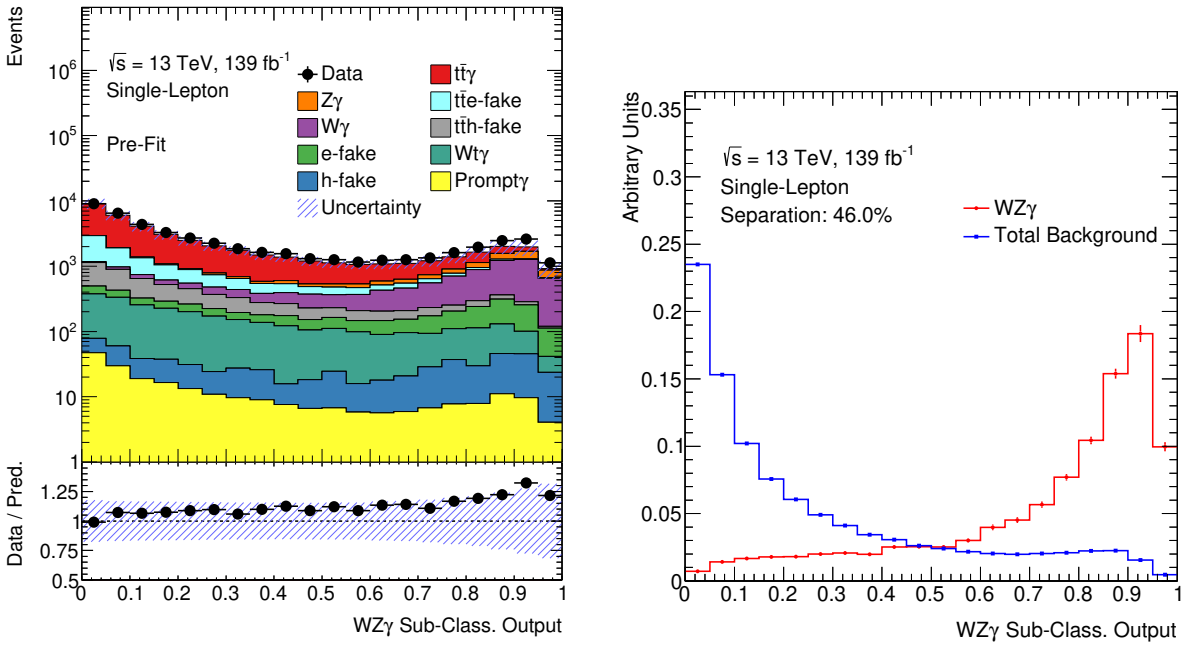

Figure 6.7.: Output distributions of the sub-classifier, responsible for separating $WZ\gamma$ events from all background events (left). Furthermore, the corresponding separation power is shown (right). This sub-classifier provides a separation power of 46.0 % between $WZ\gamma$ events and all background events.

classifier yields a separation power of 46.0 %. The shoulder in the separation plot for high output values of the sub-classifier can be explained by the presence of e-fake events coming from the dileptonic decay of a Z^0 boson. Since the output of the classifier is dominated by b-tagging information, as shown in the list of the 10 dominating Pearson correlation coefficients (PCC) [89] in Figure 6.8, e-fake events have a high probability to be labelled as a $W\gamma$ or $Z\gamma$ event. The ROC curves, as shown in Figure 6.8, are smooth and do not show significant signs of overfitting. The maximum difference between the testing and training ROC curves is 0.9 % with most differences being in the range of 0.5 %.
Therefore, the sub-classifier is kept since it is performing well on unseen data[5]. For future improvements, the contribution of $W\gamma$ and $Z\gamma$ events should be investigated further. Since

[5]The term refers to data that were not previously used in the training or validation.

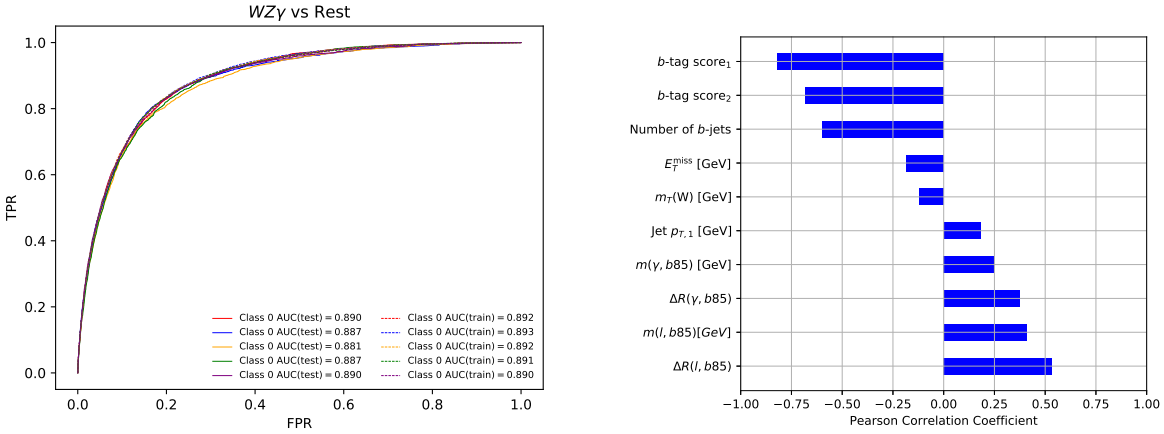

Figure 6.8.: ROC curves for all five folds of the $WZ\gamma$ sub-classifier (left) and the 10 highest PCC (right).

variables with b-tagging information provide high separation power in this case, further investigations on their use can be performed in an attempt to improve the performance of this sub-classifier. Additionally, the overall background modelling of $W\gamma$ and $Z\gamma$ events should be revisited since the shape of the distributions of these events and the expected overall event yields have a significant impact on the training itself [90]. The observed output distribution shows a slope towards higher NN output values. Since no indication for overtraining is observed, the slope can be attested to a potential improper modelling of $W\gamma$ and $Z\gamma$ events.
For the training of this sub-classifier in this analysis, no $W\gamma$ or $Z\gamma$ scale factors are applied. A dedicated study of these two backgrounds was beyond the scope of this analysis but should be considered in the future.

e-fakes-versus-all

Electronic fake photons which are not coming from a dileptonic $t\bar{t}$ decay, make up a small percentage of the predicted overall event yields. Nevertheless, they are easy to separate from all other events using the photon conversion type, therefore utilizing conversion information obtained by the tracker and calorimeters.
Low b-tag scores lead to a good separation from $t\bar{t}\gamma$ events and fake events coming from $t\bar{t}$-decays, since the jets of these events usually have higher b-tag scores. Furthermore, the missing transverse momentum and information about the distance between the photon and neighbouring jets can be used to separate e-fakes from all other events.
The distribution of the output values of the e-fake sub-classifier as well as the separation power of the classifier are shown in Figure 6.9. The sub-classifier yields a separation

power of 34.0 %, hence providing good separation of e-fake events from all other events. The ROC curves for all folds of the e-fake sub-classifier and the distributions of the PCC

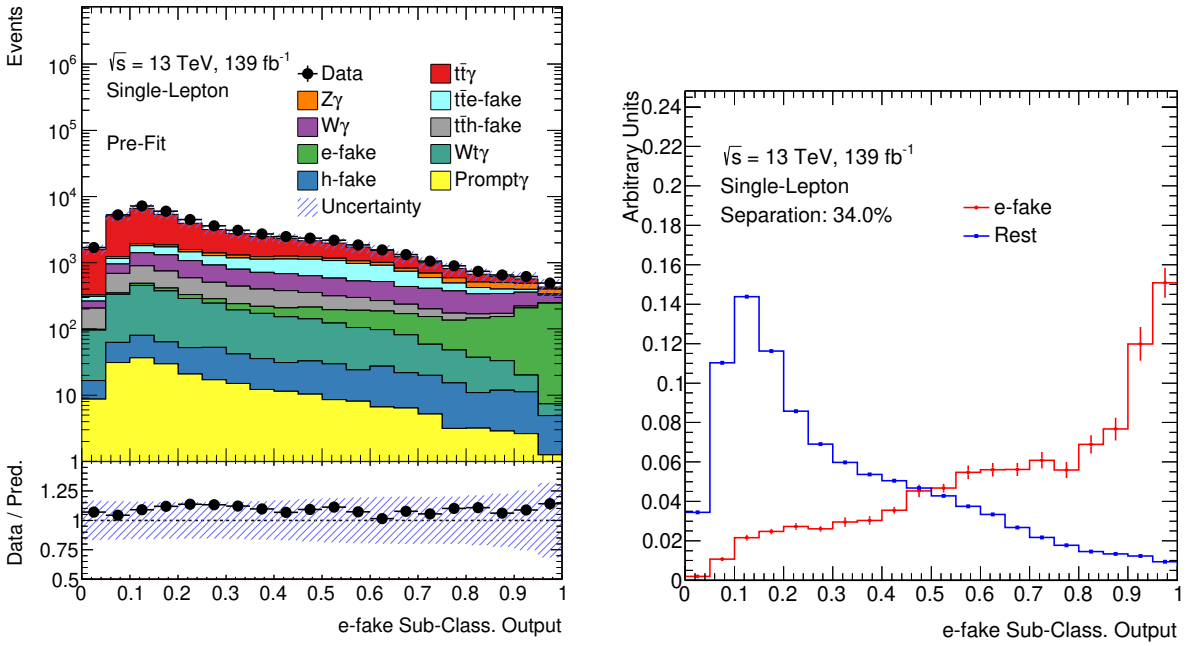

Figure 6.9.: Output distributions of the sub-classifier, responsible for separating e-fake events from all background events (left). Furthermore, the corresponding separation power is shown (right). This sub-classifier provides a separation power of 34.0 % between e-fake events and all background events.

are shown in Figure 6.10. The maximum difference between the ROC curves calculated using the testing and training set is 1.3 % with most differences at around 0.4 %. The differences originate from shape differences between the ROC curves at low FPR values. These differences between the two sets of ROC curves indicate that the classifier performs better on the training set than on the testing set. This behaviour can be explained by the composition of the e-fake event category. Since the category is composed of different types of events as mentioned in Chapters 1 and 3, the e-fake composition within the testing set might differ for each fold, which leads to small performance differences. Nevertheless, when taking the overall distribution shown in Figure 6.9 into account, it can be assumed that the sub-classifier is performing well on unseen data.
It yields results that are expected, reflecting the differences in the kinematics of this type of event from all other types of events. Although this classifier performs rather well on its own, its impact on the one-vs-all multi-class classifier regarding e-fake events is small due to the small number of e-fake events within the signal region.

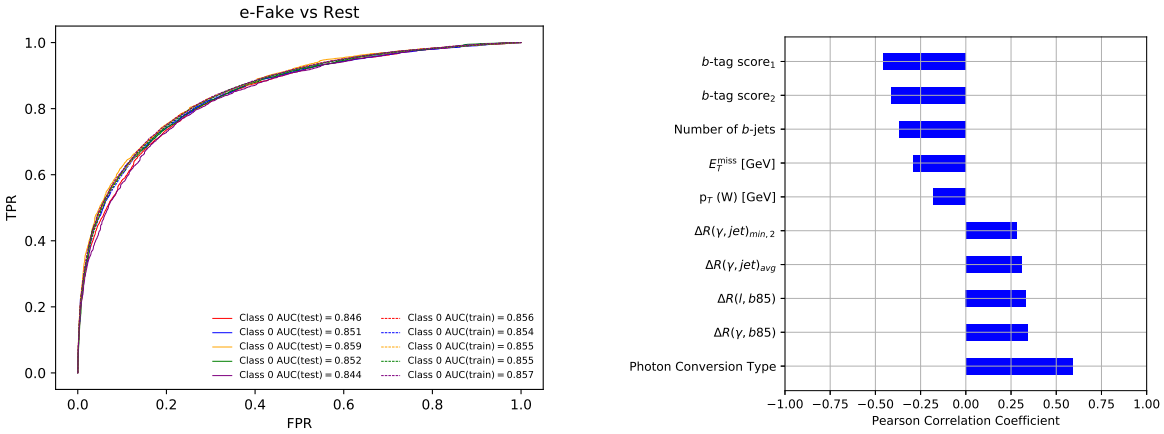

Figure 6.10.: ROC curves for all five folds of the e-fake sub-classifier (left) and the highest ranking PCC (right).

$t\bar{t}$ e-fakes-versus-all

Electronic fakes from $t\bar{t}$ decays are separated from all remaining events using a dedicated $t\bar{t}$ e-fake sub-classifier. The output distribution of this sub-classifier as well as the separation power between $t\bar{t}$ e-fakes and all other events are shown in Figure 6.11. The sub-classifier

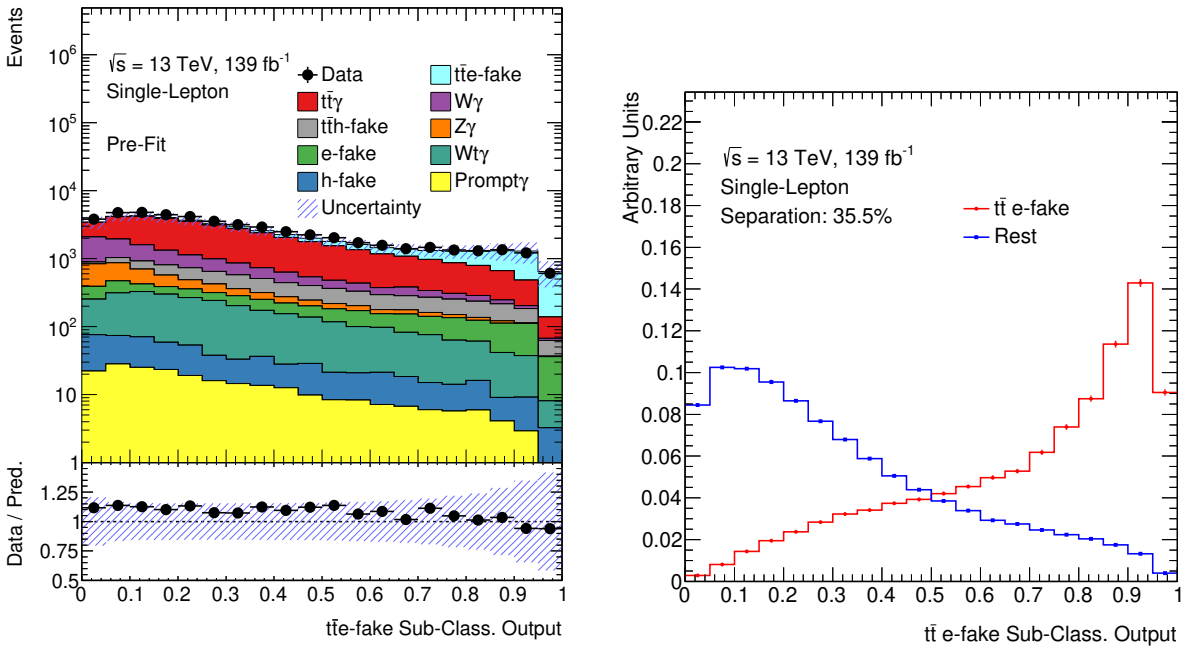

Figure 6.11.: Output distributions of the $t\bar{t}$ e-fake sub-classifier (left) and the corresponding separation plot (right).

provides a separation power of 35.5 % and separates $t\bar{t}$ e-fakes well from all remaining

events. This separation power ranks third when comparing all one-vs-all sub-classifiers, behind the $WZ\gamma$ and e-fake sub-classifiers.
Through differences in the photon conversion type, $t\bar{t}$ e-fake events can be well separated from other events. This is especially true for the separation from $W\gamma$, $Z\gamma$ and $t\bar{t}\gamma$ events. The separation is strongest between $t\bar{t}\gamma$ and $W\gamma$ events as well as $t\bar{t}\gamma$ and $Z\gamma$ events which can be attributed to the additional difference in b-tagging information between these events, since $t\bar{t}$ e-fake final states contain two top quarks, that almost always decay into two b quarks.
The ROC curves and the PCC for the $t\bar{t}$ e-fake sub-classifier are shown in Figure 6.12.

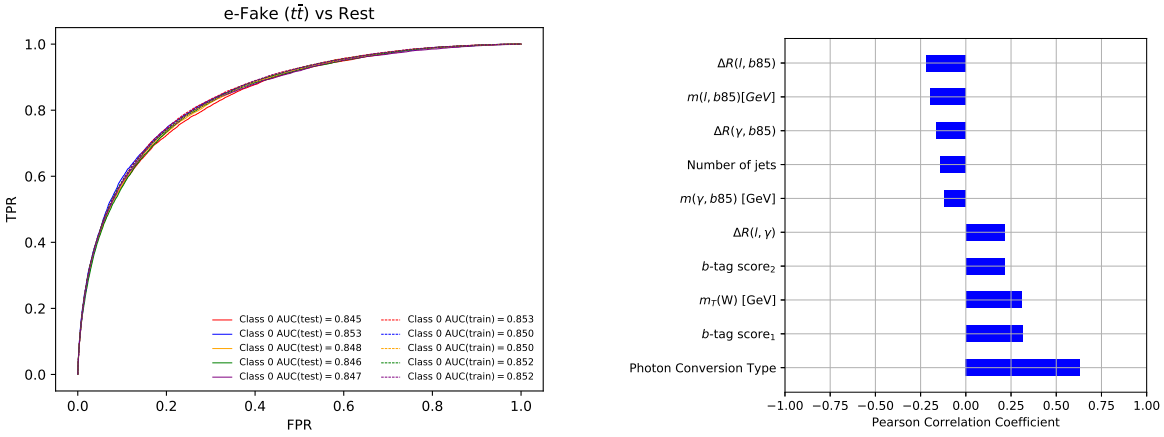

Figure 6.12.: ROC curves for all five folds of the $t\bar{t}$ e-fake sub-classifier (left) and the highest ranking PCC (left).

The calculated ROC curves are smooth and a maximum deviation of 0.8 % between the AUC score in the training and testing process is observed. The smoothness of the ROC curves can also be traced back to the high statistics of $t\bar{t}$ e-fake events from MC as well as the presence of separation variables with good separation power.

$t\bar{t}\gamma$-versus-all

The output distributions for the sub-classifier responsible for separating $t\bar{t}\gamma$ events from all background events and the corresponding separation power are shown in Figure 6.13. The conceptual structure of this sub-classifier is equivalent to the binary classifier approach pursued in the past within the scope of the $t\bar{t}\gamma$ analysis [72] performed by ATLAS using proton-proton collision data taken in 2015 and 2016.

While $W\gamma$, $Z\gamma$ as well as e-fake and $t\bar{t}$ e-fake decays are well separated from $t\bar{t}\gamma$ events, the distribution from $Wt\gamma$ events and $t\bar{t}$ h-fake events peak at higher classifier output values, leading to the observed shoulder in the total background distribution shown in

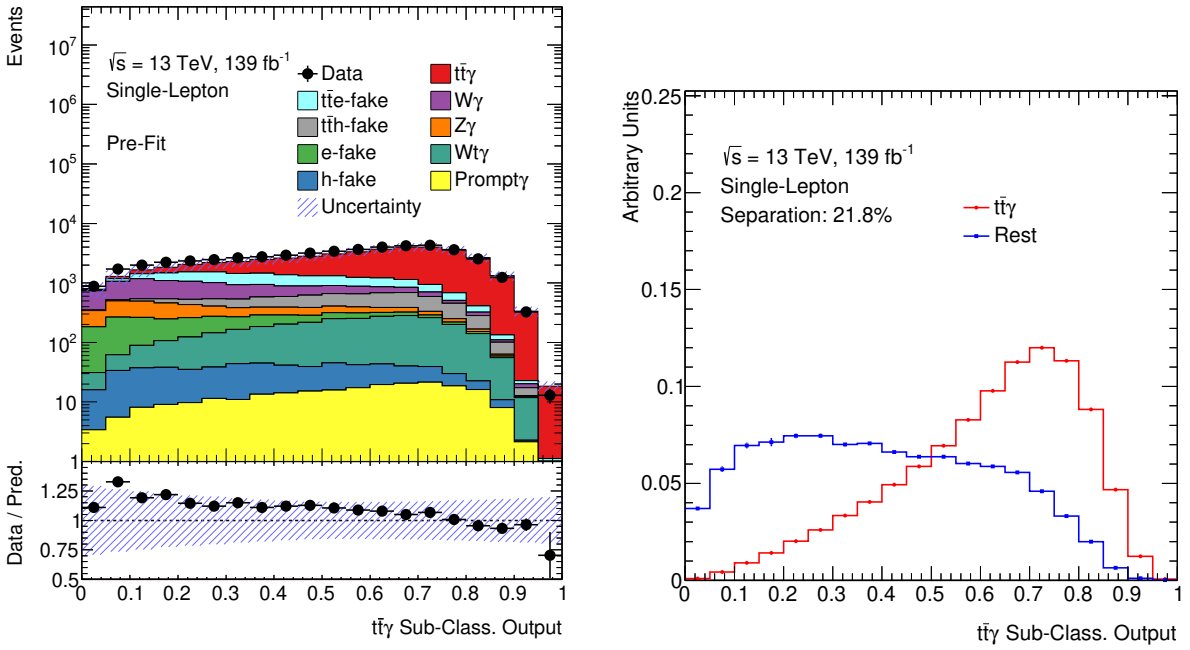

Figure 6.13.: Output distributions of the sub-classifier, responsible for separating $t\bar{t}\gamma$ events from all background events (left). Furthermore, the corresponding separation power is shown (right). This sub-classifier provides a separation power of 21.8 % between $t\bar{t}\gamma$ events and all background events.

the separation plot.

The reason for this overlap is the similarity of the kinematics of $t\bar{t}\gamma$, $Wt\gamma$ and $t\bar{t}$ h-fakes. All of these topologies have a high probability to contain b-jets, tagged with a high score. Since all variables, containing b-tagging information have a high impact on the output of this sub-classifier, their output values are shifted to the right.

