การประยุกต์แอดเวอร์แซเรียลเน็ตเวิร์กในการค้นหาการสร้างท็อปควาร์กสี่ตัวในเครื่องตรวจจับอนุภาคซีเอ็มเอส

นายวิชญนันท์ วชิรภูษิตานันท์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตร์มหาบัณฑิต

สาขาวิชาฟิสิกส์ ภาควิชาฟิสิกส์

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2561

APPLICATION OF ADVERSARIAL NETWORKS IN SEARCH FOR FOUR TOP QUARK PRODUCTION IN CMS

Mr. Vichayanun Wachirapusitanand

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Physics

Department of Physics

Faculty of Science

Chulalongkorn University

Academic Year 2018

Thesis Title	APPLICATION OF ADVERSARIAL NETWORKS IN SEARCH FOR FOUR TOP QUARK PRODUCTION IN CMS
By	Mr. Vichayanun Wachirapusitanand
Field of Study	Physics
Thesis Advisor	Norraphat Srimanobhas, Ph.D.
Thesis Co-advisor	Professor Freya Blekman, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

..... Dean of the Faculty of Science
(Professor Polkrit Sangvanich, Ph.D.)

THESIS COMMITTEE

..... Chairman
(Assistant Professor Boonchoat Paosawatyanyong, Ph.D.)

..... Thesis Advisor
(Norraphat Srimanobhas, Ph.D.)

..... Thesis Co-advisor
(Professor Freya Blekman, Ph.D.)

..... Examiner
(Pawin Ittisamai, Ph.D.)

..... External Examiner
(Assistant Professor Tanasanee Phienthrakul, Ph.D.)

วิชญนันท์ วชิรภูษิตานันท์: การประยุกต์แอดเวอร์แซเรียลเน็ตเวิร์กในการค้นหาการสร้างท็อปควาร์กสี่ตัวในเครื่องตรวจจับอนุภาคซีเอ็มเอส. (APPLICATION OF ADVERSARIAL NETWORKS IN SEARCH FOR FOUR TOP QUARK PRODUCTION IN CMS) อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก : อาจารย์ ดร. นรพัทธ์ ศรีมโนภาษ, อาจารย์ที่ปรึกษาร่วม : ศาสตราจารย์ ดร. เฟรยา เบลคแมน 56 หน้า.

ปัญหาหนึ่งในการวิเคราะห์ข้อมูลสำหรับฟิสิกส์พลังงานสูงคือความคลาดเคลื่อนในการวัด ทั้งความคลาดเคลื่อนเชิงระบบและเชิงสถิติ ถึงแม้ว่าจะใช้ระบบโครงข่ายประสาทเทียมอันซับซ้อนในการวัดของฟิสิกส์พลังงานสูง ผลที่ได้จากการวัดอาจยังได้รับผลกระทบจากความคลาดเคลื่อนทั้งสองประเภท ความคลาดเคลื่อนเชิงระบบส่วนใหญ่มีที่มาจากข้อมูลต่าง ๆ เช่นการประมาณเชิงทฤษฎี ซึ่งสามารถระบุค่าของมันได้ ในขณะนี้มีนักวิจัยกลุ่มหนึ่งที่เสนอการลดความคลาดเคลื่อนเชิงระบบนี้โดยอาศัยโครงข่ายประสาทเทียมชนิดหนึ่งเรียกว่าแอดเวอร์แซเรียลเน็ตเวิร์ก (ANN) โดยสามารถทำให้ตัวคัดแยกเหตุการณ์การชนไม่ได้รับผลกระทบจากความคลาดเคลื่อนเหล่านี้ แต่แนวทางนี้ยังไม่ได้มีการใช้งานจริงในการวิเคราะห์ข้อมูลจากเครื่องชนอนุภาคแฮดรอนขนาดใหญ่ (LHC)

งานชิ้นนี้ได้ศึกษา ANN กับการประยุกต์ใช้เพื่อการค้นหาการสร้างท็อปควาร์กสี่ตัว ซึ่งเป็นอันตรกิริยาที่เกิดได้ยากในเครื่อง LHC และยังเป็นอันตรกิริยาที่สามารถยืนยันหรือปฏิเสธความถูกต้องของแบบจำลองอนุภาคมูลฐาน ในบางกรณี การค้นหาการสร้างท็อปควาร์กสี่ตัวจะได้รับผลกระทบจากความคลาดเคลื่อนเชิงระบบในปริมาณมาก ในงานนี้ ผู้เขียนได้สร้างตัวคัดแยกเหตุการณ์การชนจากโครงข่ายประสาทเทียมทั่วไป และแอดเวอร์แซเรียลเน็ตเวิร์ก พร้อมกับฝึกฝนและคัดเลือกตัวคัดแยกเหตุการณ์โดยอาศัยการปรับไฮเปอร์พารามิเตอร์ และใช้ตัวคัดแยกเหตุการณ์ทั้งสองประเภทเพื่อคำนวณค่าคาดหวังค่าขอบบนของพื้นที่ตัดขวาง และนัยสำคัญทางสถิติของการค้นหาการสร้างท็อปควาร์กสี่ตัว โดยอาศัยข้อมูลที่ได้จากการจำลองเหตุการณ์การชนภายในเครื่องตรวจจับอนุภาคซีเอ็มเอสที่ตั้งอยู่ในเครื่อง LHC พร้อมกับนำผลที่ได้จากตัวคัดแยกเหตุการณ์ทั้งสองประเภทไปเปรียบเทียบกับผลที่ได้จากการวิเคราะห์ก่อนหน้านี้ วิทยานิพนธ์ฉบับนี้จะกล่าวถึงการปรับปรุงที่ได้และการประยุกต์ใช้สำหรับการค้นหาอันตรกิริยาที่เกิดได้ยากในอนาคต

ภาควิชา .....ฟิสิกส์..... ลายมือชื่อนิสิต .....

สาขาวิชา.....ฟิสิกส์..... ลายมือชื่ออ.ที่ปรึกษาหลัก .....

ปีการศึกษา .....2561..... ลายมือชื่ออ.ที่ปรึกษาร่วม .....

607 19935 23: MAJOR PHYSICS
KEYWORDS: ADVERSARIAL NEURAL NETWORK, STANDARD MODEL, FOUR TOP QUARK PRODUCTION, PARTICLE PHYSICS, LHC, COMPACT MUON SOLENOID

VICHAYANUN WACHIRAPUSITANAND : APPLICATION OF ADVERSARIAL NETWORKS IN SEARCH FOR FOUR TOP QUARK PRODUCTION IN CMS. ADVISOR : NORRAPHAT SRIMANOBHAS, PH.D., THESIS COADVISOR : PROFESSOR FREYA BLEKMAN, PH.D., 56 pp.

One burden of high energy physics data analysis is uncertainty within the measurement, both systematically and statistically. Even with sophisticated neural network techniques that are used to assist in high energy physics measurements, the resulting measurement may suffer from both types of uncertainties. Fortunately, most types of systematic uncertainties are based on knowledge from information such as theoretical assumptions, for which the range and behaviour are known. It has been proposed to mitigate such systematic uncertainties by using a new type of neural network called adversarial neural network (ANN) that would make the discriminator less sensitive to these uncertainties, but this has not yet been demonstrated in a real LHC analysis.

This work investigates ANNs using as a benchmark the search for the production of four top quarks, an extremely rare physics process at the LHC and one of the important processes that can prove or disprove the Standard Model. The search for four top quarks in some cases is sensitive to large systematic uncertainties. Discriminators based on traditional and adversarial neural networks are trained and chosen via hyperparameter adjustment. The expected cross section upper limit and expected significance for four top quark production is calculated using traditional neural networks and adversarial neural networks based on simulated proton-proton collisions within the Compact Muon Solenoid detector within Large Hadron Collider, and are compared to existing results. The improvement and further considerations to the search for rare processes at the LHC will be discussed.

Department : .....Physics.....	Student's Signature .....
Field of Study : .....Physics.....	Advisor's Signature .....
Academic Year .....2018.....	Co-advisor's signature .....

Acknowledgements

It is impossible to believe that an unpredictable, mood-swinging madman like me, who has been exposed to the wonderful world of Physics for almost ten years ago, will finally obtain his MSc degree. Impossible things like this would remain impossible if I have not met so many wonderful people along my ten-year journey with Physics, and I would like to thank you all for this.

First of all, I would like to thank Dr Norraphat Srimanobhas and Prof Dr Freya Blekman, who, as my thesis advisors, mentored and advised me throughout the whole year working on this thesis. Without both of you, I would be stranded alone like a shipwrecked man in the Pacific ocean. I would also like to thank my BSc academic advisor, Assist Prof Dr Narumon Suwonjandee, and my BSc thesis advisor, Assist Prof Dr Burin Asavapibhop, who mentored me during my undergraduate years, and still advises me througout bureaucracy stuffs I am required to do during my years here. This work has also been financially supported in part by the Thailand Research Fund under contract no. MRG5980195, the Chulalongkorn Academic into Its 2nd Century Project Advancement Project (Thailand), and het Fonds Wetenschappelijk Onderzoek ODYSSEUSII programme (Belgium).

I would like to give my appreciation to Dr Martijn Mulders, who, as my advisor during my days as a CERN Summer Student, introduced me to the real day-to-day work at CERN, inspiring me to work on this career path further, and making me confident that I am on the right career choice. I would also like to thank Dr Emanuele Simili, who introduced me to the field of High Energy Physics while I was just a stupid undergraduate freshman.

I would like to thank my family, who, even to this day, still supports me in this academic path unconditionally. I am sure they don't know or understand much of my work, but they still keep on believing in me doing what I do best.

Last but not least, I would like to give out my deepest love and appreciation to my teachers and friends at Mahidol Wittayanusorn School, and also the school itself that changed my life forever. I have heard that the school has been changed in recent years, but I, as an alumni, still wish the school to keep on inspiring students and making a perfect community for studying Science. Please keep on changing the lives of students with brilliant minds for the better, just like what you did to me.

Contents

List of Tables

List of Figures

CHAPTER I

INTRODUCTION, BACKGROUND THEORY, AND CHALLENGE

The latest discovery to the standard model (SM), the Higgs Boson, is discovered in 2012 [1] and makes it the most comprehensive model of elementary particles to date. Still, the model is currently under tests from many more and more complex particle processes and phenomena. SM is expected to "break", or unable to precisely predict a phenomenon, at some point of precision, as it has happened to Newtonian mechanics before. Newtonian mechanics has a predictive power at a certain scale, and predicting physical phenomena beyond that scale requires all-new theories, such as General Relativity which can predict black holes and Quantum mechanics which can explain such spooky phenomena in subatomic particles. As such, if SM fails to describe one of the critical phenomena, several theories by theoretical physicists might someday predict it, but we have to prove that SM fails to do it in the first place.

The first four chapters in this thesis are mainly literature studies in terms of High Energy Physics and Computer Science, particularly in the field of Machine Learning, while new contributions presented in this thesis starts from Chapter 5 onwards. In this chapter, one of the tests on SM, four top quark production, is presented, along with a challenge in search of the production. Following theoretical background are the specifications of the Compact Muon Solenoid (CMS) detector in Chapter 2. Chapter 3 presents basic terminology in the field of machine learning, while Chapter 4 summarises previous analyses for a search of four top quark production in the past using the CMS detector. An analysis using a traditional neural network is presented in Chapter 5. Adversarial neural networks, as a recent advancement in the field of machine learning (ML), is also introduced in this chapter, although the thorough description of this type of neural network will be presented in Chapter 6. To summarise the results of this thesis, Chapter 7 outlines the sensitivity results from both traditional and adversarial neural networks, as well as a future outlook to the application of the adversarial neural networks.

1.1 The top quark

Discovered in 1995 [2], the top quark is the heaviest quark currently discovered in SM. As of current measurements, its mass is reported by [3] as 173.0 ± 0.4 GeV, heavier than the mass of Higgs boson of 125.18 ± 0.16 GeV. Even after 20 years of discovery, the top

quark remains an important study in High Energy Physics, since its properties raise many questions to SM. According to SM, top quarks may be produced through various production mechanisms, given that there is enough energy for the top quark to be produced in the first place. A top quark instantly decays (96% of the time) to a W boson and a bottom quark, and the W boson would further decay into either a pair of quark-antiquark or a lepton and a neutrino of the corresponding flavour, as shown in Figure 1. This high branching ratio of $t \to Wb$ decay mode is explained by a specific element V_{tb} as one of the diagonal terms of Cabibbo-Kobayashi-Maskawa (CKM) matrix [4], which gives a particularly high probability of a top quark decaying into a bottom quark (and a W boson due to charge conservation). The current value of V_{tb}, derived from a global fit to all available measurements and SM constraints, is $V_{tb} = 0.999\,105 \pm 0.000\,032$ [3].