The separation of $t\bar{t}\gamma$ and $t\bar{t}$ h-fakes could be improved by using more calibrated calorimeter information in the form of photon shower shapes, thereby focusing on the major difference between those two types of events.

The ROC curves for all folds of the $t\bar{t}\gamma$ sub-classifier and the distributions of the PCC for the input variables with highest correlations are shown in Figure 6.14. The maximum difference between any test and training ROC curve is 0.8 %. Since all testing ROC curves are smooth and yield similar results, the sub-classifier is assumed to perform well on unseen data.

The distribution of the PCC show that variables containing b-tagging information have the highest impact on the output of the sub-classifier. This is expected since a significant impact on the training process comes from $W\gamma$ and $Z\gamma$ background events that can be well separated from $t\bar{t}\gamma$ events using this information.

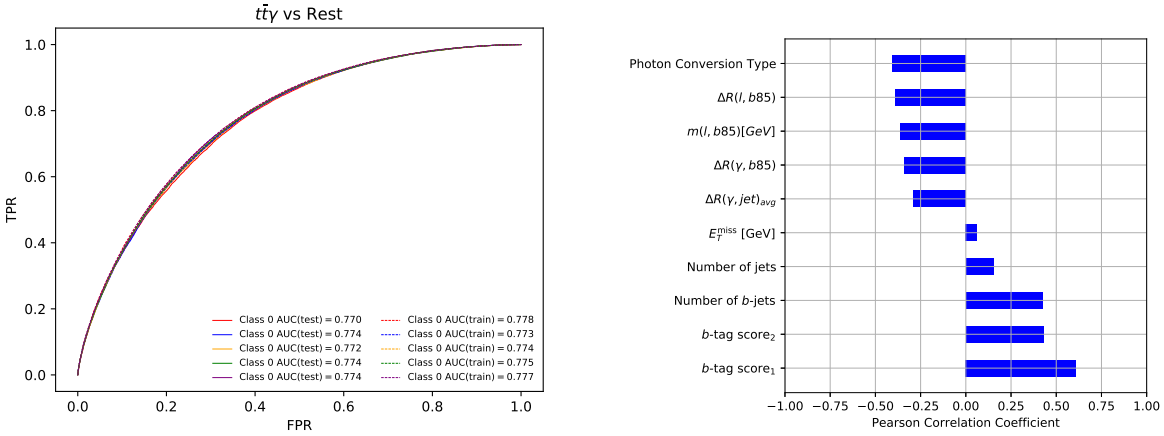

Figure 6.14.: ROC curves for all five folds of the $t\bar{t}\gamma$ sub-classifier (left) and PCC (right).

Furthermore, the photon conversion type and the distances between the photon and the closest b-jet, the distance between the lepton and the closest b-jet and the invariant mass of the lepton and the closest b-jet have high impact. The impact of the photon conversion type ranks high and is the main variable to separate e-fake and $t\bar{t}$ e-fake events from $t\bar{t}\gamma$ events.

$t\bar{t}$ h-fakes-versus-all

Hadronic fake events from $t\bar{t}$-decays are similar to $t\bar{t}\gamma$ events since they also have two decaying top quarks in the final states. The only difference between these events and the $t\bar{t}\gamma$ signal events is whether the photon is a faked photon or not. Since the training of a sub-classifier separating $t\bar{t}$ h-fakes from all remaining events largely depends on the difference of these events and $t\bar{t}\gamma$ signal events, the separation is challenging.
The resulting trained sub-classifier and its separation power for separating $t\bar{t}$ h-fakes from all other events is shown in Figure 6.15. The sub-classifier shows a separation power of 11.8 %. The performance of this sub-classifier suffers from the fact that no shower-shape variables and thus no shower information from the electromagnetic and hadronic calorimeters is used. The separation then has to rely on kinematic information alone. Nevertheless, the sub-classifier is able to provide a small separation between $t\bar{t}$ h-fakes and all other events. An upward slope is observed in Figure 6.15 for higher classifier output values, covering bins with decreasing contents and fewer than 2000 entries. Similar to the slope observed for the $WZ\gamma$ this effect does not seem to be a training effect since the ROC curves shown in Figure 6.16 do not show any signs for overfitting or underfitting. Therefore, this effect can be attested to the improper modelling of $t\bar{t}$ h-fakes. To further understand the impact of the $t\bar{t}$ h-fake events on the sub-classifier, the need of data-

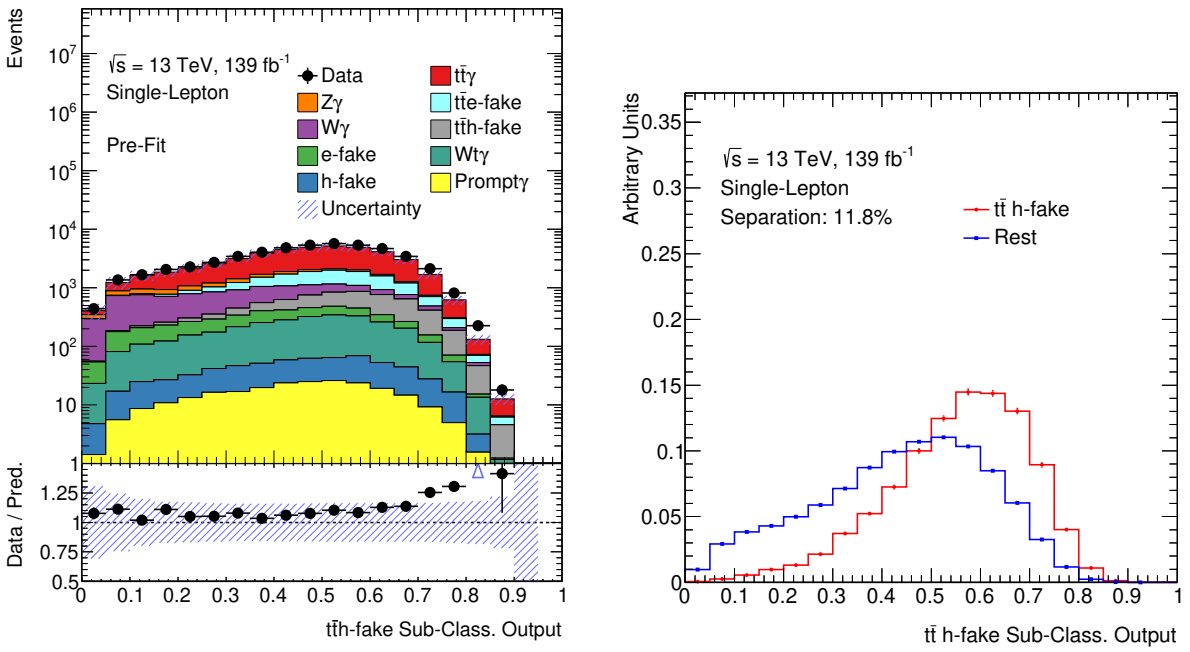

Figure 6.15.: Output distributions of the sub-classifier, responsible for separating $t\bar{t}$ h-fake events from all background events (left). Furthermore, the corresponding separation power is shown (right). This sub-classifier provides a separation power of 11.8 % between $t\bar{t}$ h-fake events and all background events.

driven fake estimates has to be investigated and their values determined. Data-driven fake estimates based on bins of photon p_{T} and photon $|\eta|$ can have a significant effect on the output of the classifier since one of the variables within the list of highest PCC is the transverse momentum of the photon. The list of the highest ranking PCC is shown in the right plot of Figure 6.16.
Since the impact of this sub-classifier in the corresponding multi-class approach is small compared to the other SCLs and it seems to perform well for values of less than 0.7, thereby including a majority of all events, the classifier is retained.

$Wt\gamma$-versus-all

The obtained output distributions for the sub-classifier responsible for separating $Wt\gamma$ events from all other events and the corresponding separation power are shown in Figure 6.17. The ROC curves for all folds of the $Wt\gamma$ sub-classifier and the distributions of the PCC are shown in Figure 6.18. The test AUC scores differ from the train AUC scores by less than 1.4 %. Although the difference between the performance on seen and unseen data is relatively small, it still indicates a small overfitting component of the sub-classifier.

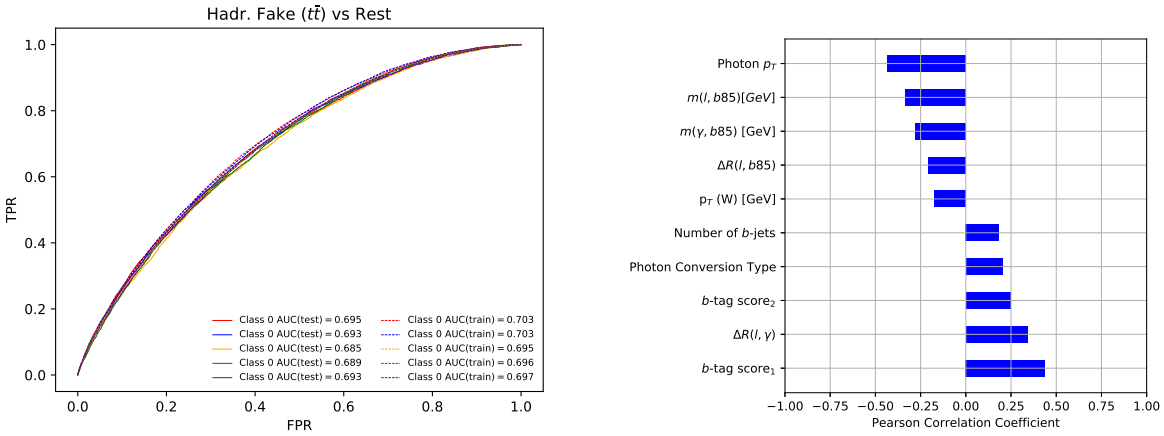

Figure 6.16.: ROC curves for all five folds of the $t\bar{t}$ h-fake sub-classifier (left) and PCC for the top 10 variables showing the highest correlation (right).

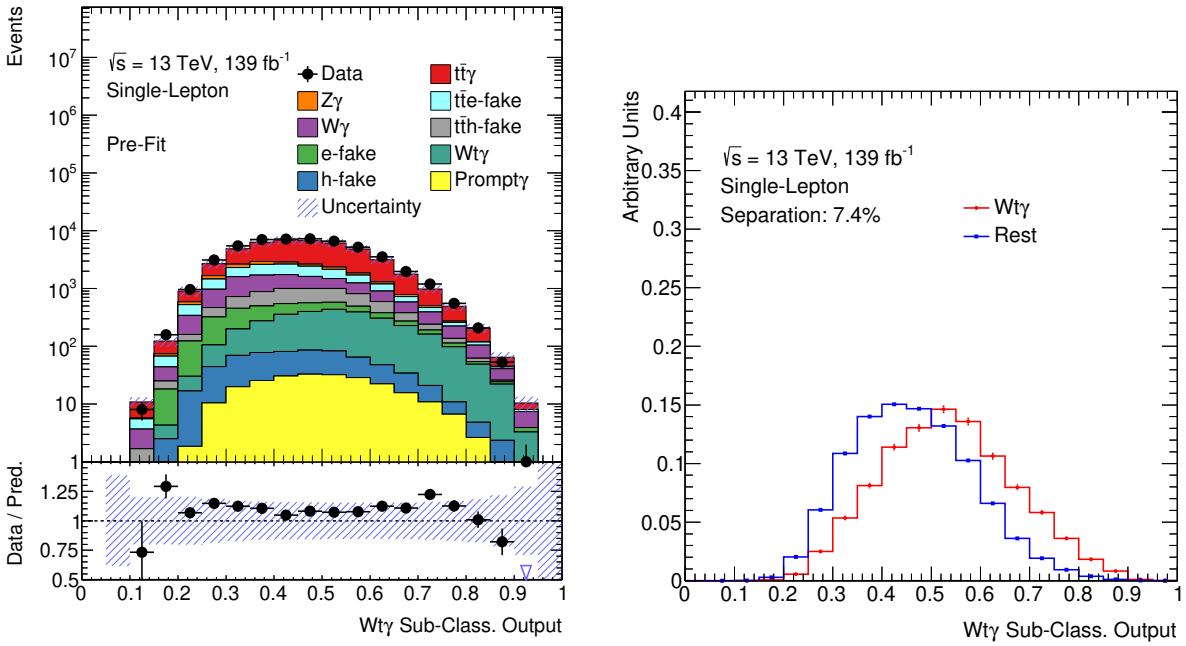

Figure 6.17.: Output distributions of the sub-classifier, responsible for separating $t\bar{t}\gamma$ events from all background events (left). Furthermore, the corresponding separation power is shown (right). This sub-classifier provides a separation power of 7.4 % between $t\bar{t}\gamma$ events and all background events.

This small amount of overfitting could not be reduced further. Nevertheless, other folds show smaller discrepancies between the testing and training performance. Therefore, this sub-classifier is retained. Detailed studies on the generalisation performance of this sub-classifier should be conducted in the future.
The overall performance of the sub-classifier is low compared to the other SCLs due to the kinematic similarities between $Wt\gamma$ events and $t\bar{t}\gamma$ signal events. All PCC are below 0.5 with the transverse momentum of the $W^{\pm}$ boson and the transverse momentum of the leading jet ranking highest.

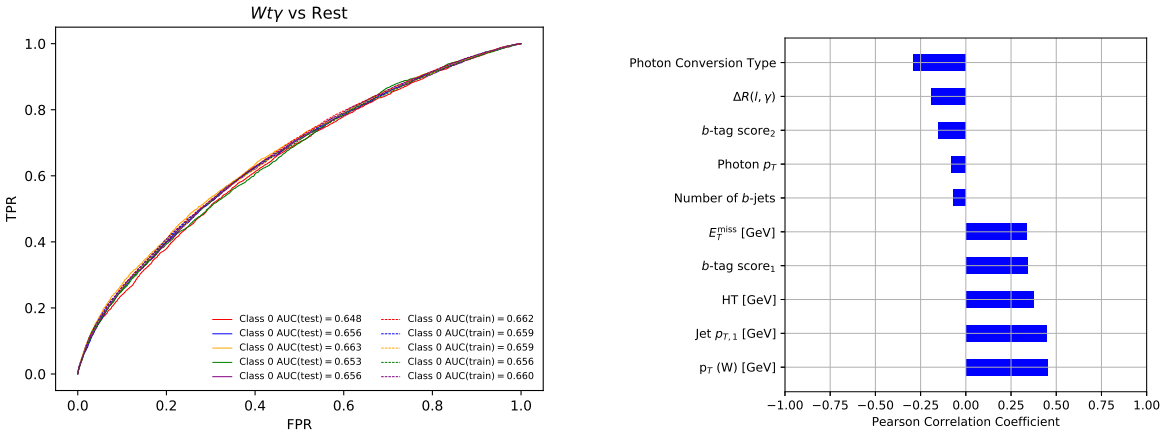

Figure 6.18.: ROC curves for all five folds of the $Wt\gamma$ sub-classifier (left) and the PCC (right).

These two variables have separation powers of 1.74 % and 2.79 % for separating $Wt\gamma$ from $t\bar{t}\gamma$. The difference in performance between this sub-classifier and the other SCLs can be understood by comparing the separation power of these two variables, to the highest ranking variables in other sub-classifiers. Since the separation power of the variables is lower for the $Wt\gamma$ sub-classifier, its performance is also lower and the two classes are harder to separate. The separation plots of the two variables are shown in Figure 6.19. The separation power of all other input variables is smaller. This low separation power leads to the the small separation power of the sub-classifier since the training is dominated by $Wt\gamma$ and $t\bar{t}\gamma$ events.

6.2.2. One-vs-One

The one-vs-one approach focuses on using SCLs as input that separate a specific type of background b from the $t\bar{t}\gamma$ signal class s. These SCLs describe the probability of an event i being a $t\bar{t}\gamma$ event assuming $i \in (s \cup b)$, $P(i \in (s \cup b)|s)$. The SCLs are trained for

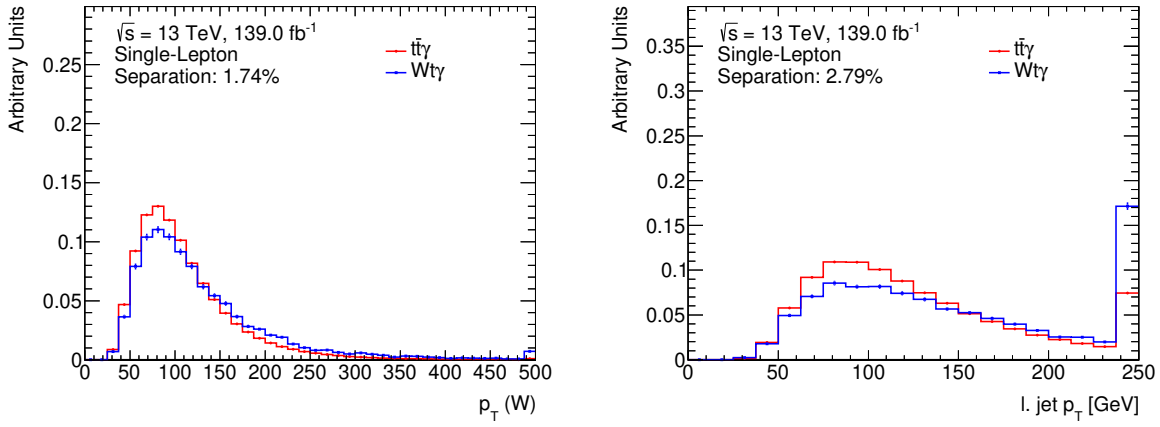

Figure 6.19.: Separation plots for the transverse momentum of the $W^{\pm}$ boson (left), $p_{\mathrm{T}}(W)$, and the transverse momentum of the leading jet (right), jet $p_{\mathrm{T},1}$.

$b =$ e-fake, $b = t\bar{t}$ e-fake, $b = t\bar{t}$ h-fake, $b = Wt\gamma$, and $b = WZ\gamma$. The impact of h-fake, $t\bar{t}V$ and diboson events is not treated in a separate classifier for the same reasons as presented in Section 6.2.1. In the following sections the results of the SCLs used in the one-vs-one approach are presented.

$t\bar{t}\gamma$-versus-$Wt\gamma$

The dedicated $t\bar{t}\gamma$-versus-$Wt\gamma$ sub-classifier aims at separating $t\bar{t}\gamma$ events from $Wt\gamma$ events for the one-vs-one multi-class classifier. All non $t\bar{t}\gamma$ and non $Wt\gamma$ events are neglected during the training and validation of the sub-classifier. The sub-classifier is therefore similar to the sub-classifier presented in Section 6.2.1 and is facing identical challenges as the $Wt\gamma$-versus-all sub-classifier.
Nevertheless, the separation of events is different for this classifier since it focuses on $t\bar{t}\gamma$ and $Wt\gamma$ events only, thereby putting more emphasis on the kinematic differences between those two similar types of events. The output distribution of the sub-classifier and the separation power achieved are shown in Figure 6.20.

The sub-classifier yields a separation power of 9.8 % which is 2.0 % higher than the separation power achieved by the $Wt\gamma$-versus-all sub-classifier used in the one-vs-all approach. The resulting ROC curves of the best performing model with the least degree of overfitting observed in all trained models is shown in Figure 6.21. The ROC curves show a smooth behaviour. The maximum difference measured between the testing and training AUC scores is 1.2 %. The variables with the highest impact on the sub-classifier are the invariant mass of the lepton and the closest b-jet tagged with the MV2C10-85% working point, $m(l, b)$. Furthermore, variables including information about jets such as their transverse momentum, jet $p_{\mathrm{T},i}$, as well as the sum of all transverse momenta, H_T,

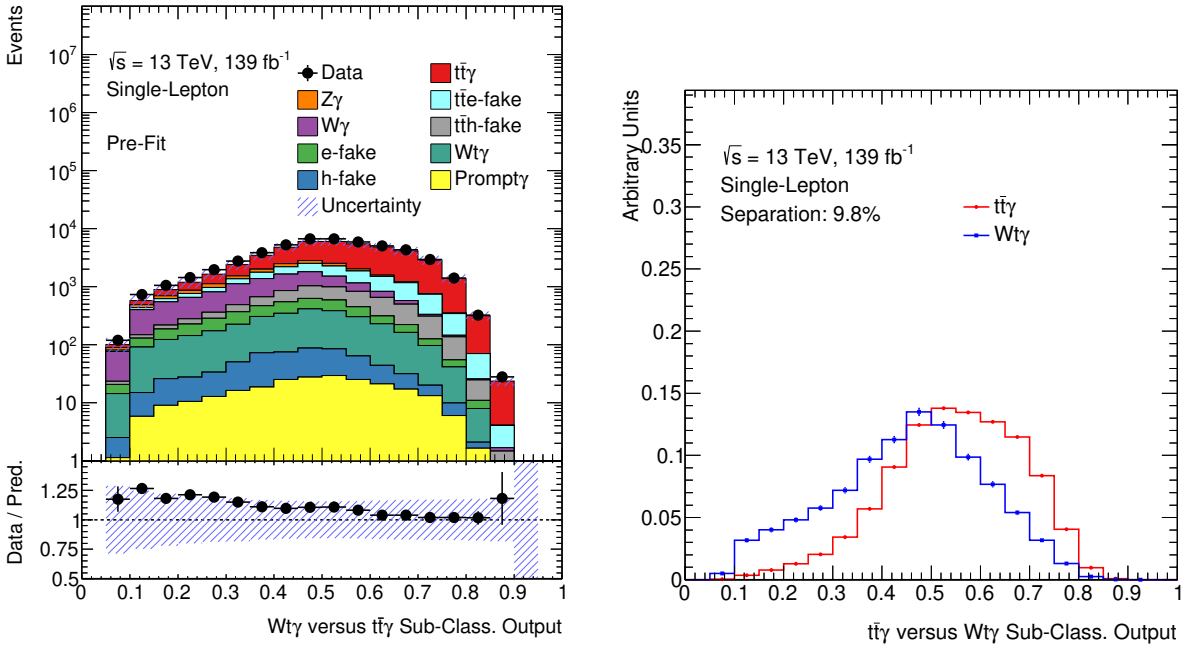

Figure 6.20.: Output distributions of the $Wt\gamma$-versus-$t\bar{t}\gamma$ sub-classifier (left) and the corresponding separation plot (right).