Quantum Chromodynamics (QCD), as the SU(3) component of SM ($\mathrm{SU}(3) \times \mathrm{SU}(2) \times \mathrm{U}(1)$), dictates that a single quark contains a quantum number called colour charge, either "red", "green", or "blue".[1] It also gives rise to gluons (or gluon fields) as strong force carriers and can alter the colour of quarks once interacted with them. A particle carrying a colour charge cannot be isolated and must be grouped up with other particles carrying colour charge to be colourless. For example, a quark with red colour charge must pair with an antiquark with antired colour charge to form mesons. A quark with red colour charge may also group up with two other quarks with blue and green colour charges respectively to form baryons. The colour charge is different from an electrical charge, since particles may possess electrical charges, such as electrons, and can be isolated freely. Due to this effect, a single quark generated from particle collisions must be grouped up with other quarks, originated from a vacuum. This process is called hadronization, and a shower of particles generated from a quark via this process is called a particle jet. The only exception to this rule from QCD is a top quark, as it decays before hadronization can occur due to its short lifetime [5].

1.2 Single top quark production in t-channel and tW-channel

A single top quark may be created via a scattering of a light quark and a bottom quark, or a light quark and a gluon, exchanging a W boson in the process and resulting in a light quark of different flavour and a top quark, as shown in Figure 2. This process is called t-channel process and is a dominant production mechanism in a single top quark production. A less dominant process is called a tW-channel process, where a gluon and a bottom quark scattered into a top quark and a W boson, as shown in Figure 3.

[1] Note that quarks do not actually have these colours.

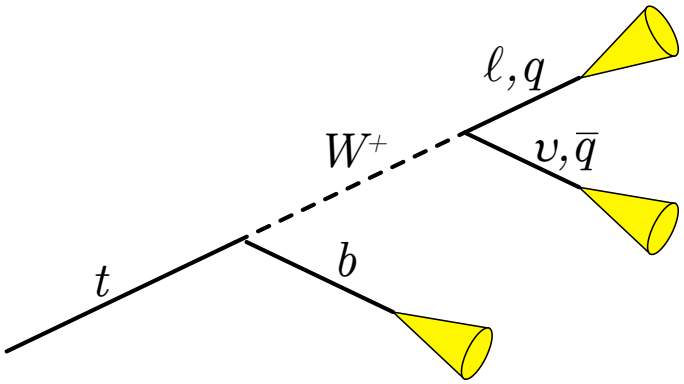


Figure 1: Decay diagram of top quark.

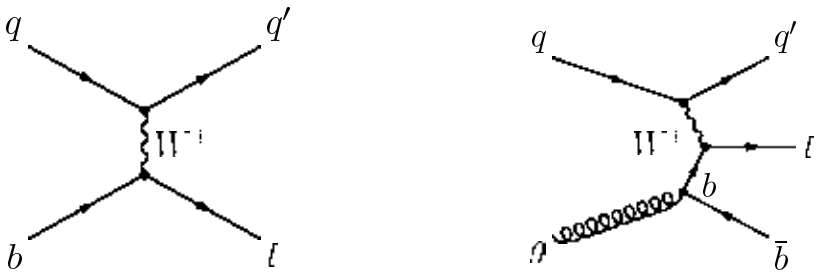


Figure 2: Tree diagrams for single top quark, t-channel process.

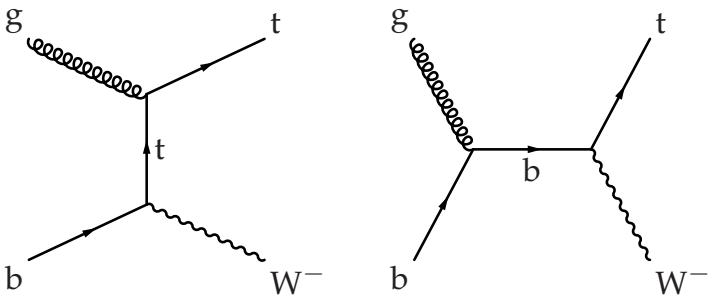


Figure 3: Tree diagrams for single top quark, tW-channel process.

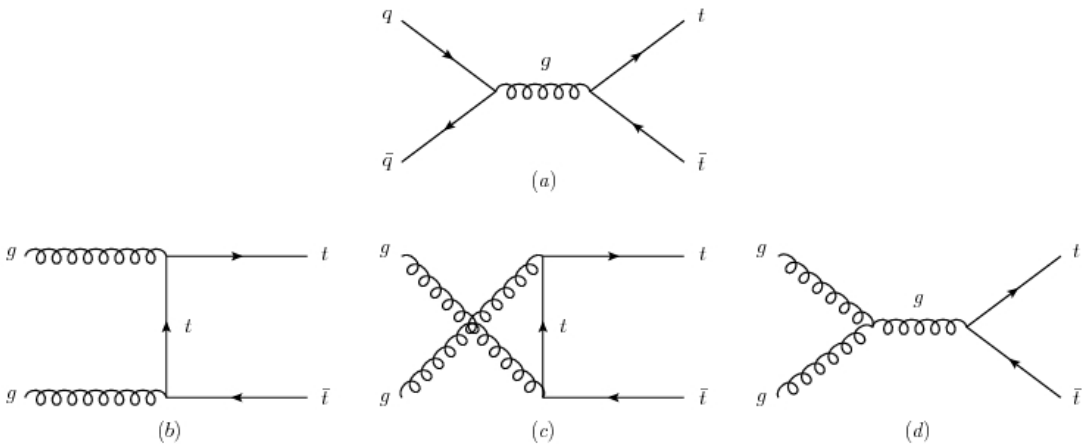


Figure 4: Tree diagrams for top-antitop quark production.

The cross sections for t-channel and tW-channel processes of single top production, measured by CMS detector with $\sqrt{s} = 13$ TeV, are 238 ± 13 (stat) ± 29 (syst) pb [6] and 63.1 ± 1.8 (stat) ± 6.4 (syst) pb [7] respectively. These measurements are reported to have an agreement with predictions made by SM.

1.3 Top-antitop quark production

A top-antitop quark pair may be produced via a collision of two quarks or an interaction between two gluons, as shown in Figure 4 [8]. The cross section for this production has been measured numerous times by the CMS detector, and has been shown to be consistent with predictions made by SM in the centre-of-mass energy of 7 TeV (such as in [9]), and 13 TeV [10], to name a few. With the measurement of the cross section of this process, the strong coupling strength α_s and the pole mass of the top quark can be extracted. [11]

1.4 Four top quark production

Within a single proton-proton collision, two pairs of top-antitop quarks may be created due to an interaction between two gluons as shown in Figure 5. Right after the production, the top quark will instantly decay into a bottom quark and either a pair of quark-antiquark or a lepton and a neutrino.

In a four top quark production, each of the four top quarks has the freedom to decay either hadronically ($t \to Wb \to q\bar{q}b$ at 66%) or leptonically ($t \to Wb \to \ell\nu_\ell b$ at 33%) via the W boson decay. For a single lepton search, we are interested in finding such

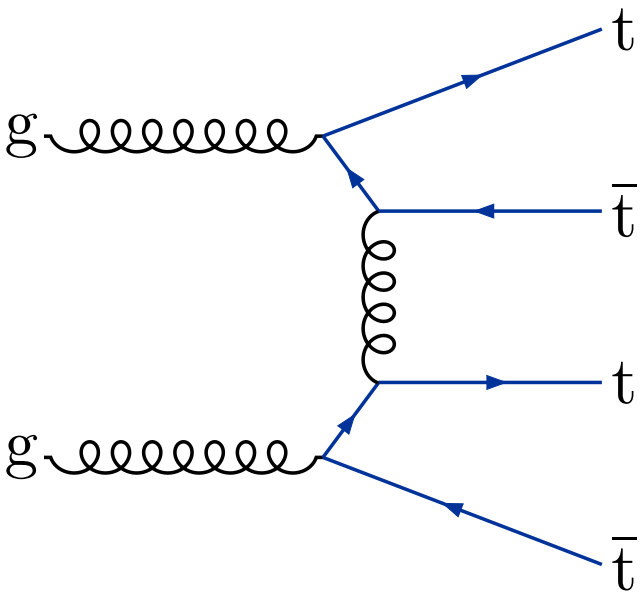


Figure 5: Tree diagram for four top quark production.

production giving a single lepton from one top quark decaying leptonically. The remaining three top quarks must decay hadronically. Hence, we will require the collision event to contain four bottom quarks (directly from top quark decay), one lepton and three pairs of quark-antiquark.

The cross section of this process, according to SM prediction at next-to-leading order (NLO) with 13 TeV centre-of-mass energy ($\sqrt{s}$), is 9.2 fb. By comparison, top-antitop production as the most dominant background has its cross section, according to SM NLO prediction at $\sqrt{s} = 13$ TeV, as high as 832 pb [10], almost 90 000 times greater than four top quark production. Previous literature [12] has determined the expected upper limit for four top quark production at $\sqrt{s} = 13$ TeV, up to 95% confidence level, as 20^{+10}_{-6} fb. To this date, the data gathered from the CMS detector is not adequate enough to determine, at enough significance, how small the cross section for four top quark production is.

Determining the four top quark production cross section is considered as a test for SM, as well as theories colloquially called Beyond Standard Model (BSM), which are possible extensions to SM itself. These theories propose "fixes" to SM should it fail to predict certain phenomena. As for the case of four top quark production, many BSM models propose that its cross section may be altered by top quark compositeness (or the top quark may contain some other particles), dark matter, supersymmetry, extra dimensions, or much more [13].

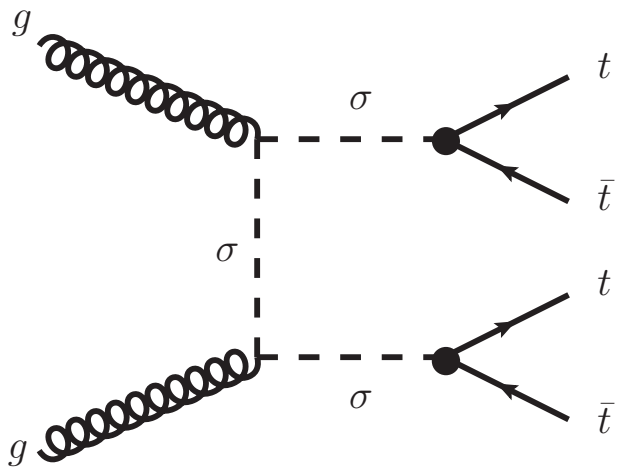


Figure 6: Feynman diagram for a possible four top quark production via supersymmetry particle. σ represents sgluons, a superpartner of gluons. [14]

1.5 Challenge for four top quark production search

As stated in Section 1.4, the dominant background for four top quark production, top-antitop production, is approximately 90 000 times larger than four top quark production. This means we have to go through a huge collection of data collected by the CMS detector containing millions of collision events just to find a handful of events that may contain four top quark production. To make matters worse, we can only observe the final state of the production, or the final decay products of the collision process, which may originate from either four top quark production or top-antitop production.

The very first discriminator used in Physics analyses is called "cut based" discriminator, which is merely a set of rules based on variables in each event. If an event contains the variables that pass all rules, the event would be classified by the cut-based discriminator as a signal-like event. On the other hand, if an event fails on *any* rule, the event would not be classified as a signal-like event. The cut-based discriminator is built by human, not computers, and in some cases, the discriminator can achieve high purity of signal events passing through the discriminator. In other cases, where high signal purity cannot be achieved or there is overwhelmingly more background, the discriminator may also leave many signal events behind to achieve high purity or allow many background events to pass if it is designed to include as many signal events as possible. To alleviate this, several machine learning (ML) techniques, such as boosted decision trees and neural networks, can better distinguish the events by analysing and assessing the events in such a way that signal events are included and background events are excluded as much as possible, based on the

training dataset used. One of the advantages of these ML techniques, and multivariate analyses (MVAs) in general is its multidimensionality: a cut-based discriminator rejects an event if it does not have a value of a variable in a certain range, whereas ML techniques tend to look at several variables at the same time to make a decision.

Boosted decision trees (BDT) [15], one of many techniques in ML, have been used in an attempt to distinguish between four top quark production and top-antitop production. By using the distribution of an output of BDT discriminator (that is trained properly with simulated collision events), we can obtain the cross section expected limit of four top quark production. However, even with traditional ML techniques, the calculated expected limit is still subject to systematic uncertainties, where most of them are based on well-known theoretical uncertainties and estimates, and can be modelled during particle collision simulations. In fact, traditional ML techniques are designed to classify events based on patterns within the data into classes such as a signal or background classes, but they are not deliberately designed to lessen the impacts from uncertainties, even with systematic uncertainties that we know their behaviours based on theoretical knowledge.

1.6 Use of adversarial neural networks

In 2017, Louppe et al. have shown [16] that it is possible to use another type of neural network, called a pivoting adversarial neural network, to adjust the discriminator output to be less susceptible to uncertainties. In essence, the adversary network will try to assess features of an input that makes the discriminator network to give certain outputs. While the goal of the discriminator is to classify the inputs, the adversary's goal is to find out why the discriminator makes such predictions. Normally adversarial neural networks are used to generate fake input into a discriminator network so that the discriminator cannot tell the difference between real input and fake input. With the normal use case, the adversary network is mostly used to generate images that humans can recognise. A light-hearted example of this use case is generating Japanese anime-styled girl portraits [17].