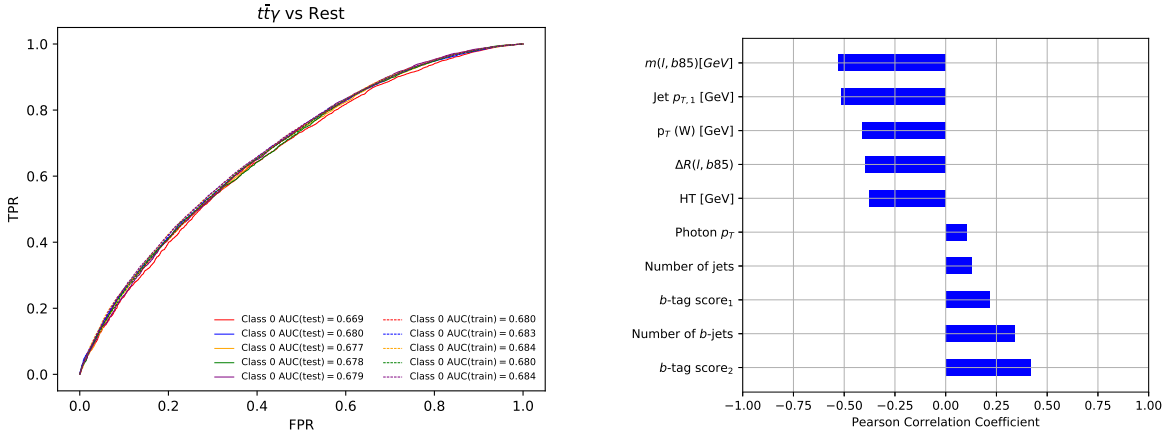

Figure 6.21.: ROC curves for all five folds of the $Wt\gamma$-versus-$t\bar{t}\gamma$ sub-classifier (left) and list of PCC (right).

rank high.

$t\bar{t}\gamma$-versus-$WZ\gamma$

The $t\bar{t}\gamma$-versus-$WZ\gamma$ sub-classifier is trained by only taking $W\gamma$, $Z\gamma$ and $t\bar{t}\gamma$ events into account. The impact of all other background events is neglected. Background events of the $W\gamma$ and $Z\gamma$ category can be easily separated from $t\bar{t}\gamma$ events by utilizing the differences in b-tagging information between these background events and the $t\bar{t}\gamma$ signal events.
The distribution of the output of the sub-classifier is shown in Figure 6.22. A clear separation of 49.8 % between $W\gamma$ and $Z\gamma$ and $t\bar{t}\gamma$ signal events is observed as shown in the separation plot on the left of Figure 6.22. While both $t\bar{t}$ h-fake and $t\bar{t}$ e-fake as well as $Wt\gamma$ events peak below the peak of the $t\bar{t}\gamma$ signal on the far right, both e-fake and h-fake events peak on the far left below the peaks of $W\gamma$ and $Z\gamma$. Both e-fake and h-fake events originate from decays of Z^0 or $W^{\pm}$ bosons, where in the case of the Z^0 boson one lepton is misidentified as a photon and in the case of the $W^{\pm}$ boson one jet fakes a photon. Their topologies are very similar to the topologies where the photon is a prompt one in these decays, which leads to the observed peak at low NN output values.
Equivalently, $t\bar{t}$ e-fakes and $t\bar{t}$ h-fakes peak below the $t\bar{t}\gamma$ signal peak since their topology is, apart from the nature of the photon, almost identical to the topology of the $t\bar{t}\gamma$ signal events.
The ROC curves for all folds of the sub-classifier separating $W\gamma$ and $Z\gamma$ events from $t\bar{t}\gamma$ events and the distributions of the PCC are shown in Figure 6.23. The difference between the training and testing ROC curves is below one percent for all ROC curves accept for the third one, where the difference is 1.5 %. Nevertheless, the sub-classifier can be seen as non-overfitting since all testing ROC curves agree well with another with differences below one percent.
The distribution of the PCC also confirms that output of the $t\bar{t}\gamma$-versus-$WZ\gamma$ sub-classifier depends heavily on variables involving b-tagging information as expected for this kind of topology.

$t\bar{t}\gamma$-versus-$t\bar{t}$ h-fakes

To separate $t\bar{t}\gamma$ events from $t\bar{t}$ h-fake events, the dedicated $t\bar{t}\gamma$-versus-$t\bar{t}$ h-fakes sub-classifier is used as an input to the one-vs-one multi-class classifier. The development of this sub-classifier is based on the fact that $t\bar{t}$ h-fakes make up the dominant part of all hadronic fake photons. Furthermore, the kinematic distributions of these events are similar to the $t\bar{t}\gamma$ signal events, thereby making their separation challenging and imposing the

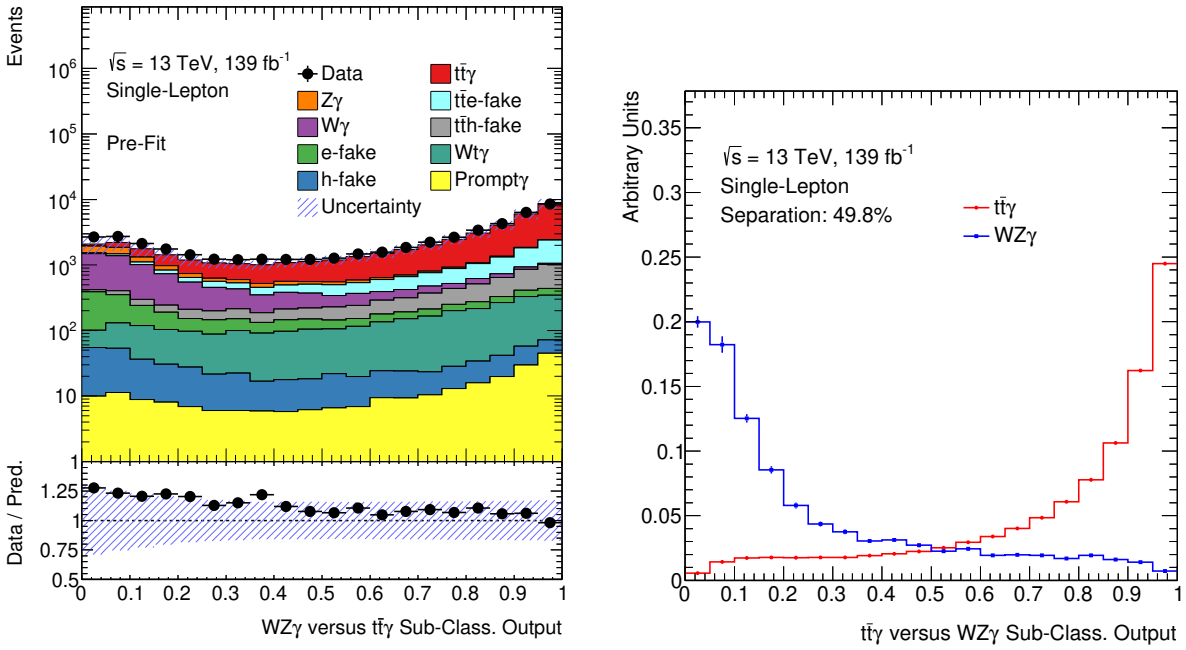

Figure 6.22.: Output distributions of the sub-classifier, responsible for separating $W\gamma$ and $Z\gamma$ events from $t\bar{t}\gamma$ events (left). Furthermore, the corresponding separation power is shown (right). This sub-classifier provides a separation power of 49.8 % between $W\gamma$ and $Z\gamma$ events and $t\bar{t}\gamma$ events.

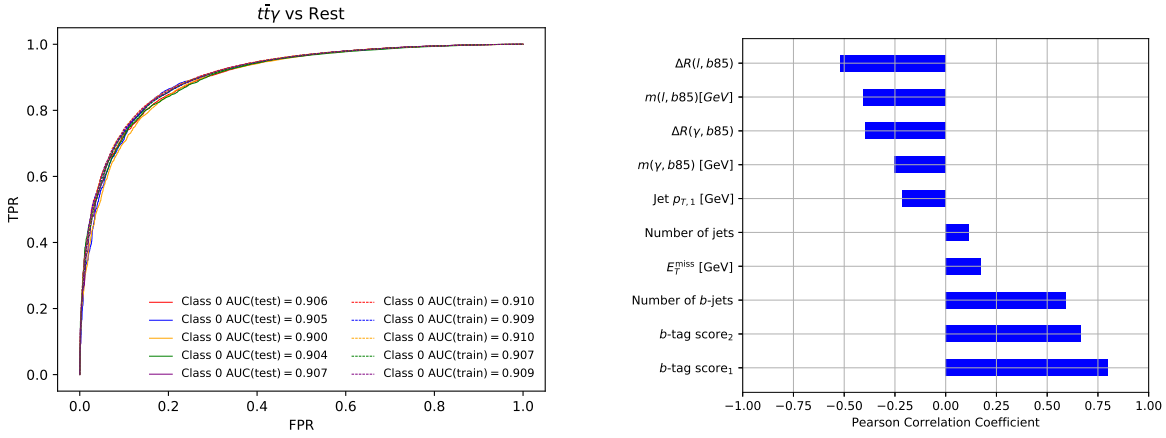

Figure 6.23.: ROC curves for all five folds of the sub-classifier separating $t\bar{t}\gamma$ events from $W\gamma$ and $Z\gamma$ events (left) and the highest ranking PCC (right)

need of a dedicated sub-classifier. The output distributions of the trained sub-classifier are shown in Figure 6.24. Additionally, the separation plot depicting the separation of $t\bar{t}\gamma$ events from $t\bar{t}$ h-fakes events is also shown in Figure 6.24. Similar to the $t\bar{t}$ h-fake

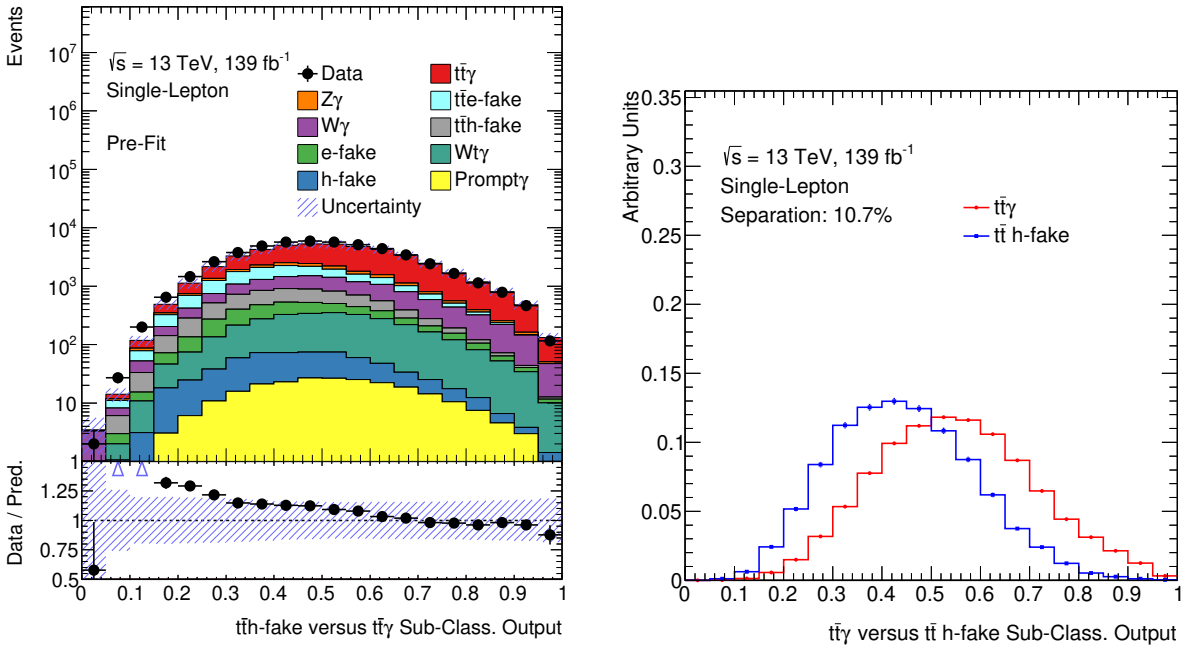

Figure 6.24.: Output distributions of the sub-classifier, responsible for separating $t\bar{t}$ h-fake from $t\bar{t}\gamma$ events (left). Furthermore, the corresponding separation power is shown (right). This sub-classifier provides a separation power of 10.7 % between $t\bar{t}$ h-fake events and $t\bar{t}\gamma$ events.

sub-classifier presented in Section 6.2.1, the sub-classifier shows a disagreement between data and MC, with a possible explanation provided in the previous section. Since the signal and background definition is reversed compared to the $t\bar{t}$ h-fake sub-classifier, the discrepancy is observed for low output values of the $t\bar{t}\gamma$-versus-$t\bar{t}$ h-fakes sub-classifier. Similar to the $t\bar{t}$ h-fakes sub-classifier, the $t\bar{t}\gamma$-versus-$t\bar{t}$ h-fakes sub-classifier is used nevertheless in a multi-class approach.

The sub-classifier provides a separation power of 15.7 % for the separation of $t\bar{t}$ h-fake and $t\bar{t}\gamma$ events. The ROC curves for all folds of the $t\bar{t}$ h-fake sub-classifier and the distributions of the PCC are shown in Figure 6.25.

The ROC curves show a maximum difference of 0.8 % suggesting that an improper modelling of the $t\bar{t}$ h-fakes is the cause of the slope observed in the distribution in Figure 6.24 and not the training process itself. Although the photon conversion type helps with separating $t\bar{t}$ h-fakes from $t\bar{t}\gamma$ events, the magnitudes of the PCC in Figure 6.25 show that most of the chosen input variables have a small impact on the output of the sub-classifier.

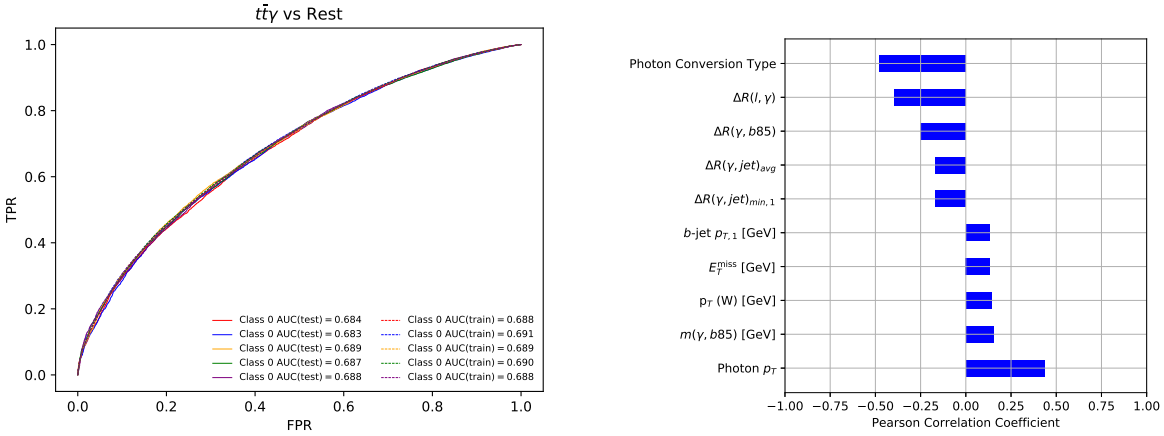

Figure 6.25.: ROC curves for all five folds of the sub-classifier separating $t\bar{t}\gamma$ events from $t\bar{t}$ h-fake events (left) and the 10 highest ranking PCC.

To improve this specific type of sub-classifier as well as the similar $t\bar{t}$ h-fake sub-classifier, new discriminating variables need to be introduced that focus on the nature of the photon to distinguish these two processes.

$t\bar{t}\gamma$-versus-$t\bar{t}$ e-fakes

The $t\bar{t}\gamma$-versus-$t\bar{t}$ e-fakes sub-classifier focuses on separating $t\bar{t}\gamma$ events from $t\bar{t}$ e-fake events. The separation of these two types of events is predominantly based on the photon conversion type of the photon since this variable provides the highest separation power for these two types of events. The second highest ranking variable is the transverse mass of the $W^{\pm}$ boson. The separation plots for these two variables are shown in Figure C.10 in the appendix. The separation power is 21.41 % and 4.64 %, respectively. These events originate from dileptonic $t\bar{t}$ decays and represent the dominant contribution to all electronic fake events. The output distribution of the $t\bar{t}\gamma$-versus-$t\bar{t}$ e-fake sub-classifier as well as the corresponding separation plot are shown in Figure 6.26. The separation power of the sub-classifier is 37.4 %. From Figure 6.26 it can also be seen that the sub-classifier performs well for separating all other backgrounds from $t\bar{t}$ e-fake events except for e-fake events. The peak of e-fake events at low output values can be attributed to the similarity of the photon conversion type for these two types of events. However, this behaviour is expected since e-fake events where neglected during the training of this sub-classifier.
Furthermore, Figure 6.27 shows the ROC curves for all folds of the e-fake sub-classifier for events coming from $t\bar{t}$ decays and the distributions of the PCC again for the variables showing the highest correlation to the output of the sub-classifier. The list of PCC also confirms the importance of the variables mentioned above. It can also be seen that vari-

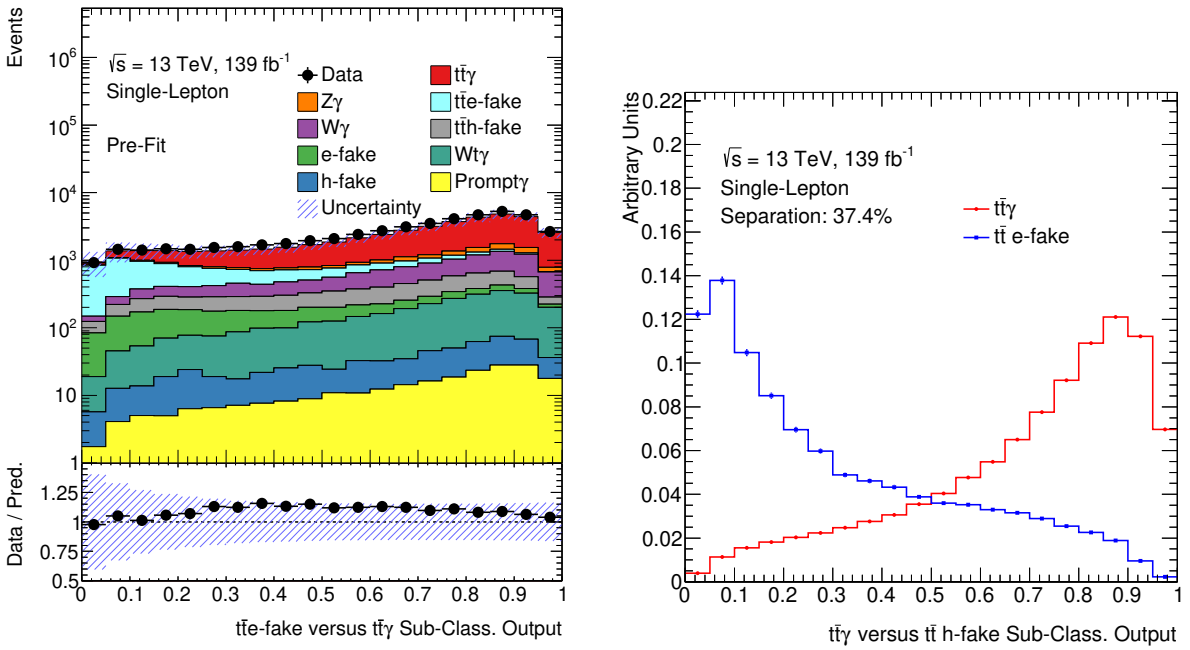

Figure 6.26.: Output distributions of the sub-classifier, responsible for separating $t\bar{t}$ e-fake from $t\bar{t}\gamma$ events (left). Furthermore, the corresponding separation power is shown (right). This sub-classifier provides a separation power of 37.4 % between $t\bar{t}$ e-fakes and $t\bar{t}\gamma$ events.

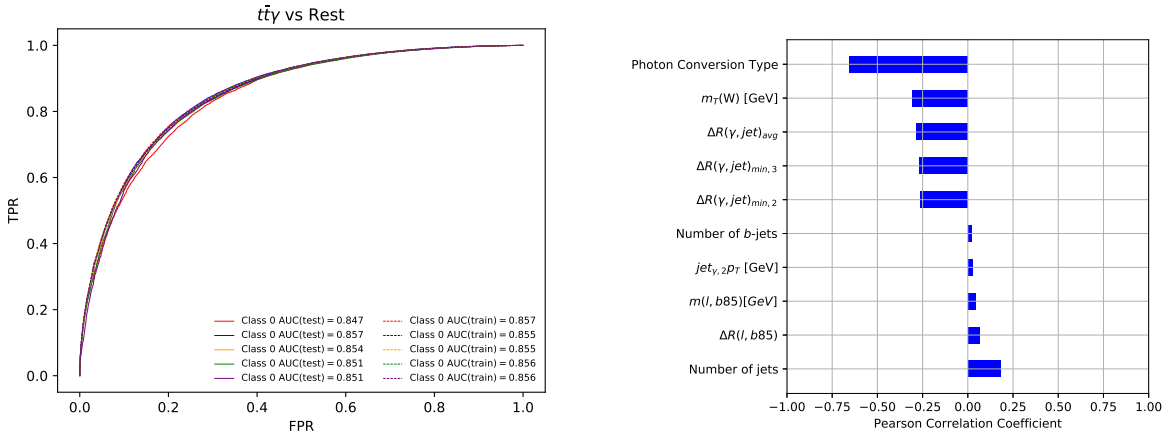

Figure 6.27.: ROC curves for all five folds of the sub-classifier separating $t\bar{t}\gamma$ events from $t\bar{t}$ e-fake events (left). In addition, the 10 highest ranking PCC are shown (right).

ables incorporating distances between the photon and the neighbouring jets as well as the invariant mass between the photon and the lepton $m(l, \gamma)$ have a significant impact on the output of the sub-classifier.
The ROC curves are smooth and the maximum difference between the AUC score for the testing and training data is 1.0 % with most other differences being in the order of half a percent. The sub-classifier is therefore assumed to perform well on unseen data and kept as an input of the one-vs-one multi-class classifier.

$t\bar{t}\gamma$-versus-e-fakes

The $t\bar{t}\gamma$-versus-e-fakes sub-classifier is similar to the e-fakes sub-classifier presented above, however only $t\bar{t}\gamma$ and e-fake events are taken into account during training for the $t\bar{t}\gamma$-versus-e-fake sub-classifier. Its purpose is to separate $t\bar{t}\gamma$ events from e-fake background events. To separate e-fake events from $t\bar{t}\gamma$ events, the photon conversion type as well is used again. Its high separation power provides the basis of separating these two types of events. Furthermore, similar to the e-fake sub-classifier, b-tagging variables such as the number of b-tagged jets as well as the first and second leading b-tagging scores provide high separation power. In addition, the presence of additional neutrinos in the event due to the main source for e-fakes being the dileptonic Z^0 decay, lead to a good separation via the missing transverse momentum, $E_{\mathrm{T}}^{\mathrm{miss}}$.
The output distribution as well as the separation plot for the $t\bar{t}\gamma$-versus-e-fake sub-classifier is shown in Figure 6.28. e-fakes are well separated from $t\bar{t}\gamma$, $Wt\gamma$ and $t\bar{t}$ h-fake events, because both b-tagging scores and photon conversion type information is significantly different for e-fakes and the listed types of events. These differences within the discriminating variables are the origin of the high separation power. Events such as $t\bar{t}$ e-fake and $W\gamma$ are less well separated because the differences in photon conversion type ($t\bar{t}$ e-fake) or b-tagging scores ($W\gamma$) are smaller. The shallow slope of non-e-fake events at low output values is dominated by $Z\gamma$ events that are incorrectly labelled as e-fake events. This miss-labelling is due to the similarities in the discriminating variables listed above.

The ROC curves of all folds and the PCC of the 10 variables that show the highest correlation are shown in Figure 6.29. While most of the difference between the two ROC curves are in the range of 0.3 %, two of the ROC curves show differences between the two sets of 1.0 and 1.4 %. These observed differences originate from difference of the training and testing ROC curves at small FPR values. The differences were previously presented in the discussion of the e-fake-versus-all sub-classifier above. In total the sub-classifier performs well and is used as an input for the one-vs-one approach. Similar to

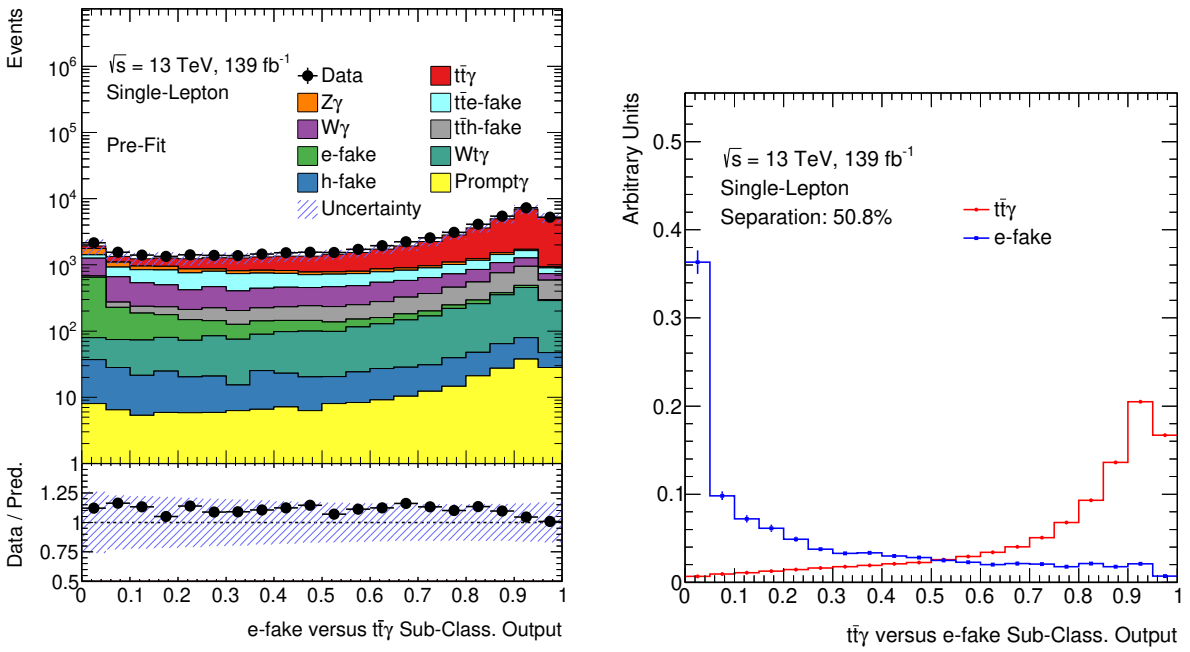

Figure 6.28.: Output distributions of the sub-classifier, responsible for separating $t\bar{t}$ e-fake from $t\bar{t}\gamma$ events (left). Furthermore, the corresponding separation power is shown (right). This sub-classifier provides a separation power of 50.8 % between e-fakes and $t\bar{t}\gamma$ events.

the e-fake-versus-all sub-classifier, its impact on the multi-class output is small because the contribution of e-fake events in the $t\bar{t}\gamma$ signal region is small.

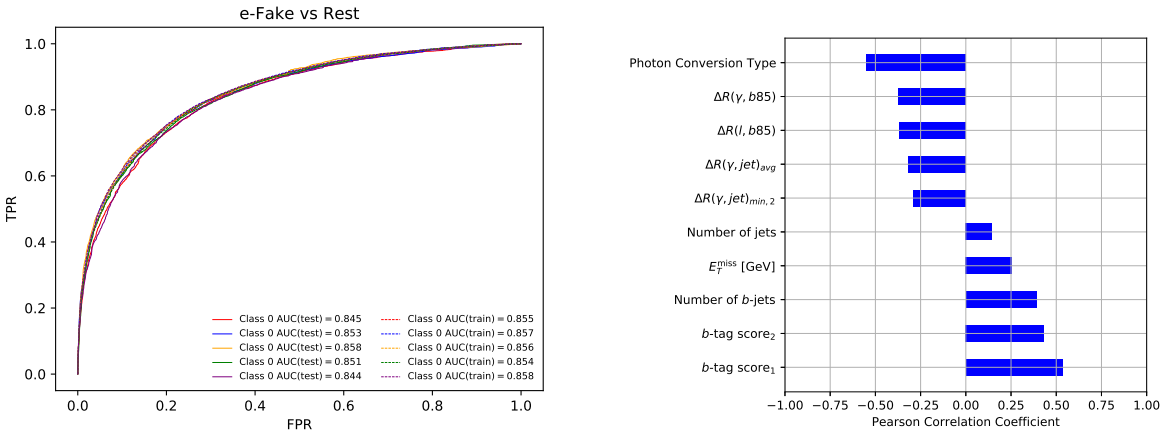

Figure 6.29.: ROC curves for all five folds of the sub-classifier separating $t\bar{t}\gamma$ events from e-fake events (left) and the list of the 10 highest ranking PCC (right)

6.2.3. Comparing different approaches

In this section, the results of the one-vs-one and the one-vs-all approach and their potential impact on the $t\bar{t}\gamma$ cross section measurement as well as their potential for use in a future ATLAS measurement of the $t\bar{t}\gamma$ process. Different metrics are introduced to compare the various aspects of the two approaches.

To directly compare the performance of the two approaches their AUC score and the corresponding ROC curves can be compared. Since the AUC score provides a aggregate measure of the performance of a classifier across all possible classification thresholds, it is well suited for comparing the two approaches. For both approaches ROC curves are calculated by the set of combinations C, defined as C =[$t\bar{t}\gamma$-versus-rest, *Fake*-versus-rest, *Other Prompt*-versus-rest], yielding five ROC curves respectively. The ROC curves for the one-vs-one and one-vs-all approach are shown in Figure 6.30.

By comparing each class with the remaining two other classes, the problem is effec-

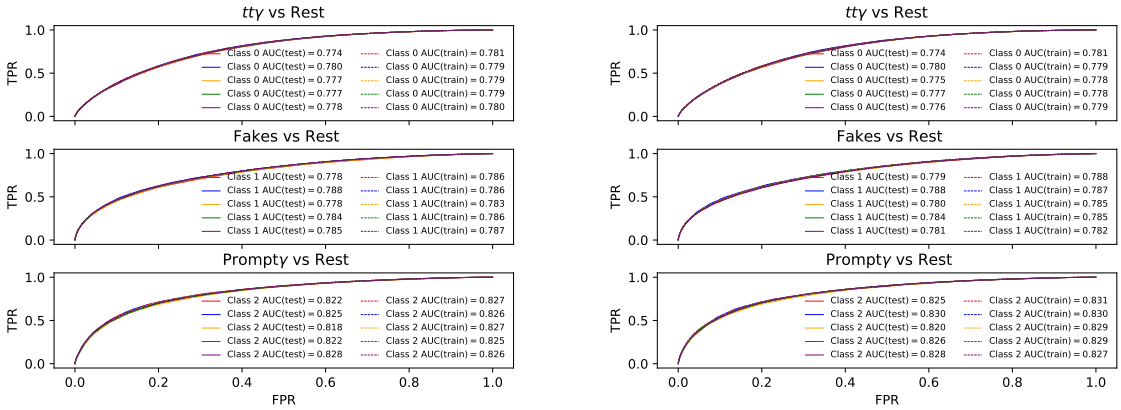

Figure 6.30.: ROC curves for all five folds of the multi-class classifier using the one-vs-one approach (left) and the multi-class classifier using the one-vs-all approach (right).

tively reduced to a binary classification problem, enabling a less difficult and piecewise binary comparison of both multi-class approaches. To negate the effect of statistical fluctuations, the mean AUC scores are compared. The mean AUC score for the *Other Prompt*-versus-rest ROC curve using the one-vs-all approach, $\overline{\text{AUC}}^{\text{ova}}_{\textit{Other Prompt}}$, ranks highest with a mean AUC score of $\overline{\text{AUC}}^{\text{ova}}_{\textit{Other Prompt}} = 0.8258$ followed by the corresponding AUC score using the one-vs-one approach, $\overline{\text{AUC}}^{\text{ovo}}_{\textit{Other Prompt}}$, with a mean AUC score of $\overline{\text{AUC}}^{\text{ovo}}_{\textit{Other Prompt}} = 0.8230$. Therefore, the one-vs-all approach performs better than the one-vs-one approach for separating $W\gamma$, $Z\gamma$ and Promptγ events from the rest.

The second highest pair of AUC scores are the AUC scores for the ROC curves describ-

ing the *Fake* versus rest separation performance, $\overline{\text{AUC}}^{\text{ovo}}_{Fake}$ and $\overline{\text{AUC}}^{\text{ova}}_{Fake}$. The one-vs-one approach achieves $\overline{\text{AUC}}^{\text{ovo}}_{Fake} = 0.7826$ and subsequently performs slightly better than the one-vs-all approach with $\overline{\text{AUC}}^{\text{ova}}_{Fake} = 0.7824$.
The remaining pair of AUC scores describes the $t\bar{t}\gamma$ versus rest separation performances for the two approaches, $\overline{\text{AUC}}^{\text{ovo}}_{t\bar{t}\gamma}$ and $\overline{\text{AUC}}^{\text{ova}}_{t\bar{t}\gamma}$. In this scenario, the one-vs-one approach yields a mean AUC score of $\overline{\text{AUC}}^{\text{ovo}}_{t\bar{t}\gamma} = 0.7772$ and therefore performs slightly better than the one-vs-all approach yielding a mean AUC score of $\overline{\text{AUC}}^{\text{ova}}_{t\bar{t}\gamma} = 0.7764$. The results are summarised in detail in Table 6.2. In addition, the configuration with the higher AUC score is highlighted. The one-vs-one approach performs better than the one-vs-all approach in two out of three compared AUC scores.

Although AUC scores provide a scale-invariant and classification-threshold-invariant

Category	one-vs-one approach	one-vs-all approach
$\overline{\text{AUC}}_{Other\ Prompt}$	0.8230	**0.8258**
$\overline{\text{AUC}}_{Fake}$	**0.7826**	0.7824
$\overline{\text{AUC}}_{t\bar{t}\gamma}$	**0.7772**	0.7764
Outperforms in	2 categories	1 category

Table 6.2.: Table summarising the AUC scores of the one-vs-one and one-vs-all approach. In addition, the configuration with the higher AUC score is highlighted. The one-vs-one approach performs better than the one-vs-all approach in two out of three compared AUC scores.

measure of comparing different classifiers, their interpretation can be difficult due to the lack of additional information providing insight on the origins of the separation performance. To provide a more detailed but less quantitative measure, confusion matrices are calculated for both approaches depicting true positive rates and true negative rates as well as false positive rates and false negative rates. In Figure 6.31 the confusion matrices for the one-vs-one (left) and one-vs-all (right) approach are shown.

To calculate the confusion matrices, the highest value of the output vector of the classifier is taken as the predicted label of the classifier. This label is then compared to the true label of the event. Each row is normalized to unity.
The mean of the 5-fold cross validation and the corresponding standard deviation are also displayed. The standard deviation should not be interpreted as an uncertainty on the mean but rather as a measure of the spread of the five folds, thereby reflecting the dependency of the prediction on the training set.
The one-vs-one multi-class classifier yields has a precision of $64.6 \pm 0.7\,\%$ of correctly predicting $t\bar{t}\gamma$ events. The precision for predicting the correct label is $55.1 \pm 1.2\,\%$ for

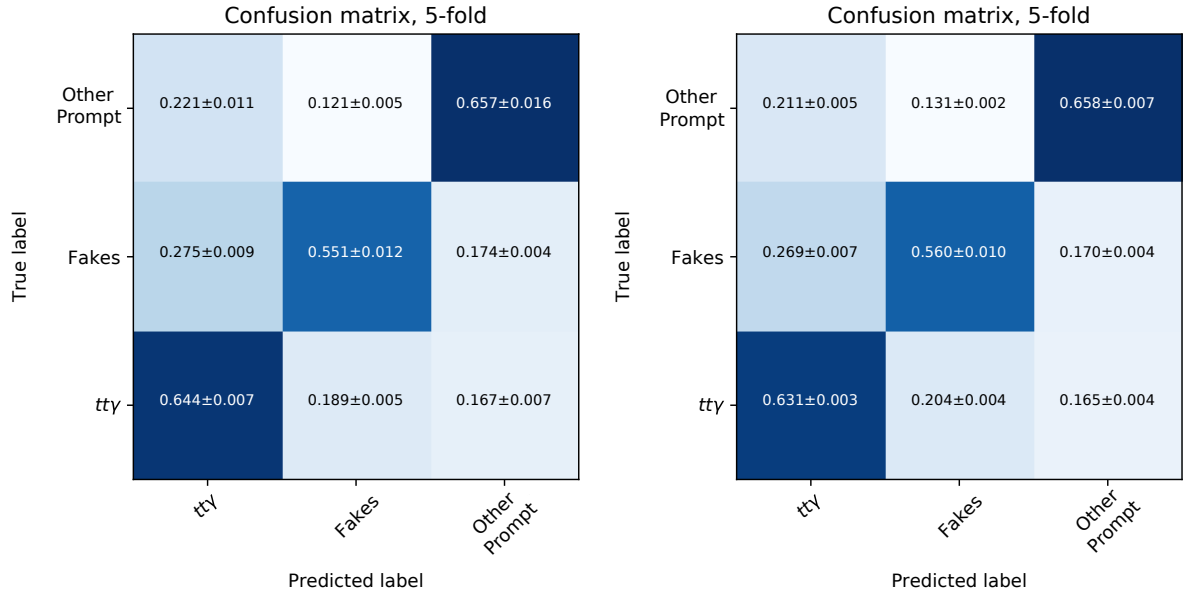

Figure 6.31.: Confusion matrices of the multi-class classifier using the one-vs-one approach (left) and the multi-class classifier using the one-vs-all approach (right).

fake photon background events and $65.7 \pm 1.6\,\%$ for Promptγ background events. In $27.5 \pm 0.9\,\%$ of all cases, fake events are predicted to be signal events. This confusion originates predominantly from $t\bar{t}$ h-fake as well as $t\bar{t}$ e-fake events that have low *Fake* output values. The second highest off-diagonal element represents the confusion between Promptγ events with $t\bar{t}\gamma$ signal events. This term mostly originates from $Wt\gamma$ events within the Promptγ that have high $t\bar{t}\gamma$ output values. The challenge of correctly predicting these events as Promptγ background events can be traced back to their similarity with the $t\bar{t}\gamma$ signal events with which the process interferes at higher order.
Another off-diagonal term that can be explained by the similarity in topologies, is the term describing the wrong prediction of fake events as Promptγ events with a value of $17.4 \pm 0.4\,\%$. This term originates from e-fake events that are classified as Promptγ events due to the similarity of the b-tagging information of these events and $Z\gamma$ events as previously described in Section 6.2.1.
The confusion matrix of the one-vs-all classifier yields similar results. The precision for predicting $t\bar{t}\gamma$ events correctly is $63.1 \pm 0.3\,\%$ which is $1.3\,\%$ lower than the precision achieved by the one-vs-one classifier. Furthermore, the precision for predicting fake and Promptγ events correctly is $56.0 \pm 1.0\,\%$ and $65.8 \pm 0.7\,\%$ respectively. Similar to the results of the one-vs-one approach, the same confusion terms are present for the one-vs-all multi-class classifier.

Separation of background and signal

As presented in Section 6.1, the separation of signal and background is one major motivation for the development of the classifiers in this analysis. The separation power of the two classifiers is investigated by comparing the output distribution of all nodes of the two classifiers and calculating the separation power. The output distributions of the nodes in the output layer are shown in Figure 6.32 for the one-vs-all and in Figure 6.33 for the one-vs-one approach including systematic and statistical uncertainties. The binning of

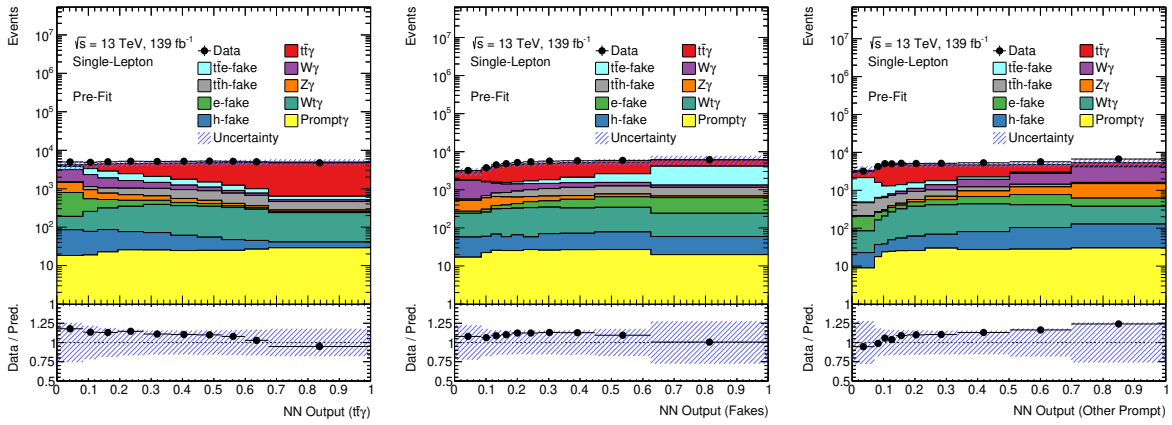

Figure 6.32.: One dimensional distributions for each output node of the multi class classifier using the one-vs-all approach. The distributions of the $t\bar{t}\gamma$, *Fake* and Promptγ output nodes are shown.

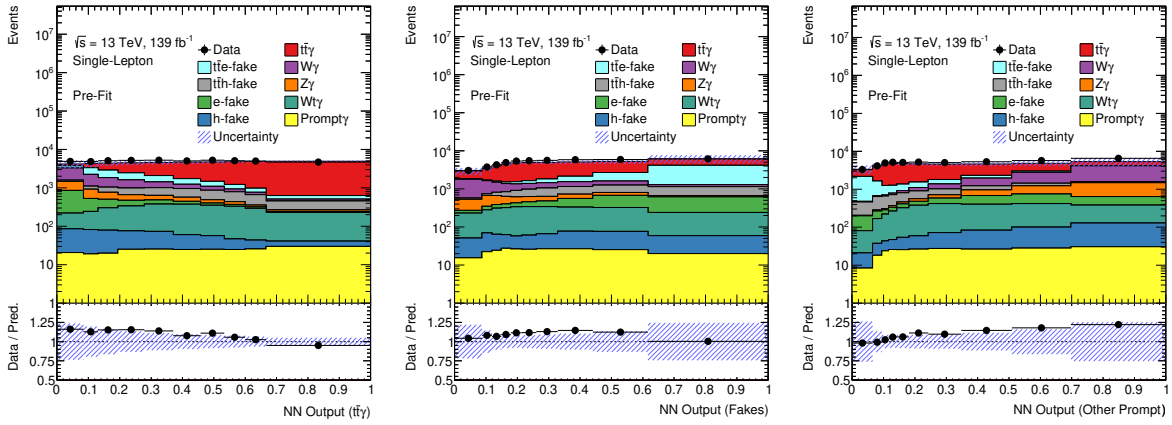

Figure 6.33.: One dimensional distributions for each output node of the multi class classifier using the one-vs-one approach. The distributions of the $t\bar{t}\gamma$, *Fake* and Promptγ output nodes are shown.

the histograms shown above is optimised to yield ten bins, five of which have a signal to background ratio greater than one and five of which have a signal to background ratio

smaller than one. The set of binning parameters is arbitrarily chosen. It is motivated by producing a relatively smooth NN-output distribution that can be used to define cuts for regions with enriched contributions of $t\bar{t}\gamma$, fake and Promptγ events. While the high NN-output region as defined by the rightmost bin in the one-vs-one approach defines a purer $t\bar{t}\gamma$ enriched region than the rightmost bin in the one-vs-all approach, the signal contamination within the high NN-output for *Fake* and *Other Prompt* is higher in the one-vs-one approach than in the one-vs-all approach.