In Louppe's work, the adversarial neural network is used in an entirely different direction. Instead of generating fake inputs to the discriminator, the adversary is instead designed to determine the uncertainty of the inputs as best as possible, by relying on the output of the discriminator alone. The discriminator is then trained to fool the adversary by giving the output that is indifferent under varying uncertainties. Shimmin et al. [18] have further developed the adversarial network training to train the discriminator network to classify particle jets without the effect of the jet's mass. Still, **this technique has been**

tested with "toy examples", or simulated data only, and has never been used in real-world cases such as a complete analysis at the Large Hadron Collider (LHC). This thesis is the first ever attempt on the application of adversarial neural networks in such analyses. With a promising use case of the adversarial network, it is possible that we can use adversarial neural networks to train a discriminator to reduce the effects of systematic uncertainties on search sensitivity, and we will investigate this application in this thesis.

CHAPTER II

THE CMS DETECTOR

The Compact Muon Solenoid (CMS) detector is one of the four main particle detectors situated on the Large Hadron Collider (LHC) along with A Toroidal LHC Apparatus (ATLAS) experiment, Large Hadron Collider beauty (LHCb) experiment, and A Large Ion Collider Experiment (ALICE). CMS collaboration, together with ATLAS collaboration, discovered the Higgs boson in 2012 based on the data collected from proton-proton collisions detected inside both of the detectors [1], confirming the almost-fifty-year-old theory at the time proposed by several theorists.

The CMS detector is designed to be a general purpose particle detector, which mainly detects photons, hadrons (both neutral and charged), electrons, and muons. The detector is not designed to detect neutrinos, and tau leptons usually decay in almost an instant. Hence, the detector is unable to detect tau leptons directly and all flavours of neutrinos. Decay products of tau leptons, however, are still detectable. The ATLAS detector and the CMS detector use different technology in its components, and both serve as a general purpose particle detector located on the LHC.

2.1 Proton beams from LHC

The LHC is a circular particle accelerator that accelerates two beams of protons inside and collides them at four main detector stations. It became operational on 19 September 2008, with an initial centre-of-mass energy of 7 TeV. Since then the accelerator has been upgraded along with colliding proton beams for Physics purposes. As of this writing, the centre-of-mass energy LHC can achieve is 13 TeV, and its peak instantaneous luminosity of 10^{34} $\text{fb}^{-1}\text{s}^{-1}$ has been achieved. These two upgrades to the LHC causes more elusive particle productions more detectable in particle detectors, including CMS.

2.2 Components of CMS detector

The CMS detector comprises four main components:

- **Silicon tracker**, the innermost component, detects tracks of charged particles that occur right after the collision. It consists of several layers of silicon pixel detectors and silicon microstrip detectors, forming the shape of a cylinder around the interaction point where proton beams collide. Three layers of pixel detectors and ten layers of

Figure 7: Aerial view of Large Hadron Collider. [19]

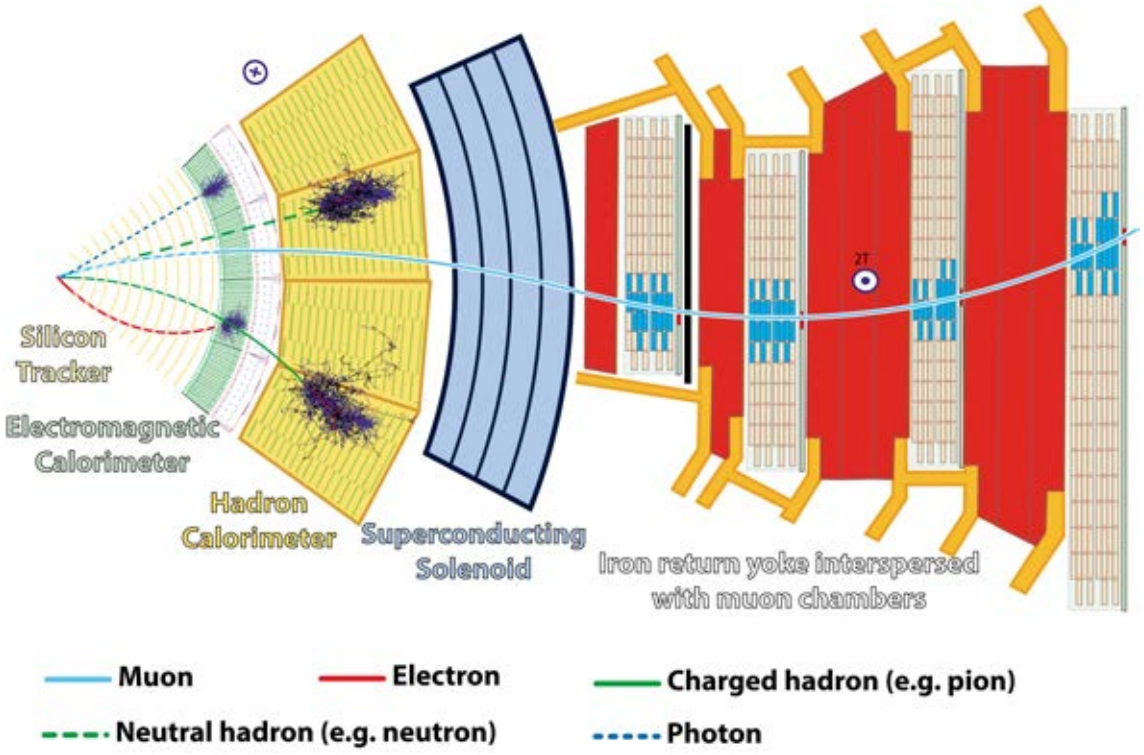


Figure 8: Schematic diagram of CMS detector with its components. [20]

silicon strip trackers form the tracker components in the shape of a cylinder with no ends. To detect forward particles at the end of the cylinder, two layers of pixel detectors and twelve layers of silicon strip trackers also cover the ends of the detector. Charged particles will leave a trace of energy in an array of small pixels and silicon strip detectors.

- **Electromagnetic calorimeter (ECAL)** detects photons and electrons. The calorimeter is made of lead tungstate ($PbWO_4$) crystals covering the pseudorapidity range of $|\eta| < 3$ around the interaction region. Both particles leave high amounts of energy and stop moving further in the detector.
- **Hadronic calorimeter (HCAL)** detects hadrons, both charged (i.e. protons) and neutral (i.e. neutrons). It is based on scintillators, and its coverage is the same as ECAL, $|\eta| < 3$. Neutral hadrons do not leave a trace in the silicon tracker and ECAL but are detected in this layer.
- **Muon chamber** exclusively detects muons. It is made of four layers of drift tubes, cathode strip chambers, and resistive plate chambers. Muons passing through this part of the detector leave energy traces in an array of pixels, similar to other parts of the detector.

The detector also contains a superconducting solenoid between HCAL and muon chamber, providing a magnetic field inside the detector. With the magnetic field generated from the solenoid, charged particles will have their trajectories curved depending on their momentum and charge. This can help to identify the charge of individual particles occurring after the collision.

To handle a large number of proton-proton collisions the LHC can deliver, the detector must be able to reject uninteresting events to save data storage. This is achieved by the use of several levels of trigger systems. Trigger systems are a combination of hardware and software, designed to accept or reject collision events as fast as possible. The hardware is also designed to resist the radiation from particle collisions over time. With the use of triggers, collision events recorded by the detector can be reduced dramatically by a factor of 10^6, allowing only interesting events to be recorded in limited storage.

A custom algorithm named Particle-flow algorithm [21] is also designed to help reconstruct particle trajectories from raw data in each component of the detector. With this algorithm, instead of dealing with raw data of energy residues in each detector pixels,

we can obtain trajectory and energy information of particles created after the collision, such as photons, electrons, muons, and particle jets. A complete description of the detector and its components can be found in [22].

CHAPTER III

BASIC TERMINOLOGY IN NEURAL NETWORK TRAINING

A neural network is a type of machine learning (ML) technique that is designed to find underlying patterns within the data. As the technique comes straight from computer sciences, there is a specific set of vocabulary that needs to be understood to work with it. This chapter will present the terminology as well as general procedures of applying neural networks to many use cases. The details of the neural network designed in this thesis will be covered in Chapter 5.

3.1 What is machine learning?

In essence, machine learning (ML) is a collection of algorithms that can predict the answer based on the inference of input data alone, without any instructions to answer. These algorithms can read the data, predict answers, and adjust itself to give more and more correct answers. Many algorithms have been devised, and the most notable algorithms are boosted decision trees (BDT) and neural networks (NN). ML algorithms are nowadays being used for a myriad of purposes, such as natural language processing, self-driving cars, and particularly discoveries in High Energy Physics, such as the discovery of Higgs Boson in 2012 by CMS [1]. However, the use of machine learning in High Energy Physics can be dated back as far as the era of Large Electron-Positron (LEP) collider, which was decommissioned in 2000.

3.2 ML classifiers vs regressors

Several ML techniques can be used to either classify input entries into certain categories, such as classifying a set of images into certain types, or predict the value of something, such as predicting house value in certain areas. The predictor in the former use case (that predicts the type, or **class**, of an input entry) is called a **classifier**, which can also be called a **discriminator** in some places. The predictor in the latter use case (that predicts arbitrary values) is called a **regressor**.

In this thesis, we will focus on training a classifier alone, although training and using a regressor is similar. The use of ML techniques in High Energy Physics is usually applied for collision events recorded from particle detectors, so an input for any ML predictor,

including neural networks, will be called an **input event** instead of an input entry in this thesis. Also, we will refer to a classifier as a discriminator, as High Energy Physics research papers tend to refer to this type of predictor as a discriminator.

3.3 Preparing the data

Just as one may not make any assumptions without supporting data, any ML discriminator, not limited to neural networks, must require a set of data (**the dataset**) to be trained on. The dataset must be separated into at least two sets: **training dataset** and **testing dataset**. The training dataset is used to train the discriminator, while the testing dataset is separated to evaluate the performance of the discriminator after the training, such as its classification accuracy. During the training of the discriminator, it is possible that the discriminator might have been trained to "memorise the answer" or trying to memorise the output of each input event. This phenomenon is called **overtraining**, and to prevent this we may reserve another set of data called **validating dataset**, which the discriminator is not to be trained with, to calculate the classification accuracy of the discriminator during the training.

3.4 Boosted decision trees

A single decision tree is a classification tree, similar to Dichotomous key in Biology. To classify an event using the tree, a decision tree starts taking decisions based on one variable at a time until a conclusion is reached. Even though a decision tree can tune itself to give more prediction accuracy, an ensemble of decision trees, sometimes called forests, are usually used, so-called *boosted* decision trees (BDT). During the training of a BDT-based discriminator, each tree in the ensemble is trained by reweighting events, and the final output of the discriminator is a weighted average of all outputs from all trees in the ensemble. [15]

BDT-based discriminators have been used in numerous High Energy Physics analyses, most notably the discovery of Higgs Boson in 2012 [1]. In these analyses, a BDT-based discriminator is trained to discriminate between signal and background events, outputting a single discriminator value which can later be used as a comparison between simulated collision events and recorded collision data.

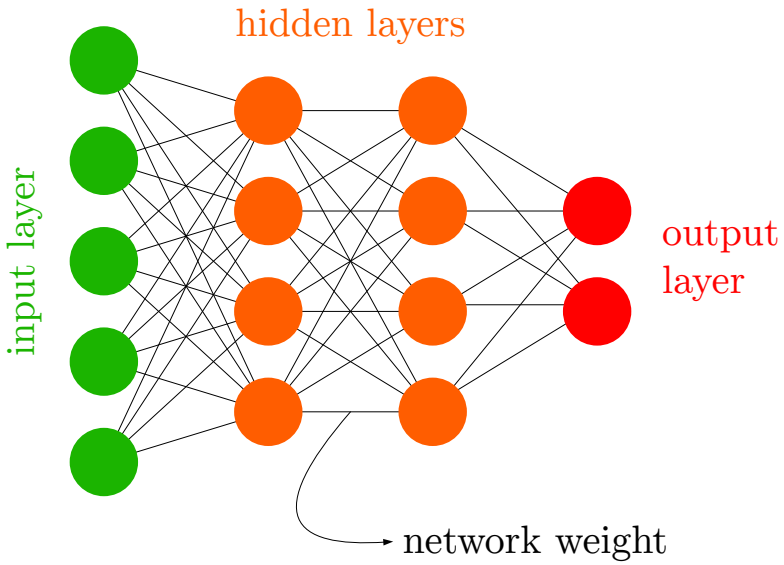


Figure 9: A brief schematic diagram of a neural network.

3.5 Anatomy of a neural network

A neural network (NN) is a mathematical model designed to mimic working neurons of the human brain. The most basic part of a neural network is a small computational unit called a **neuron**. In traditional neural networks, several neurons form a **layer of neurons**, and several stacked layers of neurons form a whole network with connections between every two layers in the stack. Each neuron in a layer will take inputs from the previous layer, multiply *each* input with individual **weights**, and calculate the weighted sum from the inputs. Each neuron may also have a **bias** which is simply a number that adds into the weighted sum. The weighted sum (with bias) is then passed to its designated **activation function**. Examples of activation function are perceptron ($f(x) = 1$ if $x > 0$ or 0 otherwise), sigmoid ($f(x) = \frac{1}{exp(x)+1}$), and hyperbolic tangent. After the calculation, the neuron will pass the output from its activation function to the next layer of neurons, and so on.