The signal to background ratio of the $t\bar{t}\gamma$ output of the classifiers in the rightmost bin is 6.81 (6.75) for the one-vs-one approach (one-vs-all approach). The signal contamination in the rightmost bin is 0.56 (0.47) and 0.27 (0.26) for the *Fake* output and *Other Prompt* for the one-vs-one approach (one-vs-all approach) respectively.

To further investigate and quantify the performance of both approaches, the separation power is calculated.

In Figure 6.34 the separation for each class compared to the other two classes is shown for the one-vs-all approach. The separation for the one-vs-one approach is separately shown in Figure 6.35. The separation power for $t\bar{t}\gamma$, *Fake* and Promptγ events is 21.2 %, 24.1 %,

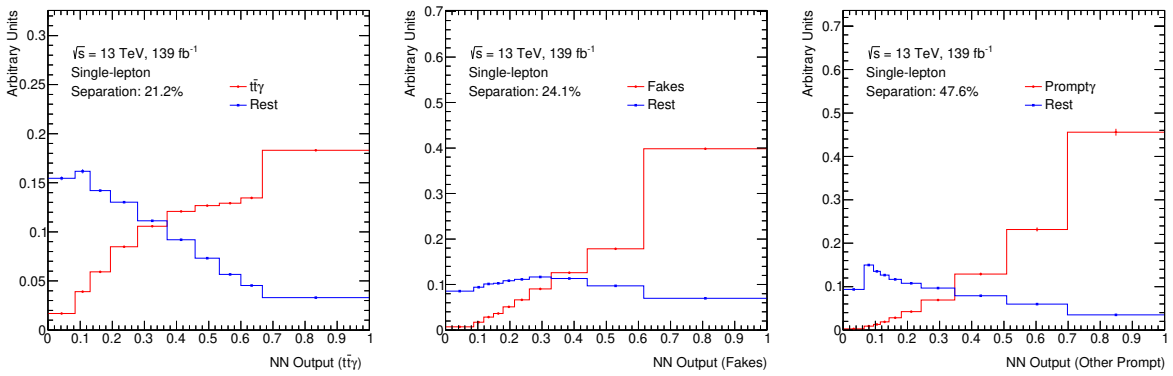

Figure 6.34.: Separation power for each output node of the multi class classifier using the one-vs-all approach. The separation power of the $t\bar{t}\gamma$, *Fake* and Promptγ output nodes are shown.

and 47.6 % in the one-vs-all approach and 21.4 %, 24.4 % and 47.0 % in the one-vs-one approach. Therefore, the one-vs-all is better at separating Promptγ events from all other events and the one-vs-one approach is better at separating $t\bar{t}\gamma$ as well as fake events from all other events.

No significant shape differences between the distributions shown in Figures 6.34 and 6.35 are observed when comparing the output distributions of the three one-vs-all output nodes with the three one-vs-one output nodes.

Both classifiers show discrepancies between MC and data that are not fully covered by the

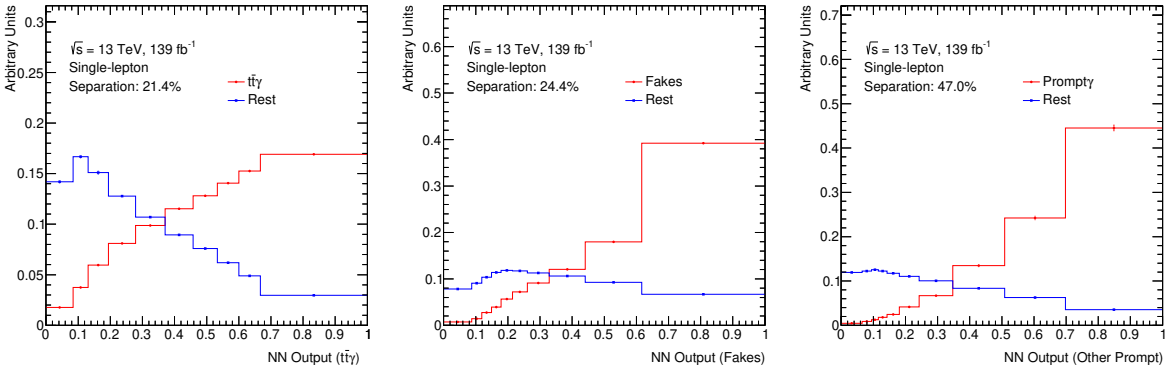

Figure 6.35.: Separation power for each output node of the multi class classifier using the one-vs-one approach. The separation power of the $t\bar{t}\gamma$, *Fake* and Promptγ output nodes are shown.

applied systematic uncertainties[6]. Based on the smoothness of the ROC curves in Figure 6.30, the assumption can be made that these discrepancies originate from improper background modelling. This assumption needs to be tested in more detail in dedicated studies.

Based on the output of the multi-class classifiers, two-dimensional separation plots can be defined utilizing the constraint imposed by the *softmax* activation function:

$$S(y_i) = \sum_{j}^{n} \frac{\exp(y_i)}{\exp(y_j)}. \tag{6.6}$$

In this case, y_i describes the output value of the neural network for node i describing. $S(y_i)$ is equivalent to the probability of the event being a member of class i.

For this purpose, two-dimensional histograms are filled by using the output value of the Promptγ node of each event as the x-value and the output value of the *Fake* node of each event as the y-value. The two-dimensional separation plots for both approaches are shown in Figures 6.36 and 6.37 highlighting the distribution of $t\bar{t}\gamma$, $Z\gamma$, $W\gamma$ and $t\bar{t}$ e-fake events.

It can be seen that both classifiers perform well for the main prompt backgrounds, $W\gamma$ and $Z\gamma$, which have their maximum peak at high *Other Prompt* and low *Fake* output values, thereby highlighting the discriminating power of both multi-class classifiers for these types of events.

Furthermore, the distribution of $t\bar{t}$ e-fake events peaks sharply at high *Fake* and low *Other Prompt* output values with small contributions in the region of the main peak of the $t\bar{t}\gamma$ events at low *Fake* and low *Other Prompt* output values. The leakage of $t\bar{t}$ e-fake events

[6]Since a proper estimation of systematic uncertainties was beyond the scope of this analysis, no final conclusion on these discrepancies can be drawn at this point.

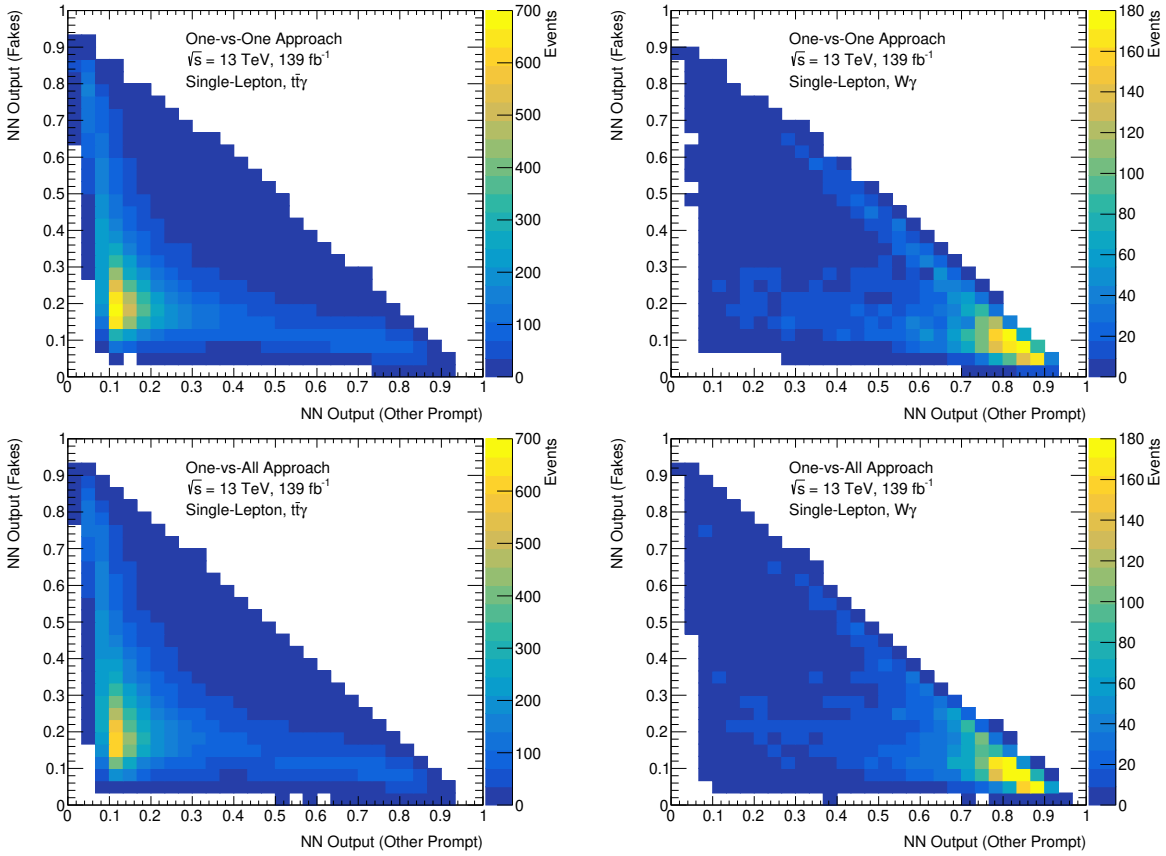

Figure 6.36.: Two-dimensional separation plots of the multi-class classifier using the one-vs-all approach (bottom row) and the one-vs-one approach (top row).

into this enriched $t\bar{t}\gamma$ defined by the low output region in these plots is higher for $t\bar{t}$ e-fakes than for $W\gamma$ and $Z\gamma$ events, leading to residual fake contributions within the enriched $t\bar{t}\gamma$ region.
The contribution of these events within the respective other regions is greatly reduced in all cases. In addition, the distributions for $t\bar{t}$ h-fake, e-fake, h-fake and Promptγ events are shown in Figures 6.38 and C.11 in the appendix.

While the presence of $t\bar{t}$ h-fakes is significantly reduced within the high *Other Prompt* output region, a major leakage of $t\bar{t}$ h-fakes into the enriched $t\bar{t}\gamma$ region at low output values can be seen, originating from the lack of separation power due to non-optimal separation variables as presented in Section 6.2.1 and Section 6.2.2. The contribution of $t\bar{t}$ h-fakes to the enriched signal region can therefore only be reduced and not eliminated. The distribution of e-fake events shows a smear between the *Fake* region in the upper left corner and the *Other Prompt* region in the lower right corner, parallel to the border region imposed by the *softmax*-constraint. This depicts cases were e-fake events are misidentified as *Other Prompt* events. While their leakage into the *Other Prompt* region

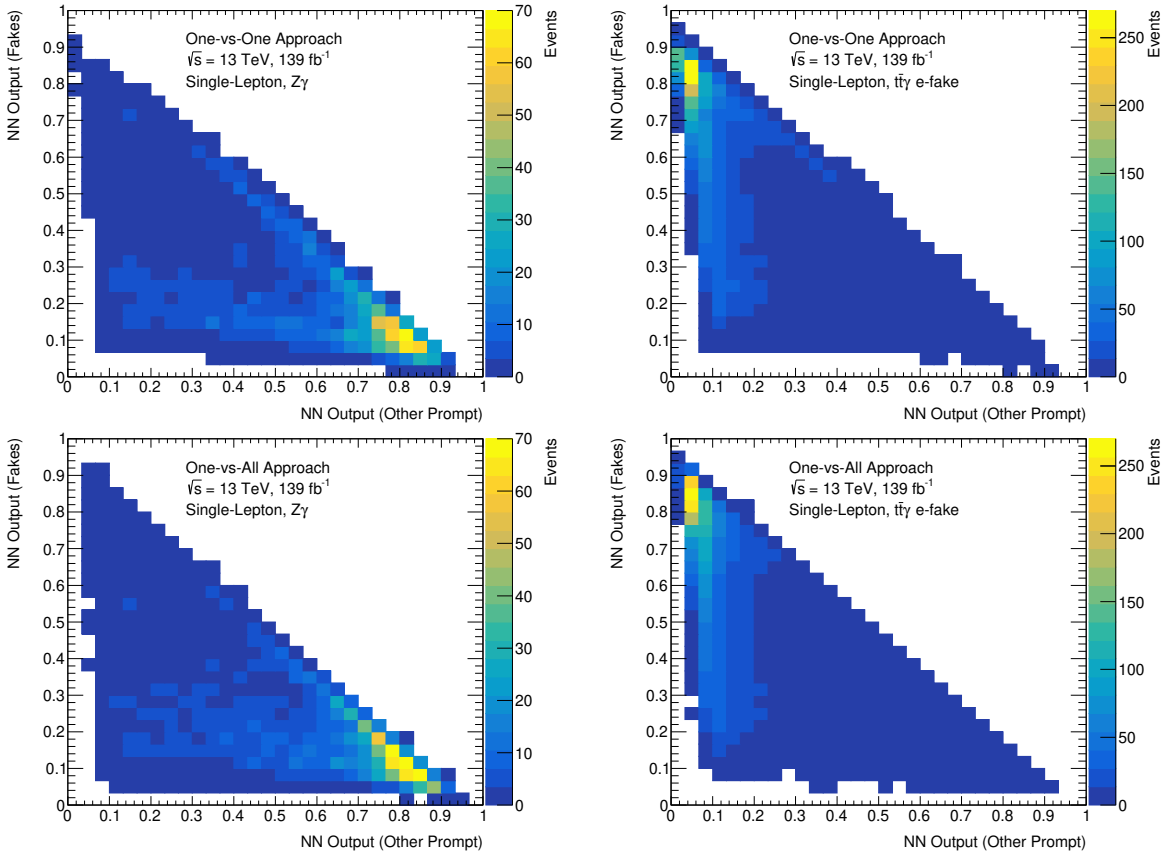

Figure 6.37.: Two-dimensional separation plots of the multi-class classifier using the one-vs-all approach (bottom row) and the one-vs-one approach (top row).

has to be taken into account, the leakage into the $t\bar{t}\gamma$ is negligible.

Hadronic fake events peak at high *Other Prompt* output values and are therefore wrongly classified in almost all cases. This confusion between h-fakes and Promptγ events originates from the fact that no dedicated sub-classifier for h-fakes was developed and their impact on the training of the other sub-classifier is minimal. In addition, most h-fake events can be associated to final states with $W^{\pm}$ or Z^0 bosons, leading to similar event topologies.

To remove this confusion, the impact of h-fakes on the training has to be increased. New ways of weighting the different samples have to be developed to handle this confusion while at the same time not introducing a bias towards h-fake events, which only have a minor overall impact due to the comparably small yields.

Promptγ events comprised of $t\bar{t}V$ and diboson events peak at low *Fake* and at low *Other Prompt* values. This leads to an overlap of these events with the $t\bar{t}\gamma$ events within the enriched signal region. The reason for this overlap is lack of a dedicated classifier for these events. Since these events make up a negligible amount of all background events, their

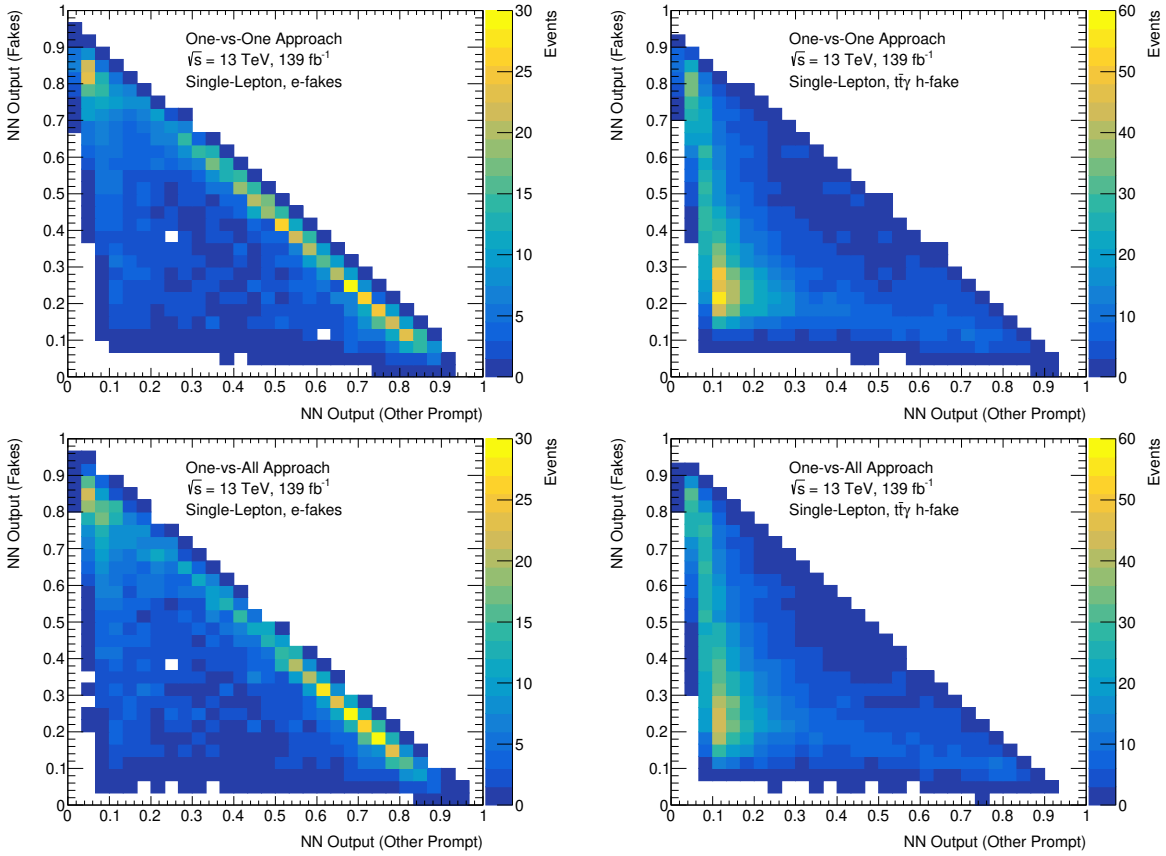

Figure 6.38.: Two-dimensional separation plots of the multi-class classifier using the one-vs-all approach (bottom row) and the one-vs-one approach (top row).

impact can be neglected. No significant improvement of the classifiers is expected when improving the separation for these events. If one would include these events directly in the training by training a dedicated sub-classifier, one would need to make sure that the imbalance between these events and the remaining other events is properly handled. An improper scaling of training weights for such a potential sub-classifier could potentially have a negative impact on the corresponding stacked classifier. As for this sub-classifier and for a potential sub-classifier for h-fakes, the development of new separation methods should always be performed with this potential negative impact in mind.

6.3. Estimating training uncertainties

Systematic uncertainties introduced by the classifier itself are difficult to estimate. Nevertheless, an attempt is made in this thesis to estimate the bin-by-bin deviations introduced by the two different multi-class classifiers. The estimation of uncertainties is based on the assumption that the classifiers are not overfitting but several possible and similar

separation boundaries exist within the hyperspace defined by the input variables. During the training the classifiers are expected to approach the minimum of the loss function, however the exact value of the loss function after training the classifier differs after each training iteration due to the random initialisation of network weights and biases. A different loss value is equivalent to a different decision boundary within the hyperspace. Nevertheless, it is expected that the distance between these decision boundaries within the hyperspace calculated using a suitable norm is small[7]. Based on this chain of thought, the dependency of the training set and the initialisation on the results calculated using the testing set is studied. For this purpose, the one-vs-one classifier and the one-vs-all classifier are trained 100 times each. During each training, the training and validation set is randomly drawn for each test fold. The variance of the predictions is then determined by evaluating the classifier using the same testing set for each of the 100 trained models and each fold separately.
The measured standard deviation is then treated as a bin-by-bin uncertainty on the predictions for the classifier due to the training process itself.
The relative uncertainties σ/μ, where σ and μ are the measured standard deviation and mean number of events per bin of the 100 trained models, are shown for the $t\bar{t}\gamma$ output node in Figure 6.39 for both approaches.

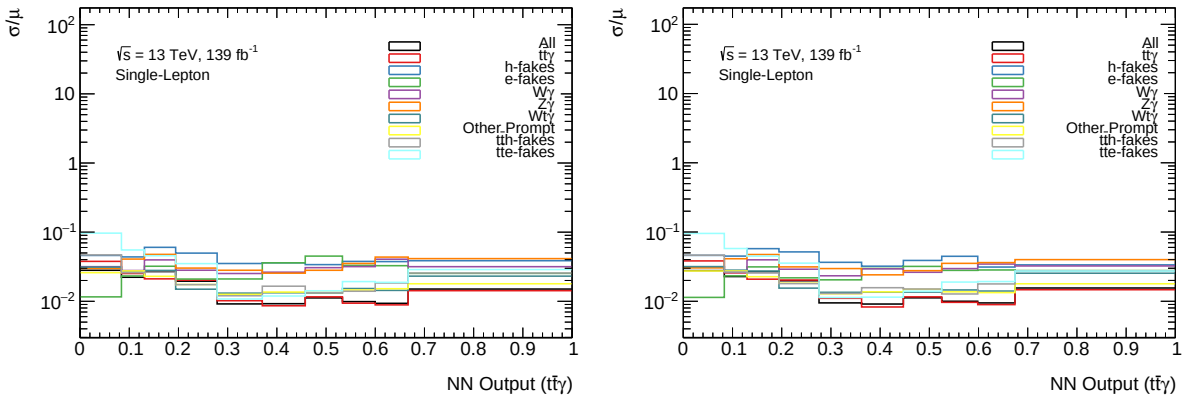

Figure 6.39.: Measured relative bin-by-bin uncertainties for the $t\bar{t}\gamma$ output class using the one-vs-one approach (left) and the one-vs-all approach (right).

Additional distributions are shown in Figures C.12 and C.13 in the appendix for the *Fake* and *Other Prompt* output nodes. The relative uncertainty is highest for h-fakes and e-fakes for most bins and lowest for $t\bar{t}\gamma$ signal events for most bins.
The interpretation of this uncertainty is non-trivial since the migration of events from one bin to another is not taken into account. Therefore, the events within each bin can differ

[7]High discrepancies between several trainings or a significant number of outliers would indicate that no global minimum of the loss function was reached during training.

from events in the same bin when using another training result. As a result, the relative uncertainties presented in Figure 6.39 should be interpreted as an additional measure of overfitting and as a tool to determine whether a global minimum was reached during the training. For the classifiers presented in this thesis the presented uncertainties support the hypothesis that a global minimum of the loss function was indeed reached.
In this thesis, a 5-fold cross validation was used. Therefore, the MC prediction for all output variables consist of stacked prediction of all five folds, where each fold represents the MC prediction equivalent to a fifth of the entire dataset. The evaluation of the proton-proton collision data is performed using the mean of all five folds since the data was not used in any training. In a case where the difference in the mean response value of these five folds is larger than the bin width used in the binned output histogram, systematic differences between the MC prediction and the evaluated data could arise. This was not directly studied within the scope of this thesis, but the effect is assumed to be small since none of the presented ROC curves suggests the presence of such significant differences.

7. Summary, conclusion and outlook

Studying the $t\bar{t}\gamma$ process provides an opportunity to study the charge of the top quark and the structure of the electromagnetic coupling of the photon to the top quark. Using the measured properties of the particles detected by the ATLAS detector does not provide sufficient discriminating power to separate different background topologies from the $t\bar{t}\gamma$ process.

Different multivariate analysis techniques can be used to improve the discrimination between signal and different types of background. One possible analysis strategy uses deep multi-class neural networks, as used in this analysis [72].

Dedicated stacked ensembles following a multi-class one-vs-one and one-vs-all approach are investigated leading to two different multi-class classifiers. Their performance in the scope of a $t\bar{t}\gamma$ cross section measurement is presented.

To optimise the multi-class classifiers, a dedicated set of input variables is chosen based on their separation power for each unique background topology as well as the topology of the $t\bar{t}\gamma$ signal. Further optimisation is based on optimising a set of sub-classifiers using a genetic optimisation algorithm which is executed using the open-source high-throughput computing software framework *HTCondor* in a diamond shaped directed acyclic graph.

The training of these networks is performed using Monte-Carlo truth information, which also incorporates the labels of the processes in question. Negative event weights are set to zero for the training. While the binary cross entropy is used as a loss function for all sub-classifiers, the categorical cross entropy is used for the multi-class classifiers.

All sub-classifiers are evaluated and discussed separately to provide a detailed description of their performance and impact on the corresponding one-vs-one or one-vs-all approach. For this purpose, the area under the receiver operating characteristics curve is used as a performance measure and overfitting test. Separation plots and binned distributions of the output value of these classifiers are computed. Pearson correlation coefficients are presented to evaluate the impact of input variables on the different classifiers.

Based on these sets of sub-classifiers, the two multi-class classifiers utilizing the presented one-vs-one and one-vs-all approach are trained. These multi-class classifiers separate events into three distinct classes, namely $t\bar{t}\gamma$, *Fake* and *Other Prompt* in a three dimen-

sional output vector with unit length.
Through the unity constraint imposed by the used *softmax* activation function in the last layer of the two multi-class classifiers, two-dimensional probability distributions are defined. Based on these two-dimensional distributions, new control and signal regions are defined, yielding one enriched signal region, a region with an enriched fake contribution and a region with an enriched contribution of prompt photon background events.
The definition of the cuts resulting in these regions are not yet optimised but similar for both multi-class approaches. By using these newly defined regions, constraints can be imposed on the prompt and fake photon background processes. Furthermore, a refined signal cut can be deployed to significantly increase the $t\bar{t}\gamma$ purity. Out of the two presented classifiers the one-vs-one performs slightly better than the one-vs-all classifier. It should be considered for a future measurement of the $t\bar{t}\gamma$ cross section in ATLAS.

Outlook

Machine learning algorithms such as deep neural networks heavily depend on the quality of input data that is fed into the network. Furthermore, the choice of labels influences the separation power and performance of the studied classifier. In this thesis, h-fake, e-fake as well as $t\bar{t}V$ and diboson events are grouped together into single sub-classes. However, these classes are comprised of different topologies and only have the nature of the photon in common. Since most of the discriminating variables used in this thesis focus on the underlying event kinematics, these mixtures of events within one sub-class is not optimal and leads to a decreased performance of the tested multi-class classifiers. For future studies, the choice of labels as well as the effect on the training should be revisited.
To take the imbalance between different classes into account, the minority classes are scaled up to adequately treat them during training. However, this rescaling of training weights was performed on a class-by-class basis where each class contained several sub-classes such as $W\gamma$ or $Z\gamma$ in case of the $WZ\gamma$ sub-classifier. With sufficient statistics within all underlying sub-classes, the rescaling should be based on these sub-classes and not the classes themselves which can be arbitrary. The impact of this type of scaling was not investigated during this thesis and should be considered in future iterations.
In addition, the search for new physics motivated variables to separate the different classes should continue. An emphasis should be put on finding more variables with high separation power to separate fake photons from prompt photons. This was previously achieved by introducing a *prompt-photon-tagger* [91]. This method unfortunately also introduced

significant new systematic uncertainties, which reduced the impact of the prompt-photon-tagger.

This problem can only be solved by studying the underlying photon shower-shape variables in detail and solving the observed discrepancies between simulation and data [66, 87].

From a ML standpoint, this analysis introduced a complex and dedicated multi-class classifier into the $t\bar{t}\gamma$ cross section analysis. However, an increase in separation performance is expected if the approach is changed by using a generative adversarial neural network (GAN) [92, 93] or a convolutional neural network (CNN) taking an abstract parameter space of detector information as its input. Combining the concept of two networks contesting with each other in a zero-sum game, a GAN approach, with the idea of a CNN using detector information to form an abstract image of the topology, might yield significant performance improvements. Therefore, a study of a GAN using a CNN should be investigated.

Bibliography

[1] S. L. Glashow, *Partial Symmetries of Weak Interactions*, Nucl. Phys. **22**, 579 (1961)

[2] A. Salam, *Weak and Electromagnetic Interactions*, Conf. Proc. **C680519**, 367 (1968)

[3] S. L. Glashow, J. Iliopoulos, L. Maiani, *Weak Interactions with Lepton-Hadron Symmetry*, Phys. Rev. **D2**, 1285 (1970)

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

[5] P. A. M. Dirac, *Quantum theory of emission and absorption of radiation*, Proc. Roy. Soc. Lond. **A114**, 243 (1927)

[6] J. Schwinger, *On Quantum electrodynamics and the magnetic moment of the electron*, Phys. Rev. **73**, 416 (1948)

[7] J. Schwinger, *Quantum electrodynamics. I. A covariant formulation*, Phys. Rev. **74**, 1439 (1948)

[8] R. P. Feynman, *Space-time approach to nonrelativistic quantum mechanics*, Rev. Mod. Phys. **20**, 367 (1948)

[9] H. D. Politzer, *Asymptotic Freedom: An Approach to Strong Interactions*, Phys. Rept. **14**, 129 (1974)

[10] W. J. Marciano, H. Pagels, *Quantum Chromodynamics: A Review*, Phys. Rept. **36**, 137 (1978)

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

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

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

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

[15] Y. Fukuda, et al. (Super-Kamiokande), *Evidence for oscillation of atmospheric neutrinos*, Phys. Rev. Lett. **81**, 1562 (1998)

[16] C. L. Bennett, et al. (WMAP Collaboration), *Nine-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and Results*, Astrophys. J. Suppl. **208**, 20 (2013)

[17] G. Hinshaw, et al. (WMAP Collaboration), *Nine-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results*, Astrophys. J. Suppl. **208**, 19 (2013)

[18] P. A. R. Ade, et al. (Planck Collaboration), *Planck 2013 results. XVI. Cosmological parameters*, Astron. Astrophys. **571**, A16 (2014)

[19] M. Tegmark, et al. (SDSS Collaboration), *Cosmological parameters from SDSS and WMAP*, Phys. Rev. **D69**, 103501 (2004)

[20] D. N. Spergel, et al. (WMAP Collaboration), *First year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Determination of cosmological parameters*, Astrophys. J. Suppl. **148**, 175 (2003)

[21] V. C. Rubin, W. K. Ford, Jr., *Rotation of the Andromeda Nebula from a Spectroscopic Survey of Emission Regions*, Astrophys. J. **159**, 379 (1970)

[22] M. Bartelmann, P. Schneider, *Weak gravitational lensing*, Phys. Rept. **340**, 291 (2001)

[23] F. Abe, et al. (CDF), *Observation of top quark production in $\bar{p}p$ collisions*, Phys. Rev. Lett. **74**, 2626 (1995)

[24] S. Abachi, et al. (D0 Collaboration), *Observation of the top quark*, Phys. Rev. Lett. **74**, 2632 (1995)

[25] ATLAS Collaboration, *Measurement of the top quark mass in the $t\bar{t} \to$ lepton+jets channel from $\sqrt{s} = 8$ TeV ATLAS data and combination with previous results*, Eur. Phys. J. **C79(4)**, 290 (2019)

[26] CMS Collaboration, *Measurement of the top quark mass in the all-jets final state at $\sqrt{s} = 13$ TeV and combination with the lepton+jets channel*, Eur. Phys. J. **C79(4)**, 313 (2019)

[27] T. Affolder, et al. (CDF Collaboration), *First measurement of the ratio* $B(t \to Wb)/B(t \to Wq)$ *and associated limit on the CKM element* $|V_{tb}|$, Phys. Rev. Lett. **86**, 3233 (2001)

[28] N. Cabibbo, *Unitary Symmetry and Leptonic Decays*, Phys. Rev. Lett. **10**, 531 (1963)

[29] M. Kobayashi, T. Maskawa, *CP Violation in the Renormalizable Theory of Weak Interaction*, Prog. Theor. Phys. **49**, 652 (1973)

[30] ATLAS Collaboration, *Measurements of the top quark branching ratios into channels with leptons and quarks with the ATLAS detector*, Phys. Rev. **D92(7)**, 072005 (2015)

[31] ATLAS Collaboration, *Observation of top-quark pair production in association with a photon and measurement of the* $t\bar{t}\gamma$ *production cross section in pp collisions at* $\sqrt{s} = 7$ *TeV using the ATLAS detector*, Phys. Rev. **D91(7)**, 072007 (2015)

[32] ATLAS Collaboration, *Measurement of the top quark charge in pp collisions at* $\sqrt{s} =$ *7 TeV with the ATLAS detector*, JHEP **11**, 031 (2013)

[33] U. Baur, et al., *Probing Electroweak Top Quark Couplings at Hadron and Lepton Colliders*, Nucl. Phys. Proc. Suppl. **160**, 17 (2006)

[34] O. Bylund, et al., *Probing top quark neutral couplings in the Standard Model Effective Field Theory at NLO in QCD*, JHEP **2016(5)**, 52 (2016)

[35] CMS Collaboration, *Measurement of the* $Z\gamma$ *Production Cross Section in pp Collisions at 8 TeV and Search for Anomalous Triple Gauge Boson Couplings*, JHEP **04**, 164 (2015)

[36] ATLAS Collaboration, *Measurement of* $W^{\pm}$ *and* Z*-boson production cross sections in pp collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, Phys. Lett. **B759**, 601 (2016)

[37] A. Kirchhoff, *Studies to classify prompt photons in the* $t\bar{t}\gamma$*-process*, Master's thesis, Georg-August-Universität Göttingen, II.Physik-UniGö-MSc-2018/04 (2018)

[38] G. Arnison, et al. (UA1 Collaboration), *Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at s**(1/2) = 540-GeV*, Phys. Lett. **B122**, 103 (1983)

[39] G. Arnison, et al. (UA1 Collaboration), *Experimental Observation of Lepton Pairs of Invariant Mass Around 95-GeV/c**2 at the CERN SPS Collider*, Phys. Lett. **B126**, 398 (1983)

[40] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3**, S08003 (2008)

[41] ALICE Collaboration, *The ALICE experiment at the CERN LHC*, JINST **3**, S08002 (2008)

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

[43] LHCb Collaboration, *The LHCb Detector at the LHC*, JINST **3**, S08005 (2008)

[44] ATLAS Collaboration, *Performance of the ATLAS Trigger System in 2015*, Eur. Phys. J. **C77(5)**, 317 (2017)

[45] S. Agostinelli, et al. (GEANT4), *GEANT4: A Simulation toolkit*, Nucl. Instrum. Meth. **A506**, 250 (2003)

[46] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, Eur. Phys. J. **C70**, 823 (2010)

[47] T. Gleisberg, et al., *Event generation with SHERPA 1.1*, JHEP **02**, 007 (2009)

[48] S. Hoeche, et al., *QCD matrix elements and truncated showers*, JHEP **05**, 053 (2009)

[49] D. J. Lange, et al., *The EvtGen particle decay simulation package*, Nucl. Instrum. Methods A **462**, 152 (2001)

[50] T. Sjostrand, S. Mrenna, P. Z. Skands, *PYTHIA 6.4 Physics and Manual*, JHEP **05**, 026 (2006)

[51] T. Sjöstrand, et al., *An Introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191**, 159 (2015)

[52] A. D. Martin, et al., *Parton distributions for the LHC*, Eur. Phys. J. **C63**, 189 (2009)

[53] J. Alwall, et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07**, 079 (2014)

[54] J. Pumplin, et al., *New generation of parton distributions with uncertainties from global QCD analysis*, JHEP **07**, 012 (2002)

[55] N. Kidonakis, *Next-to-next-to-leading-order collinear and soft gluon corrections for t-channel single top quark production*, Phys. Rev. **D83**, 091503 (2011)

[56] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms*, JHEP **11**, 040 (2004)

[57] S. Frixione, B. Webber, *Matching NLO QCD computations and parton shower simulations*, JHEP **06**, 029 (2002)

[58] S. Alioli, et al., *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP **06**, 043 (2010)

[59] S. Frixione, P. Nason, C. Oleari, *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*, JHEP **11**, 070 (2007)

[60] R. D. Ball, et al. (NNPDF), *Parton distributions for the LHC Run II*, JHEP **04**, 040 (2015)

[61] M. Czakon, A. Mitov, *Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders*, Comput. Phys. Commun. **185**, 2930 (2014)

[62] J. M. Campbell, R. K. Ellis, *An Update on vector boson pair production at hadron colliders*, Phys. Rev. **D60**, 113006 (1999)

[63] D. de Florian, et al. (LHC Higgs Cross Section Working Group), *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector* (2016)

[64] ATLAS Collaboration, *Electron and Photon Reconstruction with the ATLAS Detector*, Nucl. Part. Phys. Proc. **273-275**, 2539 (2016)

[65] ATLAS Collaboration, *Photon identification with the ATLAS detector*, PoS **EPS-HEP2017**, 760 (2017)

[66] ATLAS Collaboration, *Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run-1 data*, Eur. Phys. J. **C76(12)**, 666 (2016)

[67] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at* $\sqrt{s}$ *=13 TeV*, Eur. Phys. J. **C76(5)**, 292 (2016)

[68] M. Cacciari, G. P. Salam, G. Soyez, *The anti-*k_t *jet clustering algorithm*, JHEP **04**, 063 (2008)

[69] ATLAS Collaboration, *ATLAS b-tagging performance during LHC Run-2 with the new Insertable B-layer*, in *Proceedings, 3rd Large Hadron Collider Physics Conference (LHCP 2015): St. Petersburg, Russia, August 31-September 5, 2015*, pages 699–704, Kurchatov Institute, Kurchatov Institute, Gatchina (2016)

[70] ATLAS Collaboration, *Performance and calibration of b-tagging with the ATLAS experiment at LHC Run-2*, Technical Report ATL-PHYS-PROC-2016-193, Geneva (2016)

[71] ATLAS Collaboration, *Performance of b-Jet Identification in the ATLAS Experiment*, JINST **11(04)**, P04008 (2016)

[72] ATLAS Collaboration, *Measurements of inclusive and differential fiducial cross-sections of $t\bar{t}\gamma$ production in leptonic final states at $\sqrt{s} = 13$ TeV in ATLAS*, Eur. Phys. J. **C79(5)**, 382 (2019)

[73] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC* (2019)

[74] H. Voss, et al., *TMVA, the Toolkit for Multivariate Data Analysis with ROOT*, PoS **ACAT**, 040 (2007)

[75] L. Breiman, *Random Forests*, Machine Learning **45(1)**, 5 (2001)

[76] Baldi, et al., *Jet Substructure Classification in High-Energy Physics with Deep Neural Networks*, Phys. Rev. **D93(9)**, 094034 (2016)

[77] H. J. Kelley, *Gradient theory of optimal flight paths*, Ars Journal **30(10)**, 947 (1960)

[78] M. Abadi, et al., *TensorFlow: A system for large-scale machine learning*, in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283 (2016)

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

[80] M. Blondel et. al., *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research **12**, 2825 (2011)

[81] R. Brun, F. Rademakers, *ROOT: An object oriented data analysis framework*, Nucl. Instrum. Meth. **A389**, 81 (1997)

[82] T. E. Oliphant, *A guide to NumPy*, volume 1, Trelgol Publishing USA (2006)

[83] D. Kingma, J. Ba, *Adam: A Method for Stochastic Optimization*, Proceedings of the 3rd International Conference on Learning Representations (2014)

[84] S. J. Reddi, S. Kale, S. Kumar, *On the Convergence of Adam and Beyond*, CoRR **abs/1904.09237** (2019)

[85] G. Dahl, et al., *On the importance of initialization and momentum in deep learning*, in S. Dasgupta, D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, PMLR, Atlanta, Georgia, USA (2013)

[86] N. Srivastava, et al., *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, J. Mach. Learn. Res **15**, 1929 (2014)

[87] ATLAS Collaboration, *Measurement of the photon identification efficiencies with the ATLAS detector using LHC Run 2 data collected in 2015 and 2016*, Eur. Phys. J. **C79(3)**, 205 (2019)

[88] J. Bergstra, Y. Bengio, *Random Search for Hyper-parameter Optimization*, J. Mach. Learn. Res. **13**, 281 (2012)

[89] K. Pearson, *Note on regression and inheritance in the case of two parents*, Proceedings of the Royal Society of London **58**, 240 (1895)

[90] T. Schröer, *Studien zu Untergründen mit prompten Photonen für die Analyse von $t\bar{t}$+Photon-Ereignissen mit dem ATLAS-Experiment*, Bachelor's thesis, Georg-August-Universität Göttingen, II. Physik-UniGö-BSc-2019/01 (2019)

[91] B. Völkel, *Studies of the discrimination between prompt photons and hadron fakes using neural networks*, Master's thesis, Georg-August-Universität Göttingen, II.Physik-UniGö-MSc-2017/07 (2017)

[92] I. Goodfellow, et al., *Generative Adversarial Nets*, in Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680, Curran Associates, Inc. (2014)

[93] T. Salimans, et al., *Improved Techniques for Training GANs*, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2234–2242, Curran Associates Inc., USA (2016)

Acknowledgements

The preparation of this thesis gave me the opportunity to learn many new skills and contribute to the field of experimental particle physics.
I would like to thank Prof. Dr. Arnulf Quadt, for providing me with the unique opportunity of working in his ATLAS working group. Working within this group was a pleasure and filled with discussion about physics and non-physics topics, e.g. during coffee breaks. This thesis would also not have been possible without the help I received from Knut. You helped me a lot with all kinds of physics and computational problems I faced. I value your opinion a lot and it was really helpful to discuss ideas with you in person. Thanks to you and Thomas for proofreading the thesis. I would also like to thank everyone in the institute for a pleasant time. Working with friends just makes the day much better.
Special thanks to my family for your support throughout all the years. This thesis would not have been possible without you. Most important: Thank you Inga! You back me up a lot and your support is just awesome. It is incredibly great to have you in my life.

A. Mathematical functions

Binary cross entropy

The binary cross entropy loss function $\text{Loss}(y, y_0)$ for $m = 2$ classes in a classification problem is defined as:

$$\text{Loss}(y, y_0) = -(y_0 \log(y) + (1 - y_0) \log(1 - y). \tag{A.1}$$

Categorical cross entropy

The categorical cross entropy loss function $\text{Loss}(\vec{y}, \vec{y_0})$ for $m = 3$ classes in a classification problem is defined as:

$$\text{Loss}(\vec{y}, \vec{y_0}) = -\sum_{c=1}^{3} y_{0,c} \log(y_c). \tag{A.2}$$

A.1. ELU activation function

The "Exponential Linear Unit" (ELU), $\text{ELU}(x)$, is a function given by:

$$\text{ELU}(x, \alpha = 1.0) = \begin{cases} x & \text{if } x > 0 \\ \alpha(\exp(z) - 1) & \text{if } x \leq 0 \end{cases} \tag{A.3}$$

A.2. ReLU activation function

The " Rectified Linear Unit" (ReLU), $\text{ReLU}(x)$, is a function given by:

$$\text{ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{A.4}$$

A.3. Tanh activation function

The $\tanh(x)$ function is given by:

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}. \tag{A.5}$$

B. Additional tables

The results of the GOA are presented in Table B.1 and Appendix B.