In a metaphorical sense, as introduced in many sources, a traditional neural network is in a shape of consecutive layers of neurons as shown in Figure 9, with lines connecting two neurons between consecutive layers. These lines, or "connections", represent weights used to calculate the weighted sum of each neuron from the previous layer. Two consecutive

neuron layers are required to be interconnected, that is there must be a connection between every two neurons in between two layers.

The first layer of a neural network, where inputs are fed, is called an **input layer**. The final layer which gives out outputs is called an **output layer**. Every other layers in between are called **hidden layers**.

3.6 How neural networks calculate the output

To obtain the output, or the prediction, of a neural network, the network requires an input event to be fed into the first layer of a network called an input layer. An input event is simply a set of numbers, and each number of input event, which can be called a **feature** in computer science literature, must be fed to each neuron in the input layer. Each neuron in the first hidden layer starts calculating a weighted sum of inputs using weights assigned between itself and every neuron in the previous layer, add its bias to the weighted sum if any, and applies its activation function to it. The hidden layer will pass on the outputs from its neurons to the next layer, and the procedure repeats until the calculation reaches the output layer.

Since this thesis mainly applies NN onto collision events, we will call features as **input variables** to better represent the use of variables in Physics analysis as inputs for NN.

3.7 Loss function

A neural network is designed to predict which class the input belongs to, and the training procedure for the network is supposed to configure the weights for the network to be able to predict the input event's *correct* class. During the training of a neural network, the output, or the prediction, of a neural network can be compared to the input event's truth class. Since the training uses a large set of data, we can evaluate how close the prediction from a network is by comparing the truth class of each input event and the predicted class from the network. A **loss function** is used to compare the difference for each input event, and it must be designed to give low output if the prediction is close to the truth value, and high output if the prediction is far from the truth value. A loss function can be as simple as a fraction of incorrect predictions to total predictions, or more complex functions such as binary cross-entropy. **Total loss** of a network is the average of loss function outputs from each input event, and as a result of loss function design, the total loss will be lower as the network better predicts the output.

Examples of loss function include mean squared error $(y-p)^2$ and binary cross-entropy $-(y\log p+(1-y)\log(1-p))$, where y is truth value and p is predicted value. These functions give zero loss if the network predicts the value correctly but behave differently if the network predicts incorrectly.

3.8 Event weighting

To emphasize the importance of some events over other insignificant events, **event weights** may be introduced during the training. Normally, the total loss of a network is the average of loss function outputs from each input event. If event weights are assigned, the total loss of a network is simply a weighted average of loss function outputs from each input event. Naturally, events with bigger weights are more significant than events with smaller weights, as it has a bigger impact on the total loss of a network.

3.9 Neural network weight tuning

To optimise the performance of a neural network after training, or to reduce the network loss, we must modify the weights of that neural network. The weights should not be tuned by random, but with an algorithm with its goal of reducing the network loss. The algorithm is called an **optimiser**. The most notable optimiser is Stochastic Gradient Descent (SGD), which is an algorithm tuning network weights by calculating the gradient of the loss concerning all network weights. During the training phase of the network, the network loss is calculated from a set of training events. Network weights are then tuned with the optimising algorithm. The duration in which the network loss is calculated and the network weights are tuned is called an **epoch**. Very often one epoch of neural network training is not enough to reduce the network loss to an optimal value, so several epochs of training are required.

3.10 Neural network performance measurement using Receiver Operating Characteristic curve

To measure the performance of a neural network, or any machine learning classifier, the **Receiver Operating Characteristic (ROC) curve** may be plotted. The curve may have several variants, but the variant used in this thesis is the plot of signal acceptance of the model at certain background acceptance. Signal acceptance is the ratio of signal events being accepted as signal-like events by the model, while background acceptance is the ratio of background events being accepted as signal-like events by the model. The curve is plotted by calculating the discriminator output value giving certain levels of background acceptance.

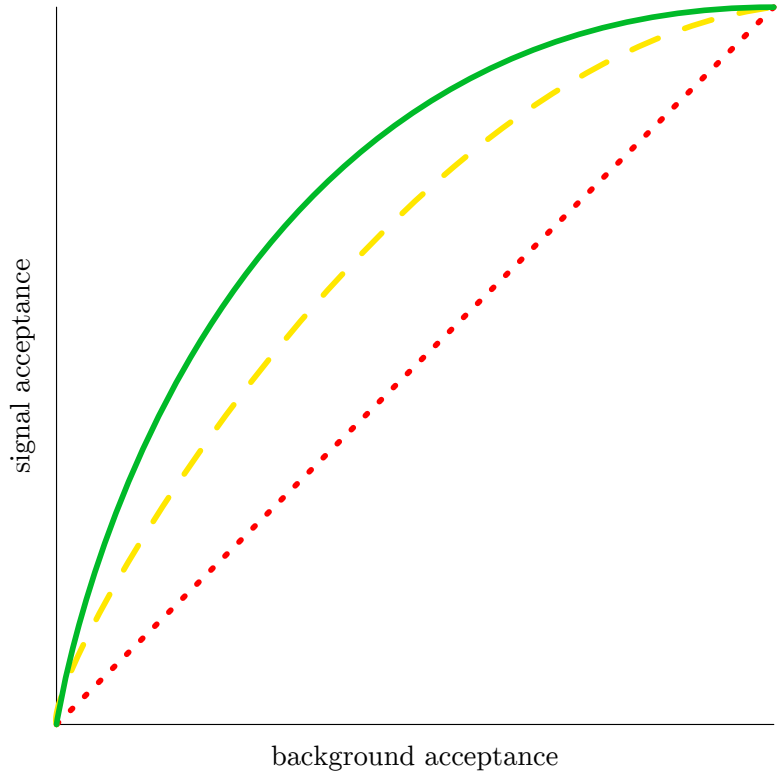


Figure 10: Receiver Operating Characteristic (ROC) curve samples. A solid green curve has a higher area under the curve than a dashed yellow curve, meaning the green curve is a better ROC curve. A dotted red curve represents a random guess.

The signal acceptance for each level of background acceptance is then calculated using the discriminator output value and plotted in a final ROC curve.

Using the ROC curve, we can calculate the area under the curve (AUC), which is a metric representing the performance of a discriminator. More area under curve signifies better signal acceptance at certain levels of background acceptance. As shown in Figure 10, the solid green line represents a better ROC curve since it contains more area under the curve than the dashed yellow curve. The dotted red curve represents random guess, which means one can obtain the equal signal and background acceptances with random guess alone.

CHAPTER IV

PREVIOUS SEARCHES FOR FOUR TOP QUARK PRODUCTION

Due to recent energy and luminosity upgrades to the LHC, four top quark production becomes more visible in the CMS detector because its cross section (according to SM at NLO) increases as the centre-of-mass energy increases from 8 TeV to 13 TeV. The integrated luminosity (a quantity describing how large the data has been collected from the detector) also increases due to recent upgrades to the LHC and longer periods of collection time. As of this writing, at least three research articles have been released by CMS collaboration that searches for four top quark production. In this thesis, we will perform an analysis based on single lepton channel; that is we are mainly looking at events from four top quark production decaying into a single lepton and jets. With this characteristic, this channel is sometimes referred in CMS literature as lepton + jets channel. This chapter will hence focus on previous analyses including the single lepton channel using the CMS detector: an analysis using the data collected with 8 TeV in 2012 [23], an analysis using the data collected with 13 TeV in 2015 [24], and an analysis using the data collected with 13 TeV in 2016 [12].

4.1 2012 analysis with $\sqrt{s} = 8$ TeV

This analysis, published in 2014, used the data recorded with a centre-of-mass energy of 8 TeV and corresponded to an integrated luminosity of 19.6 fb^{-1}. For a centre-of-mass energy of 8 TeV, the cross section of four top quark production, as predicted by SM at next-to-leading order, is extremely small at 1.3 fb. At the same centre-of-mass energy, top-antitop production cross section is 245.8 pb [25], five orders of magnitude greater than four top quark production cross section.

In this analysis, a search of four top quark production in single lepton channel (both with single electron and single muon) was performed. Boosted decision trees were used to discriminate between four top quark production and top-antitop production. The discriminator used 10 vairables based on multiplicity of top quarks generated in a collision event, the number and kinematics of jets, and the number of jets tagged as originating from bottom quark, or b-tagged jets in CMS literature. The events themselves were separated in three categories based on the number of jets: 6, 7, and > 7.

After BDT training, the *trained* BDT discriminator was used to calculate how likely a collision event is generated from four top quark production. BDT outputs of simulated collision events from signal process (four top quark production) and certain background process (including top-antitop production) were then populated in a histogram. BDT outputs of real collision events from CMS detector were also populated into the same histogram in order to compare and calculate the expected limits of four top quark production.

The BDT output distributions from simulated collision events were compared with the same distributions from real collision events. With three different event categories in two channels (single electron and single muon), the expected limit of four top quark production cross section is calculated with asymptotic CL_s [26] method from six sets of BDT output distributions. The calculated expected limit is 32 ± 17 fb, which is approximately 25 times the prediction of SM. The expected limit is, however, not a final quantity that indicates the validity of SM, as the data at this stage is inadequate for precise measurements.

4.2 2015 analysis with $\sqrt{s} = 13$ TeV

For this analysis, published in 2017, the data recorded from CMS detector was generated from proton-proton collision with 13 TeV centre-of-mass energy, and corresponded to an integrated luminosity of 2.6 fb^{-1}. With upgraded centre-of-mass energy, the predicted cross section for four top quark production, at SM NLO, jumps to 9.2 fb. Nevertheless, top-antitop production cross section also increases to 831 pb (at SM NLO), still five orders of magnitude greater than four top quark production cross section. The much higher increment ratio of four top quark production cross section, compared to top-antitop quark production, is due to its contributing terms at NLO.

This analysis presented searches for four top quark production both in single lepton channel and opposite sign dilepton channel. The analysis still useed boosted decision tree as a discriminator, which used 11 variables including number of jets, kinematic variables of jet system in an event, and variables involved with event topology. The discriminators were also trained separately for single lepton channel and opposite sign dilepton channel. The events themselves were also separated into categories based on number of jets (6, 7, 8, and > 8) and number of b-tagged jets (2, 3, and > 3).

As in the search in 2012, BDT outputs from certain processes, both from simulated and real collision events, were compared in a histogram. The expected limit was also

calculated with asymptotic CL_s method based on the BDT distributions. With BDT distributions from single lepton channel, separated into categories as described earlier, the expected limit was 151^{+90}_{-52} fb, roughly 17 times greater than the predicted cross section.

4.3 2016 analysis with $\sqrt{s} = 13$ TeV

This analysis used the data collected from CMS detector with the same centre-of-mass energy of 13 TeV. The data this time corresponded to an integrated luminosity of 35.8 fb^{-1}, almost 14 times larger than previous analysis. With the same centre-of-mass energy as in the same search in 2015, the cross sections predicted by SM at NLO for both four top quark production and top-antitop production are the same.

Fifteen variables derived from collision events were used to train the boosted decision tree, based on event topology, kinematics, and bottom quark multiplicity:

- `multitopness`, the output of another BDT responsible for determining combinations of three jets (trijet) decaying from top quark. The output is calculated from third possible system of three jets
- `HTH`, or the ratio of the scalar sum of transverse momentum of all jets to the scalar sum of momentum of all jets
- `HTb`, scalar sum of the transverse momentum values of all the b-tagged jets in the event
- `HTRat`, the ratio of the HT of the four leading jets in the event to the HT of the other jets in the event
- `HTX`, the scalar sum of transverse momentum of all jets except jets in highest ranking system of three jets (trijet)
- `SumJetMassX`, invariant mass of the system comprising all the jets except jets in highest ranking system of three jets (trijet)
- 1st, 2nd, 5th, and 6th highest jet transverse momentum
- Lepton transverse momentum
- Third and fourth highest combined secondary vertex (CSV) discriminator value
- Transverse momentum of jets with third and fourth highest CSV discriminator value

Table 1: Expected limit on the cross section of four top quark production in single lepton channel, obtained from previous analyses.

	Expected limit on cross section (fb)	Expected limit on cross section ($\times\sigma^{\mathrm{SM}}_{t\bar{t}t\bar{t}}$)
2012 search 19.6 fb^{-1} data, $\sqrt{s} = 8$ TeV	32 ± 17	24.6 ± 13.1
2015 search 2.6 fb^{-1} data, $\sqrt{s} = 13$ TeV	151^{+90}_{-52}	$16.4^{+9.8}_{-5.7}$
2016 search 35.8 fb^{-1} data, $\sqrt{s} = 13$ TeV	86^{+40}_{-26}	$9.4^{+4.4}_{-2.9}$

Prior to event-level discriminator, a combined secondary vertex v2 (CSVv2) b-tagging neural network was used to calculate how likely each jet in a collision event is caused by a bottom quark. Each jet in a collision event was then assigned a value called CSV b-tagging discriminator value.

The discriminator, based on boosted decision trees using `scikit-learn` [27] machine learning package, was trained with four top quark production samples as signal dataset against top-antitop production samples as background dataset, since top-antitop production is the most important background with the highest cross-section.

By using the same asymptotic CL$_s$ method as in previous analyses, the expected limit of four top quark production, calculated from single lepton channel, was calculated to be 86^{+40}_{-26} fb, 9.4 times greater than the predicted cross section.

The work in this thesis, which will be covered in subsequent chapters, will use the same simulated Monte Carlo dataset used in the 2016 search, and will also employ the same asymptotic CL$_s$ method in order to determine the sensitivity of the search for four top quark production. The difference between the 2016 search and this thesis is the design of machine learning discriminators. The discriminator in 2016 search is based on boosted decision trees, while the discriminator used in this thesis will be based on neural networks.

CHAPTER V

FOUR TOP QUARK PRODUCTION ANALYSIS WITH TRADITIONAL NEURAL NETWORK

As described in Section 1.6, the adversarial neural networks consist of two smaller networks: the discriminator network and the adversary network. Before we can design any adversary networks, the discriminator network must be designed in the first place. This chapter presents the design of traditional neural network discriminator as well as the implementation of the discriminator in four top quark production analysis, including calculations on the expected limit of four top quark production and significance.

5.1 Monte Carlo simulated datasets used in this analysis

As with every ML technique training, our neural network requires a dataset to train itself on. However, it is impossible to use real collision data to train our neural network to discriminate between top-antitop production and four top quark production, since we cannot determine exactly which process a collision event recorded with CMS detector comes from. Instead, simulated collision events with the predetermined processes are used to train the neural network. Simulated events generated from four top quark production are used as signal events, while simulated events generated from top-antitop production are used as background events, since top-antitop production is a dominant background. The Monte Carlo dataset is simulated under the same CMS virtual detector with 13 TeV centre-of-mass energy and is required to contain exclusively one lepton (electron or muon). The simulated samples are produced and approved by CMS collaboration to use.

5.2 Data preselection

All simulated datasets used in neural network training must pass the preselection criteria, the same as in 2016 analysis in Section 4.3, as follows:

- For single electron
  - Missing transverse energy (MET) > 50 GeV
  - Scalar sum of transverse momentum (HT) > 500 GeV
  - Lepton pseudorapidity $|\eta| < 2.1$
  - Lepton transverse momentum $p_T > 35$ GeV

  - Number of reconstructed jets ≥ 8
  - Passes trigger `HLT_Ele32_eta2p1_WPTight_Gsf`

- For single muon
  - Missing transverse energy (MET) > 50 GeV
  - Scalar sum of transverse momentum (HT) > 450 GeV
  - Lepton pseudorapidity $|\eta| < 2.1$
  - Lepton isolation < 0.15
  - Lepton transverse momentum $p_T > 26$ GeV
  - Number of reconstructed jets ≥ 7
  - Passes either trigger `HLT_IsoMu24` or `HLT_IsoTkMu24`

All reconstructed jets in this analysis are required to have $p_T > 30$ GeV and $|\eta| < 2.4$. Furthermore, in real data analysis, triggers are used to filter collision events from the detector. Since simulated datasets do not come from the detector itself, trigger emulations are applied in the simulated datasets instead of triggers. Trigger emulations applied on the simulated datasets provide roughly the same acceptance rate as in triggers used in real data analysis.

5.3 Custom input variables

In the previous analysis, 15 variables available in the dataset, described in Section 4.3 are used as inputs. To better analyse for differences between four top quark production (the signal process) and top-antitop production (the background process), more variables are needed to be derived from each simulated collision event. These variables may contain kinematics information of particles and the overall topology of the event.

Following is the list of variables that can be extracted directly from the simulated collision events.

- 3rd and 4th highest jet transverse momentum
- `LeadingBJetPt`, highest value of transverse momentum among all jets having CSV b-tagging discriminant higher than 0.8484
- `jet5and6pt`, a sum of fifth and sixth jet transverse momentum
- 1st to 4th highest CSV discriminator value

- Transverse momentum of reconstructed jets with first and second highest CSV discriminator value
- Number of reconstructed jets

Following is the list of variables further derived from the simulated collision events.

- `H` is the scalar sum of momentum from all jets
- `pt3HT` and `pt4HT` are the ratio of 3rd or 4th highest jet transverse momentum to the sum of transverse momentum from all jets (`HT`) respectively
- `sphericity` is defined as $S = (3/2)(\lambda_2 + \lambda_3)$ where λ_2 and λ_3 are the two smallest eigenvalues of a matrix $S^{\alpha\beta} = (\sum_i p_i^\alpha p_i^\beta)/(\sum_i |\vec{p}_i|^2)$. Indices α and β represent one of three components of a momentum vector for each jet. The definition is taken from [24].
- `invmass_34`, `invmass_35`, `invmass_36`, `invmass_45`, `invmass_46`, `invmass_56` represent the invariant mass of each dijet system (a system of two jets) where each jet in a system has third to sixth highest CSV b-tagging discriminant. In other words, a group of jets in each event are sorted based on CSV b-tagging discriminant, and jets having third to sixth highest CSV b-tagging discriminant are chosen, two at a time, to calculate the invariant mass of a dijet. For example, `invmass_35` represents the invariant mass of a dijet system composed of a pair of jets with third and fifth highest CSV b-tagging discriminant.
- `HT2M` is the scalar sum of transverse momentum from all jets, minus transverse momentum of two jets with highest CSV b-tagging discriminant (where the discriminant must be greater than 0.8484).
- `mean_csv` is the mean of CSV b-tagging discriminant of all jets. For some jets, the CSV b-tagging discriminator has insufficient information, and thus recorded the discriminant as -10. Such jets will have its discriminant counted as 0.
- `trijet1st_invmass`, `trijet2nd_invmass`, and `trijet3rd_invmass` are the invariant masses of each trijet (a system of three jets) chosen from an MVA discriminator responsible for finding the most probable set of three jets that may originate from a top quark. The MVA first picks the most probable set of three jets out of all jets in an event and repeats the process for the most probable set out of remaining jets.

Table 2: Number of signal and background events by number of jets.

	Number of events in		
Number of jets	Signal	Background	Signal/total events ratio
7	23 972	150 535	0.14
8	51 874	266 969	0.16
9	44 288	110 178	0.29
10 or more	53 776	44 145	0.55

- `angletoplep` is the angle between the most probable trijet system with the only lepton in an event.
- `angletop1top2` is the angle between two trijet system having the most and second most probable to be originated from a top quark.
- `pTRat1st` and `pTRat2nd` are the ratios between transverse momentum from the first and second trijet to the sum of transverse momentum of all jets.
- `topness`, `ditopness`, and `tritopness` are discriminant values from the MVA responsible for choosing the trijets. They correspond to the first, second, and third trijet in the event respectively.

With 15 original variables, 11 extracted variables, and 22 derived variables added to the list, we now have 48 variables that may be used as input variables for our neural network.

5.4 Training data choices

Since the neural network's performance depends on the training dataset, we may choose the dataset to have certain criteria on top of preselection criteria or remove some input variables that do not give a positive contribution to the performance of the neural network. This section discusses such choices for the training dataset.

5.4.1 Number of jets criteria

It is observed that, within the training dataset, events with 9 or more jets have more signal/total events ratio, as shown in Table 2. With this fact, it is possible for us to choose only events with 9 or more jets to train our neural network or choose all events regardless of the number of jets. In this analysis, we have trained neural networks with

both variants of the training dataset, and neural networks trained with each dataset will have their performances compared.

5.4.2 Input variables used

To test whether an input variable has a positive contribution to the neural network's performance, a neural network is set up and trained by using all 48 variables as input variables. A set of 48 new neural networks, with the same structure as the first neural network, are then trained with all except one input variable, to determine the decrease in network performance. It is found that a neural network with `SumJetMassX` input variable removed and another neural network with `pTRat2nd` input variable removed has higher area under curve (AUC) than a neural network with 48 input variables. We can then deduce that both `SumJetMassX` and `pTRat2nd` input variables give a negative contribution to neural network performance. With this discovery, we may choose to train our neural network to include all 48 input variables or only 46 variables that don't give a negative contribution to the performance. As per the criteria of the number of jets, neural networks trained with different sets of input variables will have their performances compared.

5.5 Event weighting

To better represent the proportions of signal and background events in real collisions, events are weighted based on the process's cross section. They are weighted in a way that events from top-antitop production have 831 760 events per 9.2 events from four top quark production. These numbers are based on the top-antitop production cross section and the four top quark production cross section, which are 831 760 fb and 9.2 fb respectively. In practice, however, event weighting only helps the network loss to converge much quicker, with a tradeoff of having outputs of an event being skewed. With this advantage of event weighting, it will be used in hyperparameter training only to save time. The final neural network that is used to calculate discriminator output values will not be trained with event weights.

5.6 Neural network structure

The neural network is built with Keras Machine Learning package in Python [28]. It consists of one batch normalization layer as an input layer, which "applies a transformation that maintains the mean activation close to 0 and the activation standard deviation close to 1". With this input layer, the input variables are automatically normalised and do not require us to preprocess them. The adjustment of this layer is carried out by using the procedure described in [29].

After the batch normalization layer, a number of hidden layers and neurons are permutated with hyperparameter search, which will be described in Section 5.8. With this hyperparameter search, we don't have to rely on only one particular network configuration, and instead, we can find the configuration with the optimal performance we are looking for. Finally, after several hidden layers, one final output layer will have one neuron with sigmoid activation, which guarantees the output to be in the range of 0 and 1. The final neuron will give its output close to 0 or 1 if the input event is background-like or signal-like respectively.

The whole neural network will have binary cross-entropy function, $-(y \log p + (1 - y) \log(1 - p))$, as its loss function, where y is the event's truth value and p is the predicted value from the network. Since logarithm function is used, this loss function will heavily punish the network if wrong predictions are made, making the training loss to converge faster.

5.7 Performance evaluation with Receiver Operating Characteristic (ROC) curve

In this analysis, several ROC curves (introduced in Section 3.10), as well as their area under each curve, are plotted using the following categories of collision events as inputs for the model:

- Low-jet multiplicity category (8 jets or lower, 8-J)
- 9 jets and 3 medium b-tagged jets (9J3M) category
- 9 jets and 4 or more medium b-tagged jets (9J4M) category
- 10 or more jets and 3 medium b-tagged jets (10J3M) category
- 10 or more jets and 4 or more medium b-tagged jets (10J4M) category

Table 3: AUC calculated from two boosted decision tree discriminators used in previous analyses, by jet categories.

	AUC from category				
	8-J	9J3M	9J4M	10J3M	10J4M
BDT (2015)	0.6701	0.6102	0.6019	0.6426	0.6400
BDT1 (2016)	0.7244	0.6444	0.6375	0.6597	0.6272

Categories with 9 or more jets are considered as categories with higher probability to contain signal events. The ROC curves are plotted using simulated dataset form top-antitiop production as background events since the production is a dominant background and simulated dataset four top quark production as signal events in corresponding categories.

In addition to calculating the AUC from neural networks, the ROC curve, along with its AUC, is also calculated from BDT discriminators in previous analyses. Table 3 shows the performance of the official BDT discriminators used.

5.8 Hyperparameter search

There is no absolute best neural network structure to work with, i.e. there is no golden rule to determine the number of neuron layers, the number of neurons in each layer, and many other properties in a neural network. Hence, we can experiment certain designs of a neural network and test which design delivers the optimal performance, which is measured by the area under the ROC curve in this analysis as described in Section 5.7. Such methods to iterate over a set of possible configurations is called hyperparameter search. In this analysis, configurations made throughout the hyperparameter search are:

- Two or three hidden layers
- 50, 100, or 200 neurons in each layer

These two possible sets of configurations lead to $2 \times 2 + 2 \times 2 \times 2 = 36$ permutations in total. Each neural network created from a single permutation is equally trained within 30 epochs.

5.9 Performance evaluation results

With 2 sets of training data, 2 sets of input variables, and 36 permutations in hyperparameter search, we can have 144 possible neural network combinations in total.

Table 4: AUC calculated from 144 possible neural network configurations, sorted by AUC from 10J4M category. Shown in table are 20 neural network configurations giving best AUC in 10J4M category.