	Test AUC	Test AUC Std	Train AUC	Train AUC Std	Difference
$t\bar{t}\gamma$ Best AUC	0.7791	0.0017	0.7816	0.0020	0.0025
$t\bar{t}\gamma$ Chosen	0.7722	0.0011	0.7728	0.0010	0.0006
$WZ\gamma$ Best AUC	0.8800	0.0016	0.8836	0.0003	0.0036
$WZ\gamma$ Chosen	0.8750	0.0009	0.8757	0.0004	0.0007
$t\bar{t}$ h-fake Best AUC	0.7207	0.0031	0.7295	0.0028	0.0088
$t\bar{t}$ h-fake Chosen	0.6833	0.0009	0.6839	0.0031	0.0006
$t\bar{t}$ e-fake Best AUC	0.8349	0.0025	0.8382	0.0024	0.0034
$t\bar{t}$ e-fake Chosen	0.7900	0.0009	0.7903	0.0020	0.0003
e-fake Best AUC	0.8462	0.0021	0.8533	0.0035	0.0071
e-fake Chosen	0.8182	0.0093	0.8190	0.0120	0.0007
$Wt\gamma$ Best AUC	0.6797	0.0040	0.6926	0.0027	0.0129
$Wt\gamma$ Chosen	0.6443	0.0079	0.6446	0.0028	0.0004

Table B.1.: Table summarising the performance results of the GOA for the SCLs of the one-vs-all approach.

	Layers	Nodes	Activation	Dropout
$t\bar{t}\gamma$ Best AUC	3	100, 100, 100	*ReLU*,*ReLU*,*tanh*	2
$t\bar{t}\gamma$ Chosen	4	100, 30, 200, 50	*tanh*,*ReLU*,*ReLU*,sigmoid	0, 2, 3
$WZ\gamma$ Best AUC	2	200, 50	*ReLU*,*ReLU*	1
$WZ\gamma$ Chosen	2	30, 30	*tanh*,*tanh*	2
$t\bar{t}$ h-fake Best AUC	4	200, 50, 100, 50	*tanh*,*tanh*,*ReLU*,*ELU*	2
$t\bar{t}$ h-fake Chosen	2	100, 100	sigmoid,*ELU*	1
$t\bar{t}$ e-fake Best AUC	4	200, 200, 50, 50	*ELU*,*tanh*,*ReLU*,*ReLU*	0
$t\bar{t}$ e-fake Chosen	2	30, 50	*tanh*,*tanh*	2, 4
e-fake Best AUC	4	50, 30, 100, 30	*tanh*,*ReLU*,*ELU*,*ELU*	1, 2
e-fake Chosen	3	30, 30, 30	*ELU*,*ELU*,sigmoid	0, 2
$Wt\gamma$ Best AUC	3	100, 50, 30	*tanh*,*ReLU*,*tanh*	0, 1, 2
$Wt\gamma$ Chosen	2	100, 50	*ELU*,sigmoid	1

Table B.2.: Table summarising the architecture results of the GOA for the SCLs of the one-vs-all approach.

	Test AUC	Test AUC Std	Train AUC	Train AUC Std	Difference
$WZ\gamma$ Best AUC	0.9075	0.0039	0.9121	0.0008	0.0046
$WZ\gamma$ Chosen	0.9054	0.0048	0.9093	0.0009	0.0039
$t\bar{t}$ h-fake Best AUC	0.7332	0.0020	0.7430	0.0011	0.0098
$t\bar{t}$ h-fake Chosen	0.7218	0.0052	0.7245	0.0034	0.0027
$t\bar{t}$ e-fake Best AUC	0.8538	0.0032	0.8605	0.0015	0.0068
$t\bar{t}$ e-fake Chosen	0.8508	0.0039	0.8547	0.0009	0.0039
e-fake Best AUC	0.9238	0.0026	0.9340	0.0016	0.0102
e-fake Chosen	0.9099	0.0057	0.9153	0.0020	0.0054
$Wt\gamma$ Best AUC	0.6960	0.0061	0.7078	0.0019	0.0118
$Wt\gamma$ Chosen	0.6852	0.0086	0.6924	0.0093	0.0071

Table B.3.: Table summarising the performance results of the GOA for the SCLs of the one-vs-one approach.

	Layers	Nodes	Activation	Dropout
$WZ\gamma$ Best AUC	3	50, 50, 50	ELU,ELU,ELU	0, 2
$WZ\gamma$ Chosen	2	50, 50	ELU,ELU	0
$t\bar{t}$ h-fake Best AUC	3	100, 50, 50	ELU,ELU,$tanh$	1
$t\bar{t}$ h-fake Chosen	2	30, 50	ELU,ELU	0, 1
$t\bar{t}$ e-fake Best AUC	2	100, 100	$tanh$,$tanh$	0
$t\bar{t}$ e-fake Chosen	2	50, 30	$tanh$,$tanh$	0
e-fake Best AUC	3	100, 50, 30	$tanh$,$ReLU$,$tanh$	0, 1
e-fake Chosen	2	50, 30	ELU,$tanh$	1
$Wt\gamma$ Best AUC	3	100, 100, 100	ELU,$tanh$,ELU	0, 1, 2
$Wt\gamma$ Chosen	2	50, 30	ELU,$tanh$	0, 1

Table B.4.: Table summarising the architecture results of the GOA for the SCLs of the one-vs-one approach.

C. Additional figures

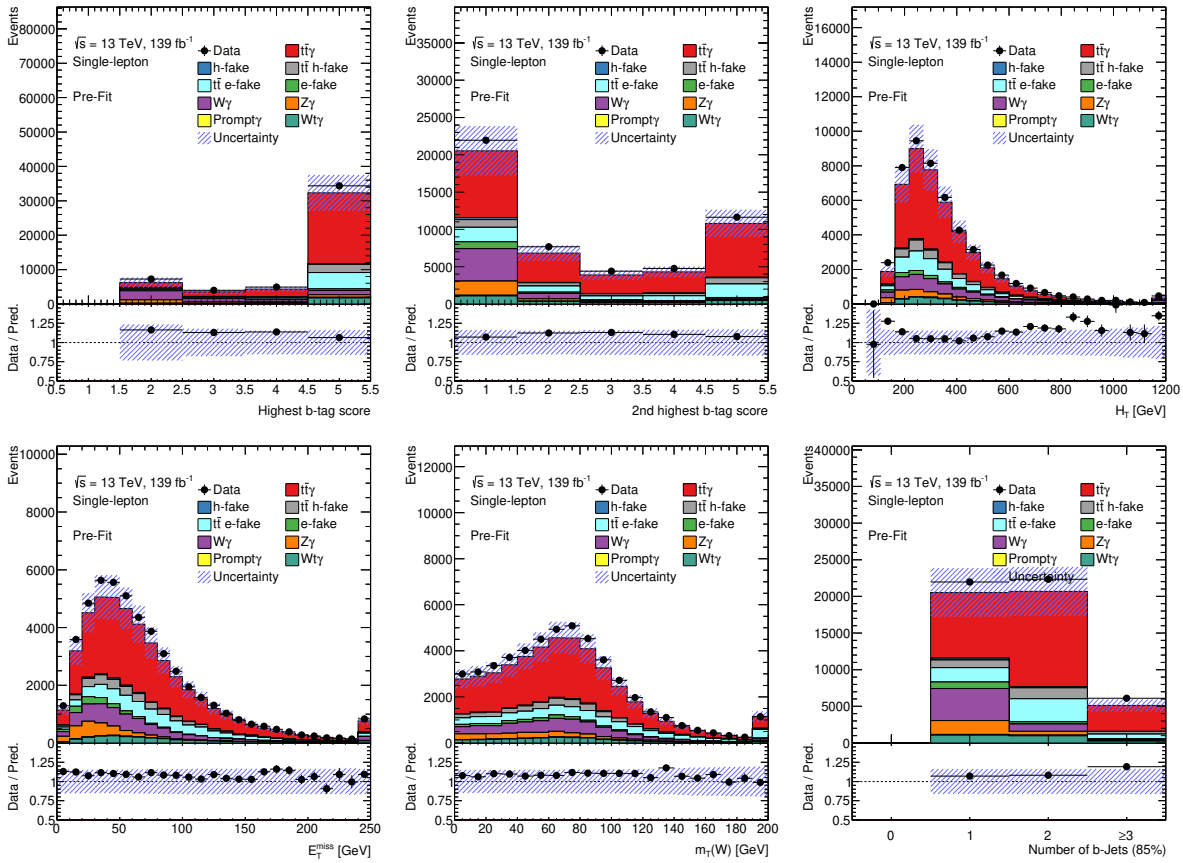

Figure C.1.: Kinematic distributions of the used input variables.

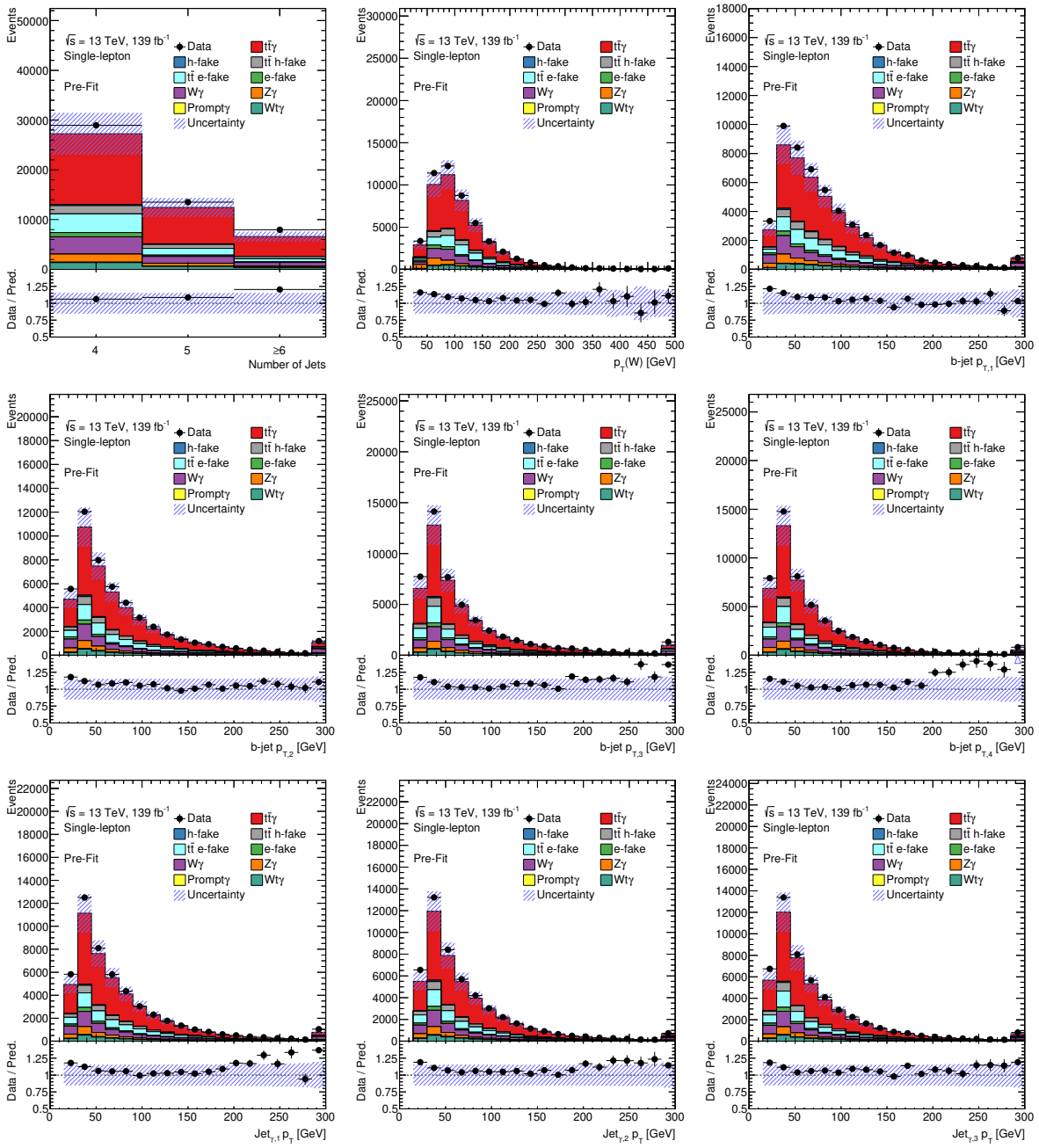

Figure C.2.: Kinematic distributions of the used input variables.

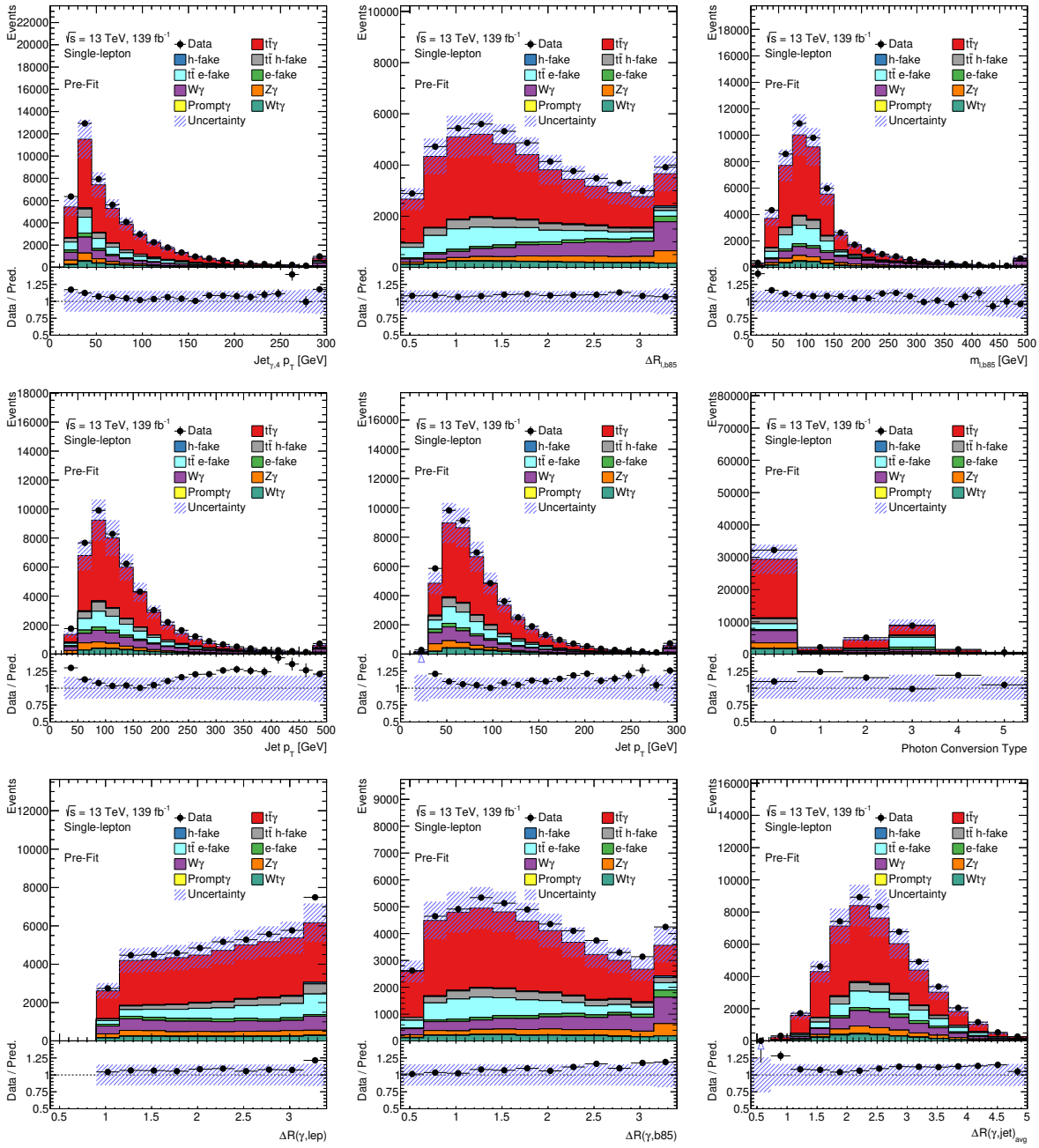

Figure C.3.: Kinematic distributions of the used input variables.

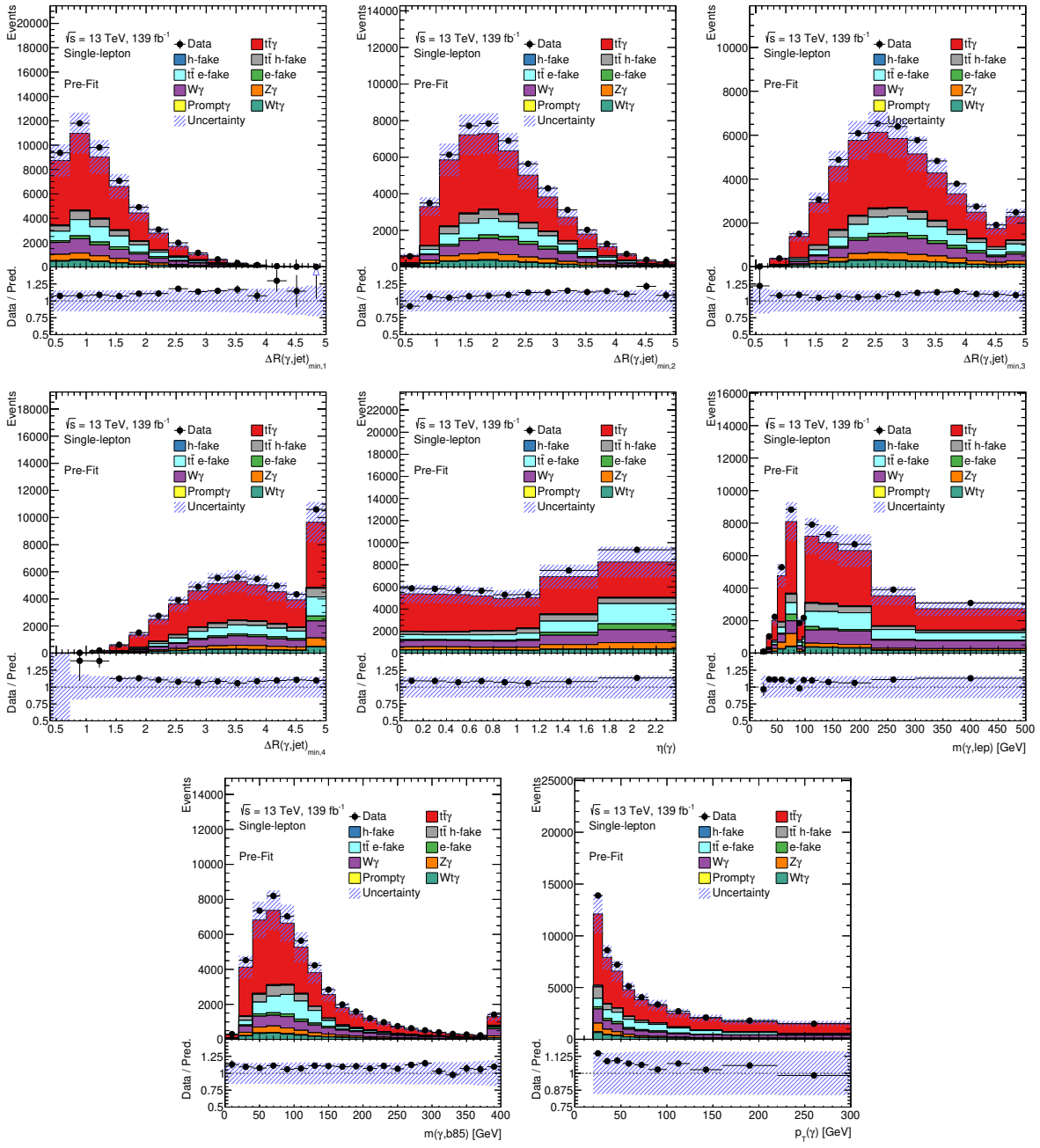

Figure C.4.: Kinematic distributions of the used input variables.