	8-J	9J3M	9J4M	10J3M	10J4M
FourTopsHypermodelAlt/9up_vars_with_extras_200_200_100_.hdf5	0.6676	0.61825	0.62649	0.66664	0.66723
FourTopsHypermodelAlt/9up_vars_remove_less_AUC_200_100_.hdf5	0.66141	0.60823	0.60847	0.66406	0.66453
FourTopsHypermodelAlt/9up_vars_with_extras_200_100_50_.hdf5	0.68391	0.60688	0.61138	0.65822	0.66332
FourTopsHypermodelAlt/9up_vars_remove_less_AUC_100_200_200_.hdf5	0.65179	0.59955	0.61729	0.65536	0.66324
FourTopsHypermodelAlt/9up_vars_remove_less_AUC_200_200_200_.hdf5	0.66613	0.59917	0.605	0.65705	0.6596
FourTopsHypermodelAlt/9up_vars_with_extras_200_200_50_.hdf5	0.67838	0.61945	0.6183	0.65883	0.65823
FourTopsHypermodelAlt/vars_with_extras_200_50_200_.hdf5	0.67654	0.61355	0.61618	0.65904	0.65798
FourTopsHypermodelAlt/9up_vars_with_extras_100_200_200_.hdf5	0.66749	0.56915	0.59742	0.63355	0.65749
FourTopsHypermodelAlt/9up_vars_with_extras_200_100_100_.hdf5	0.65926	0.6077	0.59782	0.65721	0.65739
FourTopsHypermodelAlt/vars_remove_less_AUC_100_200_50_.hdf5	0.66908	0.61602	0.607	0.65599	0.65651
FourTopsHypermodelAlt/9up_vars_with_extras_100_100_50_.hdf5	0.62773	0.5709	0.5813	0.64196	0.65579
FourTopsHypermodelAlt/vars_remove_less_AUC_200_200_100_.hdf5	0.67344	0.61516	0.61294	0.6582	0.65377
FourTopsHypermodelAlt/9up_vars_remove_less_AUC_200_200_50_.hdf5	0.69619	0.60747	0.61485	0.65506	0.6533
FourTopsHypermodelAlt/9up_vars_remove_less_AUC_200_200_.hdf5	0.64131	0.57333	0.57943	0.63763	0.65026
FourTopsHypermodelAlt/9up_vars_with_extras_50_50_50_.hdf5	0.63286	0.59137	0.5957	0.64947	0.65005
FourTopsHypermodelAlt/9up_vars_remove_less_AUC_200_100_100_.hdf5	0.66794	0.60769	0.60271	0.65634	0.64889
FourTopsHypermodelAlt/9up_vars_with_extras_200_50_100_.hdf5	0.66235	0.60154	0.60424	0.64741	0.64816
FourTopsHypermodelAlt/9up_vars_with_extras_200_50_50_.hdf5	0.67362	0.60978	0.5907	0.65947	0.64715
FourTopsHypermodelAlt/9up_vars_with_extras_50_200_50_.hdf5	0.65272	0.60242	0.58806	0.65114	0.64662
FourTopsHypermodelAlt/vars_with_extras_200_100_200_.hdf5	0.6915	0.60806	0.62259	0.64042	0.64632

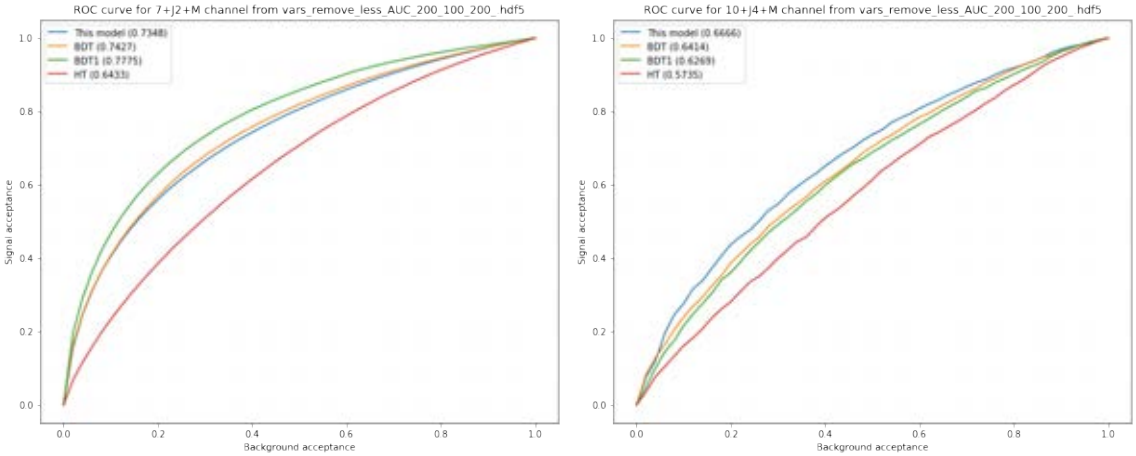
Table 4 shows the best 20 out of 144 possible neural network combinations, sorted by AUC from 10J4M category.

Coloured indicators in Table 4 represent the performance compared to BDT discriminators used in previous analysis shown in Table 3. The green indicator represents better performance than both official discriminators, yellow indicator represents the performance in between two official discriminators, and red indicator represents worse performance than both official discriminators. From the table, the best neural network configuration in terms of AUC from 10J4M category has three hidden layers, with 200, 200, and 100 neurons in each layer. The network has a better performance in the 10J4M category, albeit performing slightly worse in low-jet multiplicity category (with 8 jets or less). The ROC curve from the best network is shown in Figure 11.

5.10 Neural network discriminator structure conclusion

From the performance evaluation shown in Sections 5.6 and 5.9, we may conclude that the optimal neural network structure in which we will use in this analysis contains hidden layers as follows:

- One batch normalization layer
- Two hidden layers with 200 neurons each



(a) inclusive category

(b) 10J4M category

Figure 11: ROC curves from the best neural network configuration, calculated from inclusive and 10J4M category.

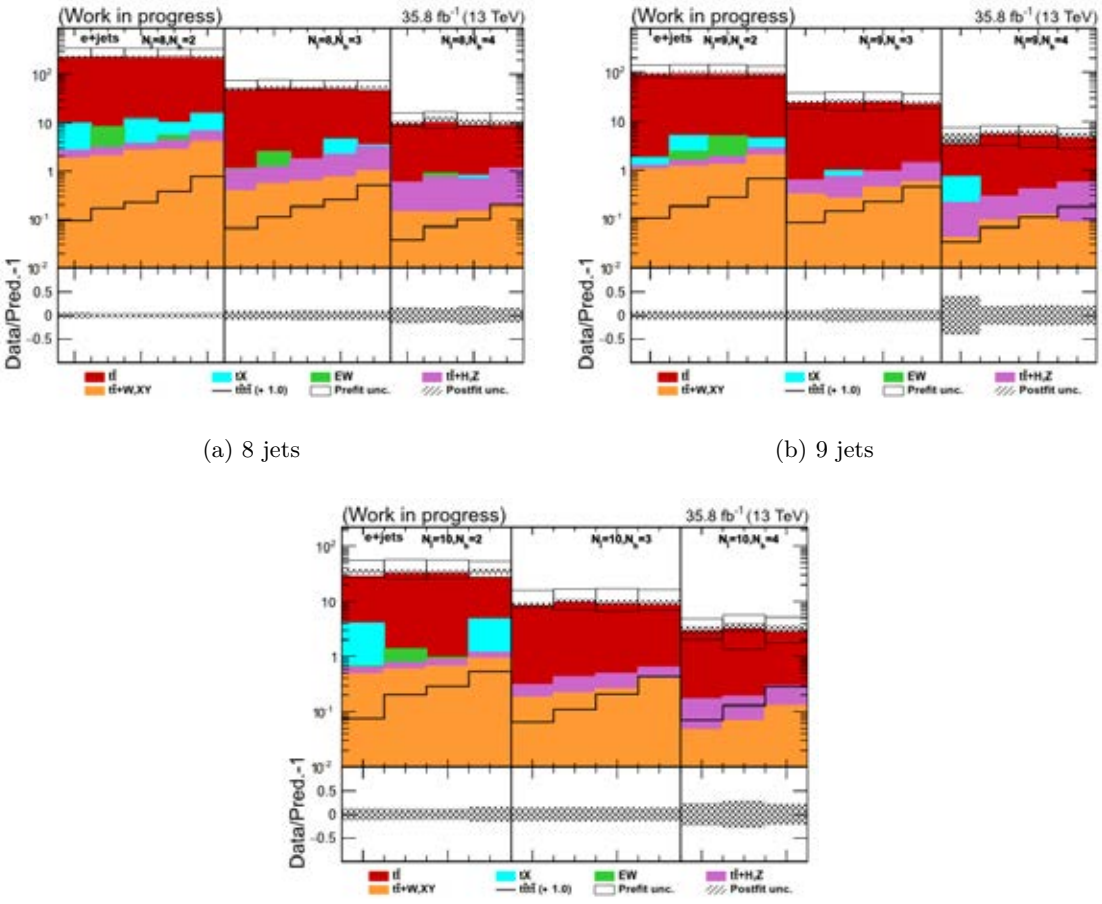
- One hidden layer with 100 neurons and tanh activation
- One final layer with one output neuron and sigmoid activation

The neural network is also trained with a dataset containing events with 9 or more jets only, using all 48 variables as described in Subsection 5.4.2. The final neural network discriminator will have the structure as described earlier and retrained with the same dataset, albeit without any event weighting.

5.11 Using neural network output in four top quark analysis

After obtaining the final neural network mentioned in Section 5.10, the neural network output is then calculated from the final network. All simulated events, including events generated from four top quark production as signal events, top-antitiop quark production as background events, and other background processes, are given a neural network output value for each process. The output values are then populated on histograms, categorised by single electron channel and single muon channel, number of jets, and number of b-jets.

The neural network output distribution, shown in Figures 12 and 13, can be used to calculate the cross section expected limit for the four top quark production, using asymptotic CL_s method. [26] This method calculates the expected limit of a parameter called signal strength μ. For event measurements, the expectation value n_i for the ith bin can be written as $E[n_i] = \mu s_i + b_i$, where s_i and b_i are the numbers of signal and background events in

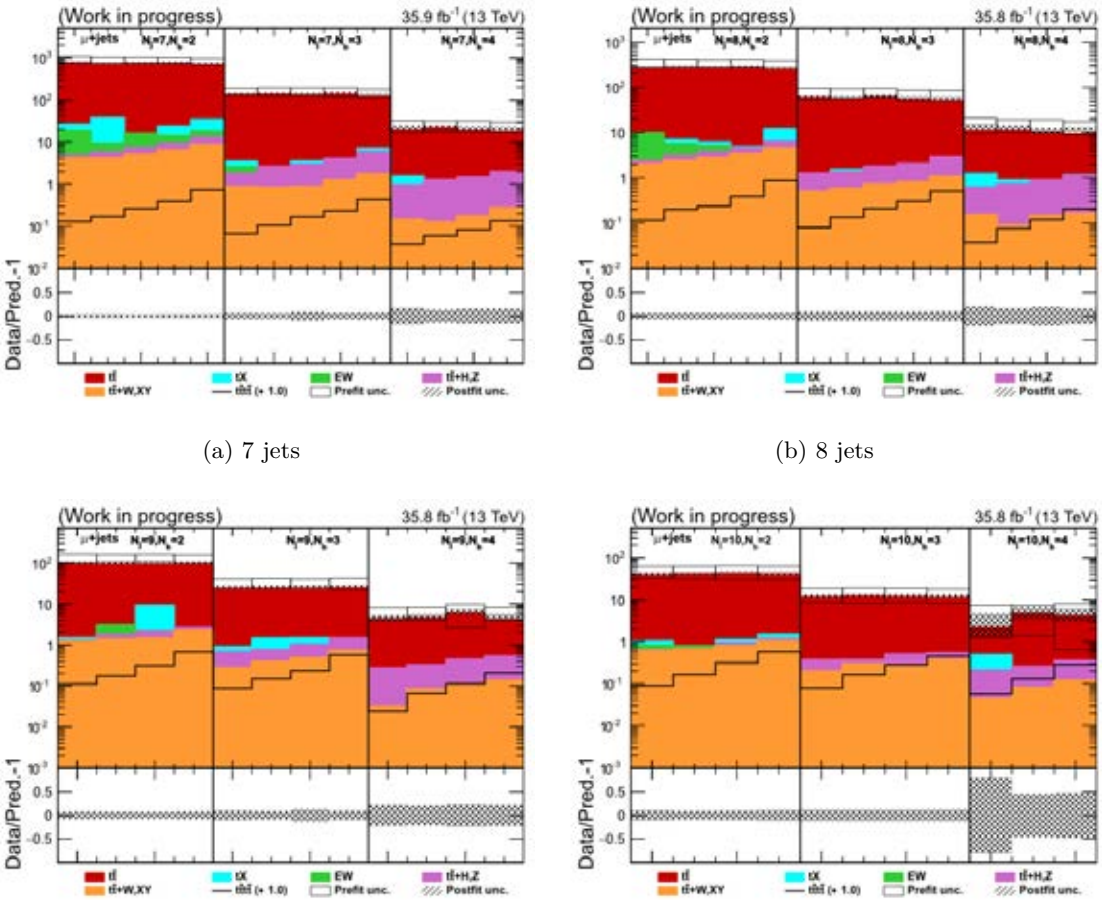


(a) 8 jets

(b) 9 jets

(c) 10+ jets

Figure 12: Histograms of traditional neural network output distribution from single electron channel.



(a) 7 jets

(b) 8 jets

(c) 9 jets

(d) 10+ jets

Figure 13: Histograms of traditional neural network output distribution from single muon channel.

the ith bin respectively. $\mu = 0$ represents the case where signal events are not expected, so-called background-only hypothesis, where $\mu = 1$ represents the case where the number of signal events is expected to be observed as described in signal bins. The expected limits for signal strength are calculated from profile likelihood, a quantity that describes how likely signal strength would have a certain value, based on the Poisson probabilities from each histogram bin.

In this work, *blind* expected limits will be calculated. In many High-Energy Physics analyses, *blind* analysis means an analysis in which data events, or the events recorded directly from CMS, are not used or *blinded*. For blind expected limits, this means that the expectation value for each bin in each histogram is merely from background histograms, not data histograms. Since the data is blinded in this analysis, Figures 12 and 13 do not have data points present in them.

The CL_s method requires the histogram of neural network output from certain background processes, such as top-antitop production, and the signal process (four top quark production). It also requires the histogram not to contain any bins with zero events from all background processes. To obtain stable statistical behaviour of the simultaneously binned maximum-likelihood fit used in this method, each histogram must be binned in such a way that each bin would contain a roughly equal amount of background events. With the CL_s method, systematic uncertainties also appear in the form of the uncertainty of the number of events in each histogram bin. The uncertainty in the histogram bins can impact the statistical fit for the expected limit.

The expected limit of four top quark production cross section calculated from the traditional neural network output discriminator is compared against the expected limits calculated in previous analysis in Table 5, comparing the results from this analysis to the results from boosted decision tree (BDT) discriminator used in 2016 analysis as described in Section 4.3.

As shown in Table 5, the expected limit *uncertainty range* calculated from the traditional neural network is smaller compared to the *uncertainty range* from BDT discriminator outputs. Our goal in this thesis is not to reduce the central value of the expected limit (90 fb from BDT versus 78 fb from traditional neural network), but to reduce the range of the upper limit, which is $37 + 24 = 61$ fb for traditional neural network versus $42 + 27 = 69$ fb for BDT.

Table 5: Expected limit and significance of four top quark production cross section, at 35.8 fb^{-1}, calculated from traditional neural network output (NN), compared to sensitivity of boosted decision tree (BDT).

	For 35.8 fb^{-1} data	
	BDT	NN
Expected limit (fb)	90^{+42}_{-27}	78^{+37}_{-24}
Expected significance	0.21	0.25

Table 6: Expected limit and significance of four top quark production cross section, at 200 fb^{-1}, calculated from traditional neural network output (NN), compared to the sensitivity of boosted decision tree (BDT).

	For 200 fb^{-1} data	
	BDT	NN
Expected limit (fb)	46^{+21}_{-14}	36^{+17}_{-11}
Expected significance	0.41	0.52

5.12 Expectations on 200 fb^{-1} data

We may also expect to use more data recorded from the CMS detector over time. The data with integrated luminosity of 200 fb^{-1} is expected to achieve with Run 2 CMS dataset recorded in 2015 - 2018, and until we obtain such amount of data, we may calculate expected limits and significances achievable at this amount of integrated luminosity, in the same manner explained in Section 5.11.

As shown in Table 6, the expected limit uncertainty range becomes smaller. Once again, we do not aim to reduce the central value of the expected limit (46 from BDT versus 36 from traditional neural network), but we focus on lessening the uncertainty range from $21 + 14 = 35$ fb for BDT to $17 + 11 = 28$ fb for traditional neural network. Also, the expected significance becomes higher for the traditional neural network. This slight improvement leads to another question: would the expected limit uncertainty decrease if we design a neural network to be resilient to the systematic uncertainty with most impact? The analysis involving adversarial neural networks, which aims to design such neural network, will be discussed in the next chapter.

CHAPTER VI

FOUR TOPS PRODUCTION ANALYSIS WITH ADVERSARIAL NEURAL NETWORK

In the previous chapter, we have seen that a traditional neural network can give better results than boosted decision tree in terms of expected limits and significance. Still, the neural network is susceptible to systematic uncertainties, as we did not design it to be resilient to any of them. This chapter discusses the design of adversarial neural networks, as well as the analysis using a discriminator based on the adversarial networks.

6.1 What are adversarial networks?

As described in Section 1.6, an adversarial neural network (ANN) are an extra neural network that determines why a discriminator network classify an event as signal-like or background-like events. The adversary network takes only the outputs of the discriminator as inputs and predicts certain features depending on its design, as shown in Figure 14. With the adversary network, we now have two networks to train with: a discriminator neural network and an adversarial neural network, with each network containing its loss. The loss for this discriminator network is calculated from how incorrectly the discriminator predicts the class of individual entries, while the loss for the adversary network is calculated from how incorrectly the adversary predicts the features of the input event.

In our case of adversary network training, the total loss for both discriminator and adversary networks is usually $L = L_D - \lambda L_A$, where L_D is discriminator network loss and L_A is adversary network loss. One parameter λ signifies the importance of adversary

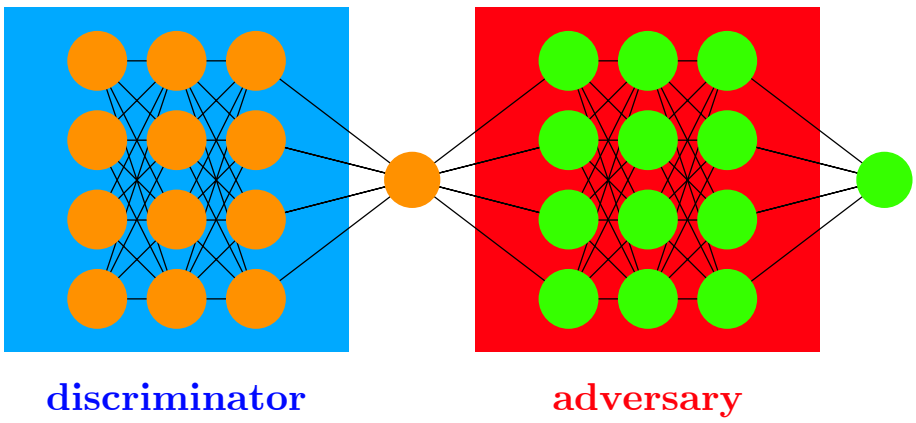


Figure 14: Simple diagram of adversarial networks.

network loss over discriminator network loss. Given the total loss in this form, the entire network can reduce the total loss by either reducing the discriminator loss, which improves the accuracy of discriminator network classifying the events, or increasing the adversarial loss, which worsens the accuracy of the adversary network. Our goal of training with the adversary network is we want the output of the discriminator network to be resilient to the most impactful systematic uncertainty. We also want to design the adversary network to guess whether an event contains that uncertainty. If we can train the discriminator network to increase the adversarial loss, we will have the discriminator network that is resilient to such systematic uncertainty.

6.2 Systematic uncertainty with most impact

Before designing the adversary network, we must first determine which systematic uncertainty induces most impact on the signal strength (as described in Section 5.11). To calculate an impact each particular uncertainty induces, we may calculate the deviation of signal strength when the uncertainty is added to the variable distribution (which, in this case, is traditional neural network output) by $+1\sigma$ or -1σ.

In this work, we have also calculated the impacts upon the signal strength of each uncertainty, both systematic and statistical. The summary plot is shown in the right part of Figure 15. From the summary plot, it is shown that **HeavyFlav uncertainty**, which is the uncertainty in the rate of $t\bar{t} + b\bar{b}$ in top-antitop production, has the most impact on signal strength.

With this information, it is clear that we may choose to design our adversary network to mitigate the impacts on HeavyFlav uncertainty only. Hence, for the rest of this chapter, HeavyFlav uncertainty will be focused, and the adversary network will be designed to handle this particular uncertainty.

6.3 Adversary network structure

The adversary network, like the discriminator network described in Chapter 5, is also built using Keras ML package written in Python [28]. Due to the constraint of time, permutating through several hidden layer structures is not possible. The adversary network contains an input layer, which directly takes the output of the discriminator network as its input. The input layer is then followed by a batch normalization layer, which normalizes the input in the same way as in the discriminator network. Three hidden layers with 50 neurons, each having a sigmoid activation function (in the form of $1/(1 + e^x)$) as its

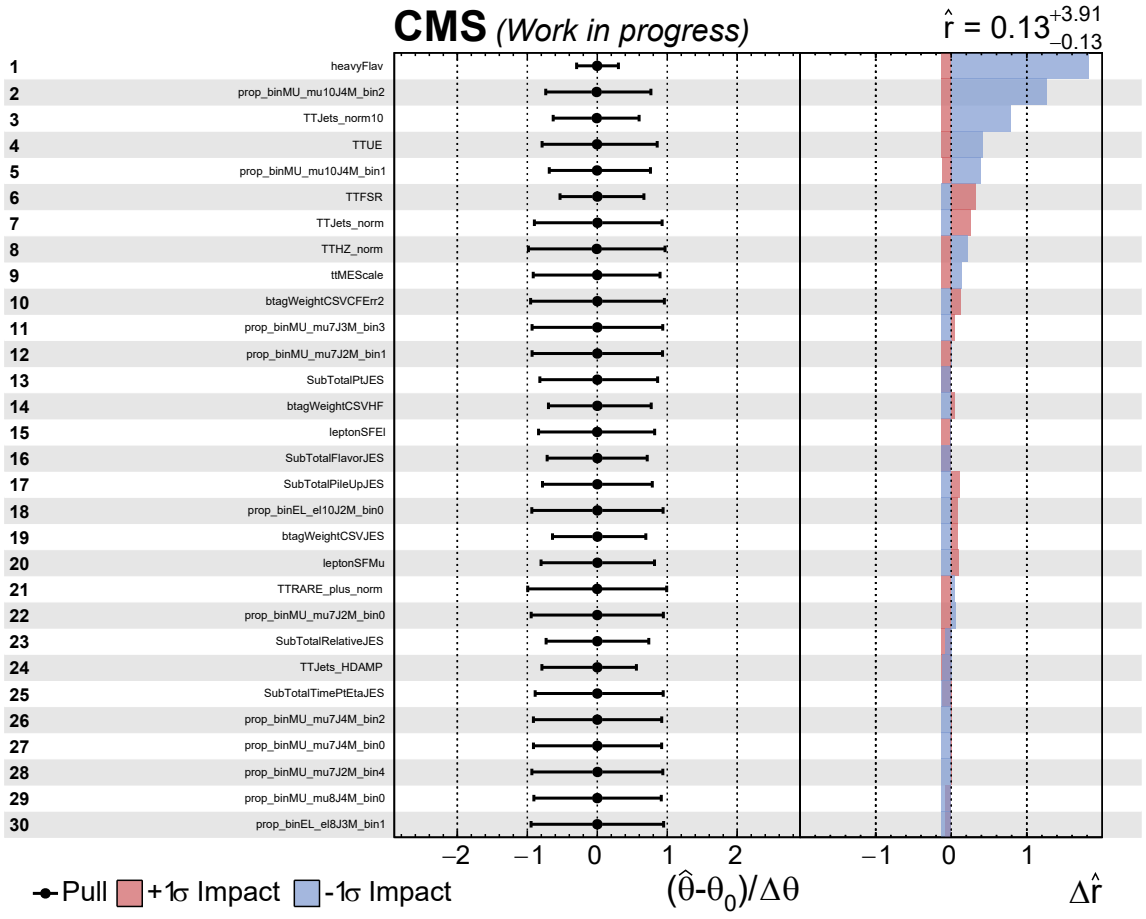


Figure 15: Impacts plot of signal strength from each uncertainty.

activation function, follows the batch normalization layer. A final output layer contains two (or three) neurons with softmax activation function, guaranteeing that the sum of all outputs of the network will be 1. Each neuron in the final output layer represents the probability in which an event falls into a particular category. A diagram depicting the adversary network structure is shown in Figure 16.

In this work, two variants of total losses of the entire network are designed and will be discussed in Section 6.6.

6.4 Training data used in adversary network training

The dataset used to train our adversary network is the same dataset used to train our final neural network described in Section 5.10. This means the dataset will have events with 9 or more jets and passes the same preselection criteria as described in Section 5.2. Furthermore, to add useful information for an adversary network, the information on whether or not an event contains HeavyFlav uncertainty is also recorded and can be used as truth values for the adversary network.

Since the training dataset is derived from simulated Monte Carlo datasets, each event is originated from a preassigned process, along with preassigned uncertainties. In other words, we know exactly which process an event in training dataset is originated from, and what kind of uncertainties does the event contain. This allows us to label truth values of each training event for the adversary network.

6.5 Adversary network training procedure

The adversary network training procedure, adapted from Louppe's algorithm in [16], is as follows:

1. Pretrain the discriminator network with the training dataset. In this case, we have already obtained the *pre-trained* discriminator network from Section 5.10, so we can use it as a basis for the discriminator network.

2. Pretrain the adversary network. The adversary network in this work is trained for 30 epochs. During the training of the adversary network, each training event is inputted into the discriminator network. The output of the discriminator network is then inputted into the adversary network. Finally, the adversary network will be adjusted based on the output from the adversary network, compared to the truth value (i.e.