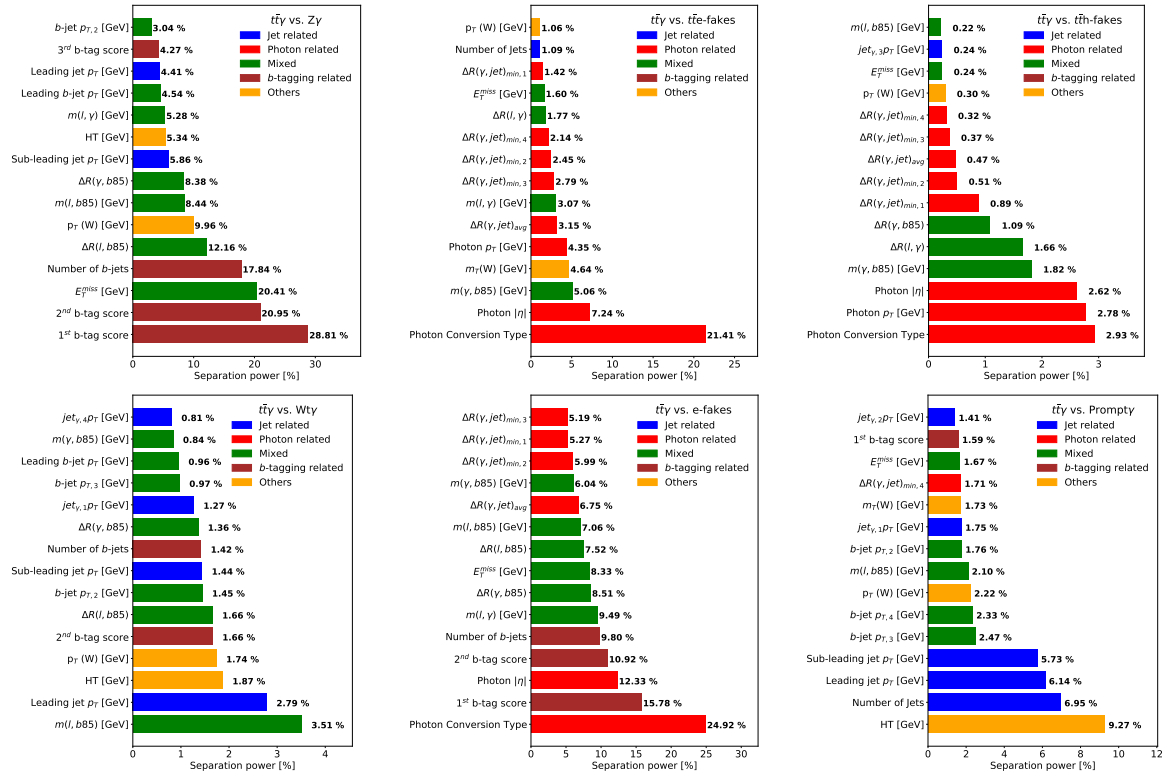

Figure C.5.: Additional ranked separation power plots of the input variables based on the study of the $t\bar{t}\gamma$ topology and the $Z\gamma$, $t\bar{t}$ e-fake, $t\bar{t}$ h-fake, $Wt\gamma$, e-fake and Promptγ event topology

C. Additional figures

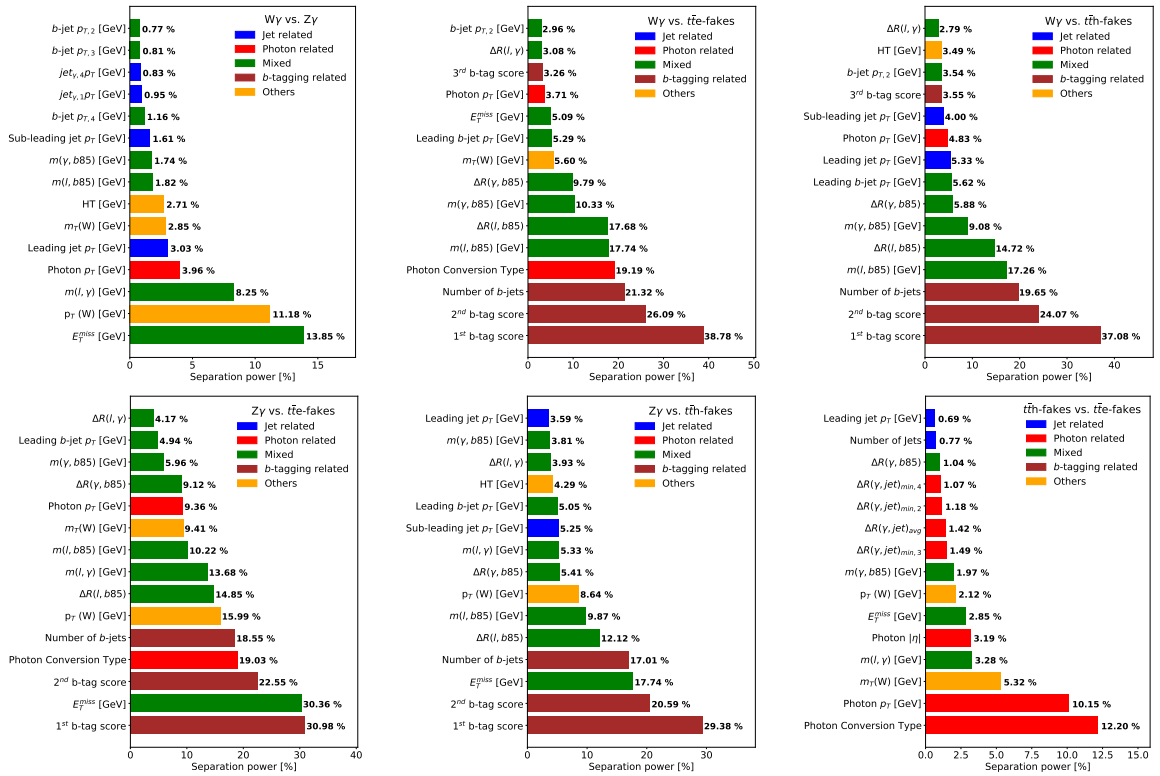

Figure C.6.: Additional ranked separation power plots.

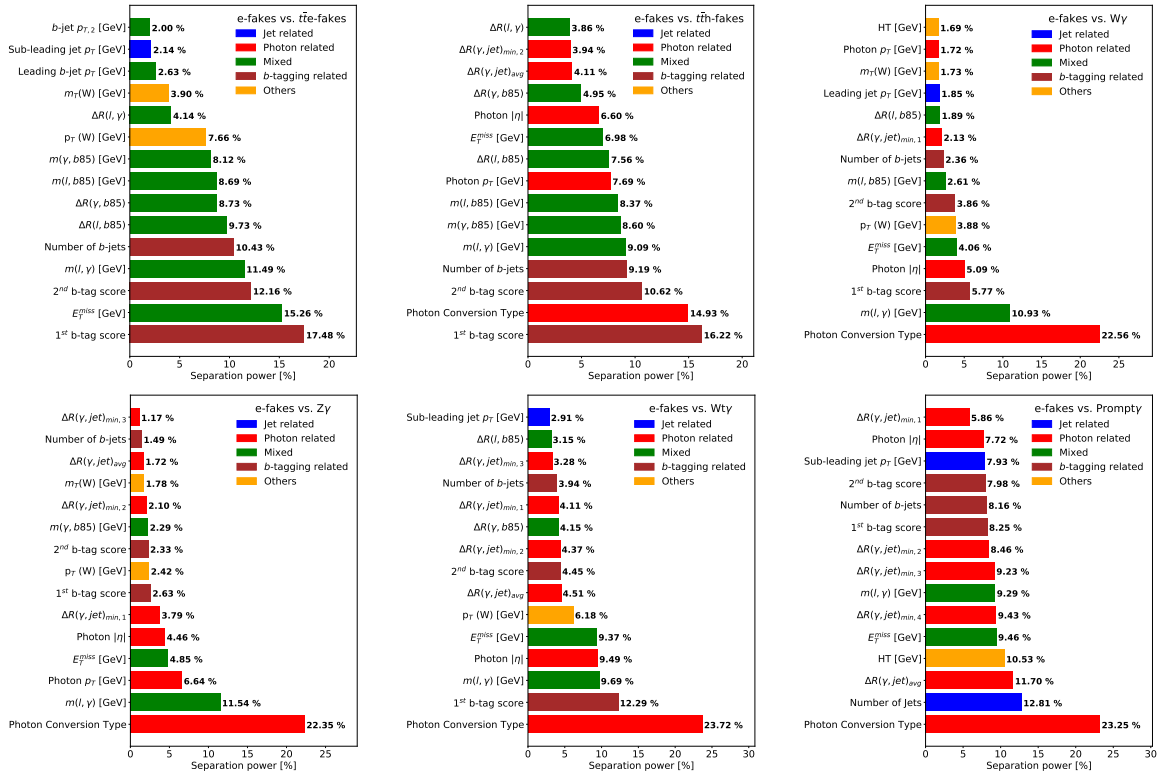

Figure C.7.: Additional ranked separation power plots.

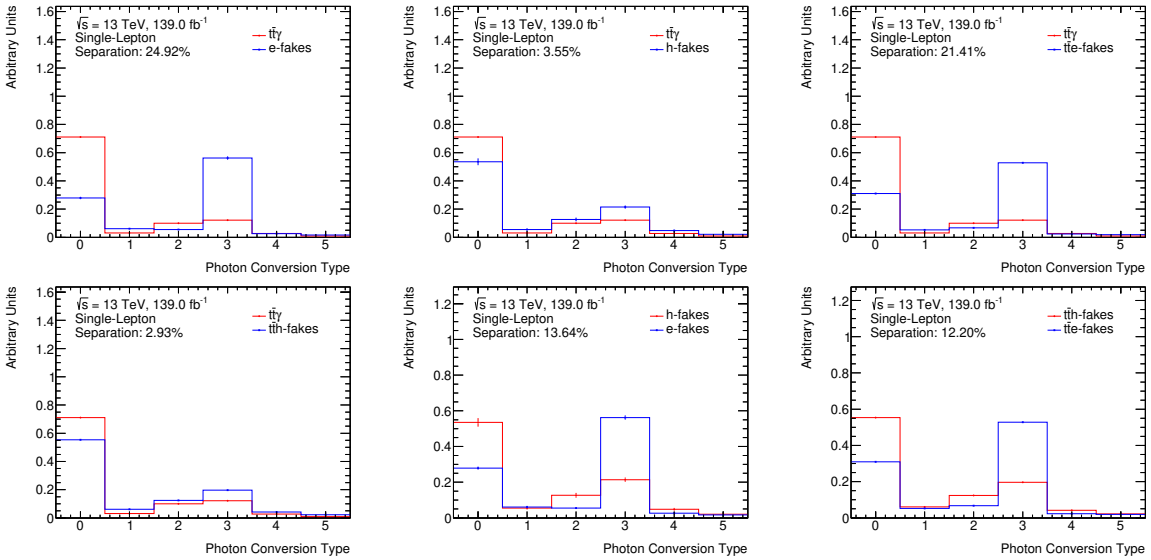

Figure C.8.: Kinematic distributions of the photon conversion type depicting the separation of $t\bar{t}\gamma$ events from h-fake, e-fake, $t\bar{t}$ h-fake and $t\bar{t}$ e-fake events. Additionally, separation plots comparing $t\bar{t}$ h-fakes and $t\bar{t}$ e-fakes as well as e-fakes and h-fakes are also shown for reference.

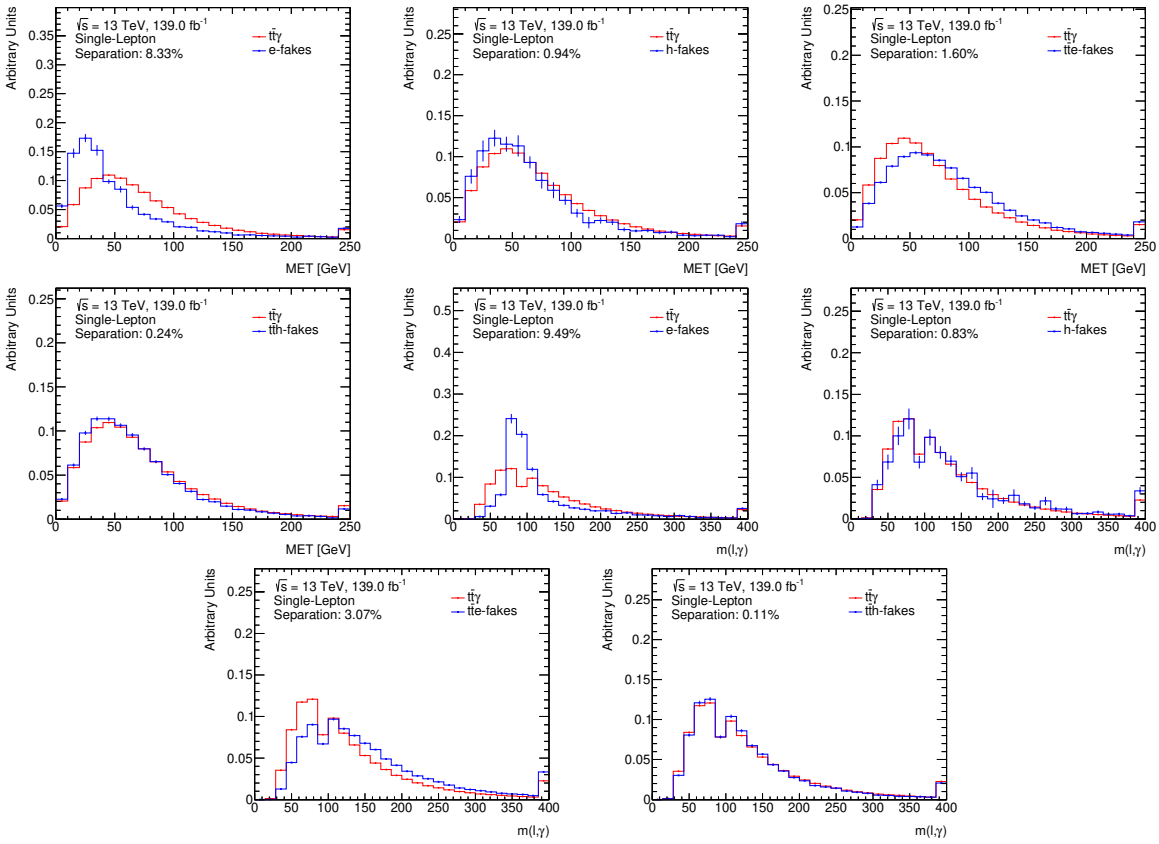

Figure C.9.: Kinematic distributions of $E_{\mathrm{T}}^{\mathrm{miss}}$ and $m(l,\gamma)$, depicting the separation of $t\bar{t}\gamma$ events from h-fake, e-fake, $t\bar{t}$ h-fake and $t\bar{t}$ e-fake events.

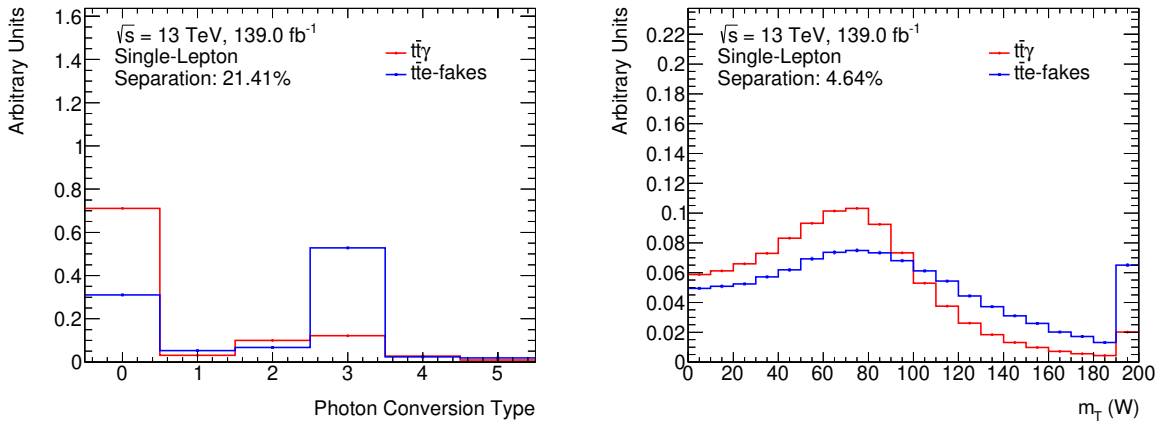

Figure C.10.: Separation plots for the photon conversion type (left) and the transverse mass of the $W^{\pm}$ boson, $m_{\mathrm{T}}(W)$, right. The separation power is 21.41 % and 4.64 % respectively.

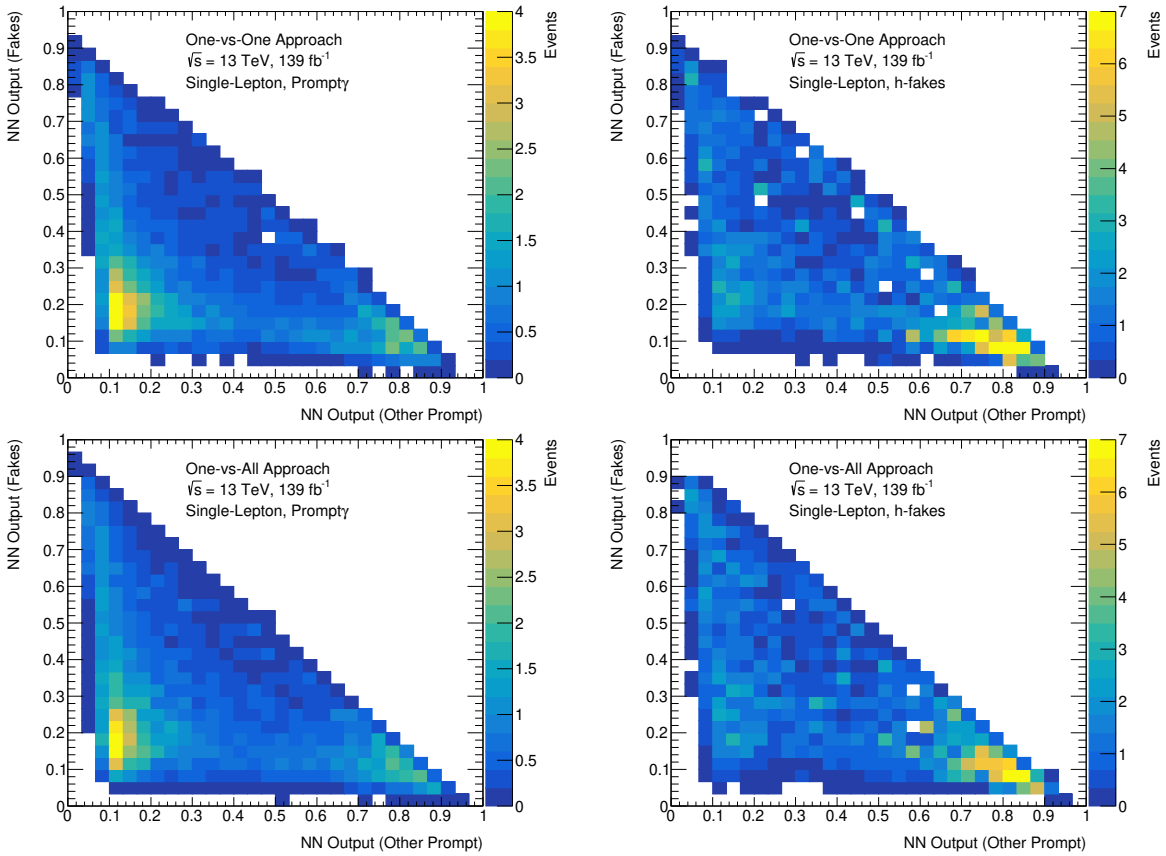

Figure C.11.: Two-dimensional separation plots of the multi-class classifier using the one-vs-all approach (bottom row) and the one-vs-one approach (top row).

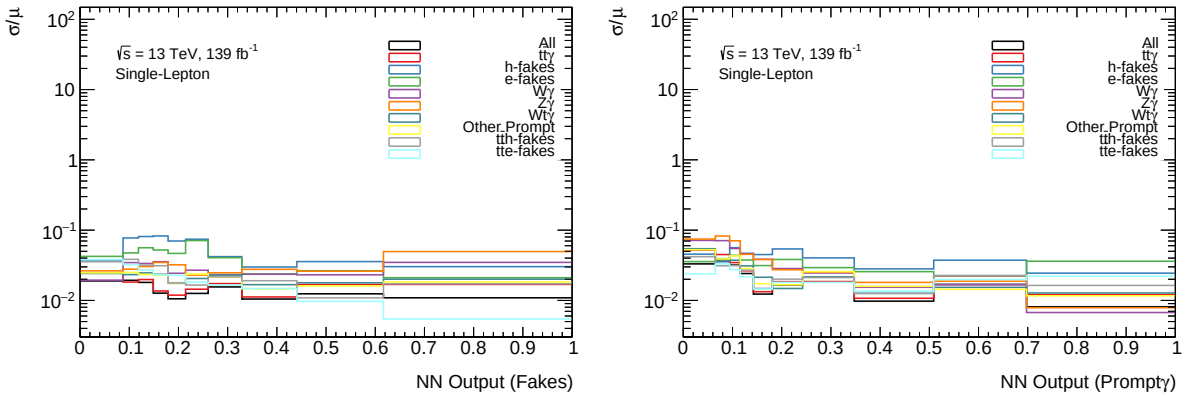

Figure C.12.: Measured relative bin-by-bin uncertainties for the Fakes (left) and Promptγ (right) output classes using the one-vs-one approach. The relative uncertainties are calculated separately for all sub processes.

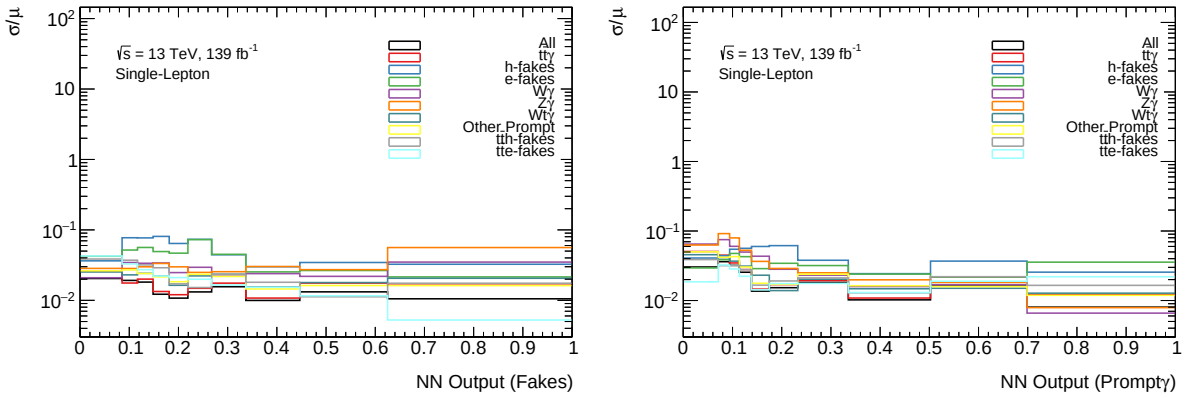

Figure C.13.: Measured relative bin-by-bin uncertainties for the Fakes (left) and Promptγ (right) output classes using the one-vs-all approach. The relative uncertainties are calculated separately for all sub processes.

Erklärung nach §17(9) der Prüfungsordnung für den Bachelor-Studiengang Physik und den Master-Studiengang Physik an der Universität Göttingen:

Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe.
Darüberhinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, im Rahmen einer nichtbestandenen Prüfung an dieser oder einer anderen Hochschule eingereicht wurde.

Göttingen, den January 24, 2020

(Steffen Korn)