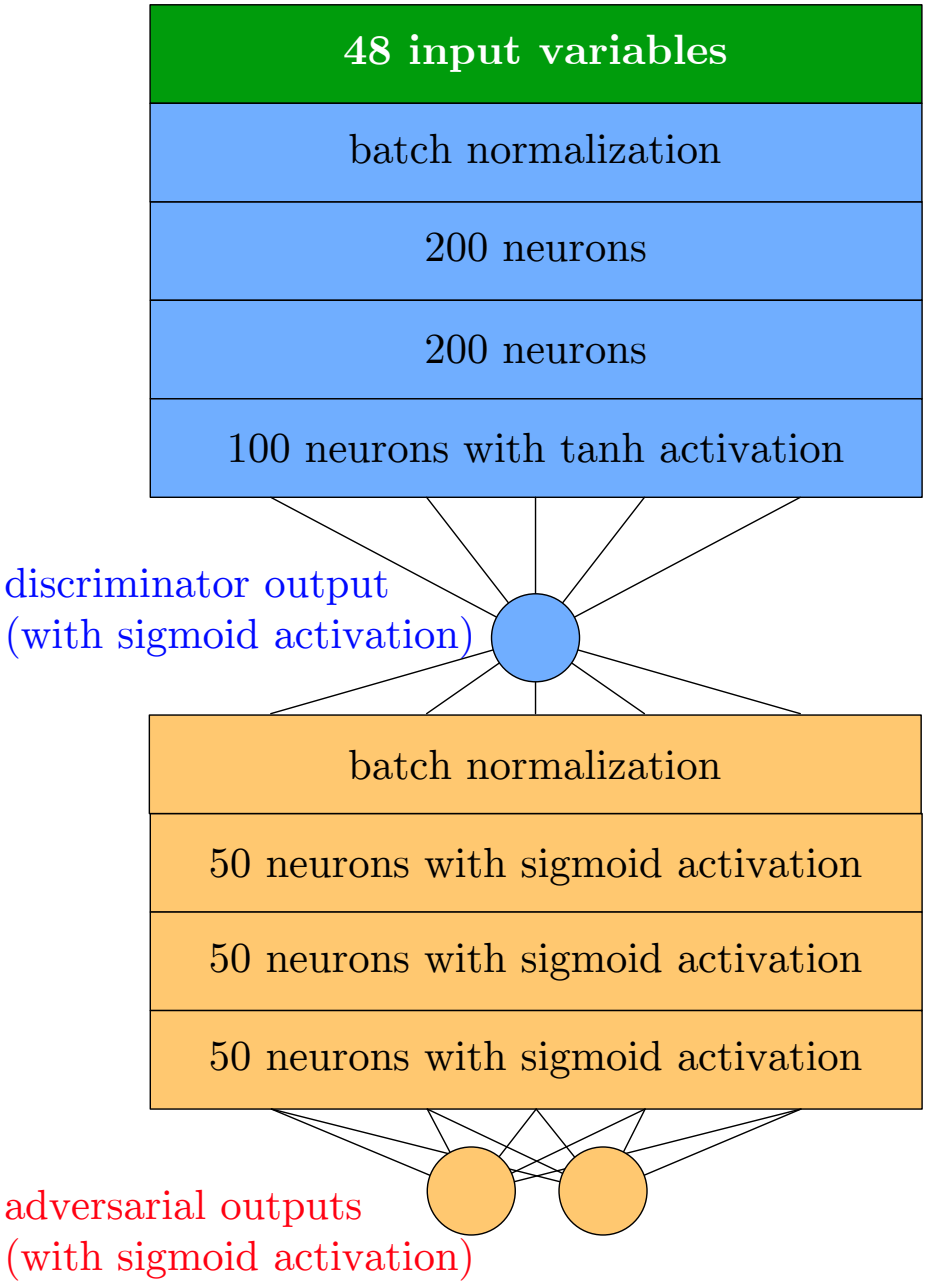


Figure 16: Adversary network structure, shown in orange below discriminator network structure, shown in blue. The output layer of adversary network is shown with two neurons in this figure.

whether or not the event contains the systematic uncertainty). The discriminator network is not yet adjusted in this step of the training.

3. For several epochs, such as 200 epochs,

   (a) Train the discriminator network for one epoch, using a small batch of training data. During the training, the adversary network is locked, which means the weights in the adversary network does not change.

   (b) Train the adversary network for five epochs, using another batch of training data with the same size. As with discriminator network training, the discriminator network is locked this time, fixing weights in the discriminator network.

4. After a sufficient number of epochs in the previous step, the discriminator network is extracted. The extracted discriminator network is said to be trained with an adversary network.

6.6 Network loss schemes

For the following loss function definitions, let HF, NHF, and sig denote background events with HeavyFlav uncertainty, background events without HeavyFlav uncertainty, and signal events. Also, let y and p denote truth and predicted value respectively. For the discriminator network, let y_D and p_D denote discriminator truth and predicted values.

Two network loss variants are devised for the adversary network. They are:

1. **Standard variant**

   In this case, the adversary network is naïvely designed, using simple loss function. This variant of the adversary network is trained with all events, both background and signal events. As such, the adversary network has **two** outputs that determine whether an event contains HeavyFlav uncertainty or not, based on the information of discriminator output alone. The discriminator loss function is binary cross-entropy, while the adversarial loss function is categorical cross-entropy since there are multiple outputs from the adversary network.

   The total loss function during discriminator network training is simply $L_D - \lambda L_A$, where L_D and L_A are discriminator loss and adversarial loss respectively. During adversary network training, the adversary network will be trained based on adversarial loss alone. With λ parameter present in a total loss, this parameter serves as the importance of adversary training in this variant of the adversary network.

The problem with this variant, however, is that only background events may have HeavyFlav uncertainty, while signal events are not affected by this uncertainty. HeavyFlav uncertainty is derived from the uncertainty of the rate of $t\bar{t} + b\bar{b}$ in top-antitop production (the dominant background process), which does not affect four top quark production (the signal process). This difference between background and signal events means events not containing HeavyFlav uncertainty may be background *or* signal events, which should have different discriminator output distributions. The discriminator output distribution for events without HeavyFlav uncertainty is, therefore, a mix between background events and signal events, and this will confuse the training of discriminator network during adversary training. This problem becomes a motivation for the second loss variant.

2. **"New loss" variant**

For this variant of the adversary network, the adversary network is designed to make discriminator outputs from background events with and without HeavyFlav uncertainty to be the same, while retaining the shape of discriminator outputs from signal events to be as close to 1 as possible. With this goal, the adversary network contains **three** outputs that determine whether an event is a background event with HeavyFlav uncertainty, a background event without HeavyFlav uncertainty, or a signal event. As with standard variant, the adversary network will use the information of discriminator output alone.

The total loss function during the training is defined separately for discriminator network training and adversary network training. While adjusting the adversary network, the loss is simply in the form of categorical cross-entropy:

$$-\left[y_{HF}\log p_{HF} + y_{NHF}\log p_{NHF} + y_{sig}\log p_{sig}\right]$$

On the other hand, the total loss while adjusting the discriminator network is

$$-\left[y_D \log p_D + (1-y_D)\log(1-p_D)\right] + \lambda\left[y_{HF}\log p_{HF} + y_{NHF}\log p_{NHF}\right]$$

Once again, λ parameter signifies the importance of adversary training.

6.7 Hyperparameter training for adversary network

As shown in Section 6.6, there are two variants of network loss with different training goals. This means we must use different criteria to choose the optimal value of the hyperparameter λ for each network variant.

Table 7: AUC of neural networks trained with the adversary network, standard variant, and several values of λ hyperparameter. AUC calculated from different event categories based on the number of jets and the number of b-tagged jets. AUC calculated using boosted decision tree in [12] (BDT) and neural network not trained with the adversary network (No tuning) are also shown.

λ	8-J	9J3M	9J4M	10J3M	10J4M
0.01	0.72082	0.69336	0.68709	0.73012	0.71473
0.05	0.72198	0.69406	0.68778	0.73028	**0.71597**
0.1	0.71974	0.69260	0.68672	0.72952	0.71438
0.5	0.72242	0.69259	0.68466	0.72944	0.71304
1	0.72499	0.69305	0.68346	0.72947	0.71414
5	0.72893	0.69044	0.67866	0.72953	0.71445
10	0.73095	0.68729	0.67454	0.72866	0.71306
50	0.73808	0.68392	0.66987	0.72645	0.70465
100	0.73819	0.68420	0.67088	0.72450	0.70569
BDT	0.72440	0.64440	0.63750	0.65970	0.62720
No tuning	0.74333	0.68775	0.68170	0.72805	0.71151

For the standard variant, the goal of hyperparameter tuning is to choose the optimal value of λ that gives the optimal AUC value in 10J4M category (events containing 10 or more jets, with 4 or more jets tagged as b-tagged jets). Due to a limited time in this thesis, the optimal value of λ is chosen from a limited set of values from 0.01 to 100, as shown in Table 7.

By inspecting the AUC calculated from 10J4M category in Table 7, a neural network trained with λ value of 0.05 has the best AUC in the 10J4M category. With this result, we may use a neural network trained with adversary network and with $\lambda = 0.05$ as the best possible network in the standard variant of the adversary network. As stated earlier, the λ parameter only indicates the importance of adversary training. It does not represent any weighting of systematic uncertainties over statistical uncertainty. Rather, having unsuitable values of λ can give too much or too little significance of the training towards the goal to increase the adversarial network loss. If the value of λ is too much, the discriminator will abandon the goal of discriminating the events to make the output indistinguishable to adversary network. On the other hand, if the value of λ is too little, the effect of adversary training will be unnoticeable, similar to the case with no training (or $\lambda = 0$).

For the "new loss" variant of adversary network training, the goal is set differently. Since the discriminator network is designed to give output distribution from background

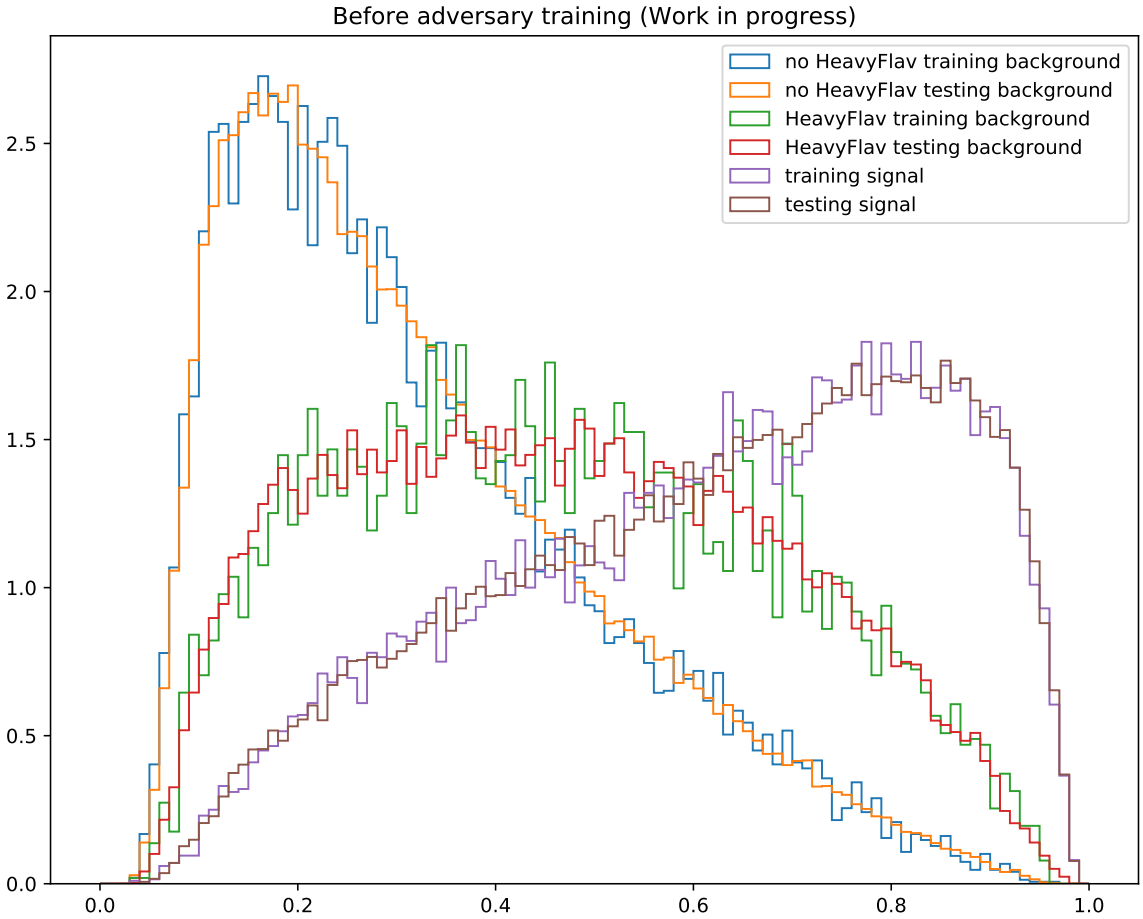


Figure 17: Neural network output distribution before adversary training, normalised.

events with and without HeavyFlav uncertainty, while retaining the output distribution from signal events to be the same, we will primarily inspect the output distribution of three categories of events: background events with HeavyFlav uncertainty, background events without HeavyFlav uncertainty, and signal events. We expect that the AUC achievable by NN discriminators after adversary training will be reduced, as a tradeoff to obtain the discriminator with the exact behaviour we want.

The output distribution for the traditional neural network is shown in Figure 17, separated by three categories of events. To check for signs of overtraining, the distribution is further separated into training and testing datasets. The distribution for training and testing datasets in the same category must be identical, which means overtraining does not occur.

Adversarial network pivoting with varying values of λ has been performed to obtain the discriminator that perfectly accomplishes our goal. By tuning the values of λ, the best value of λ giving the perfect distribution is found to be $\lambda = 0.3$, and the output distribution

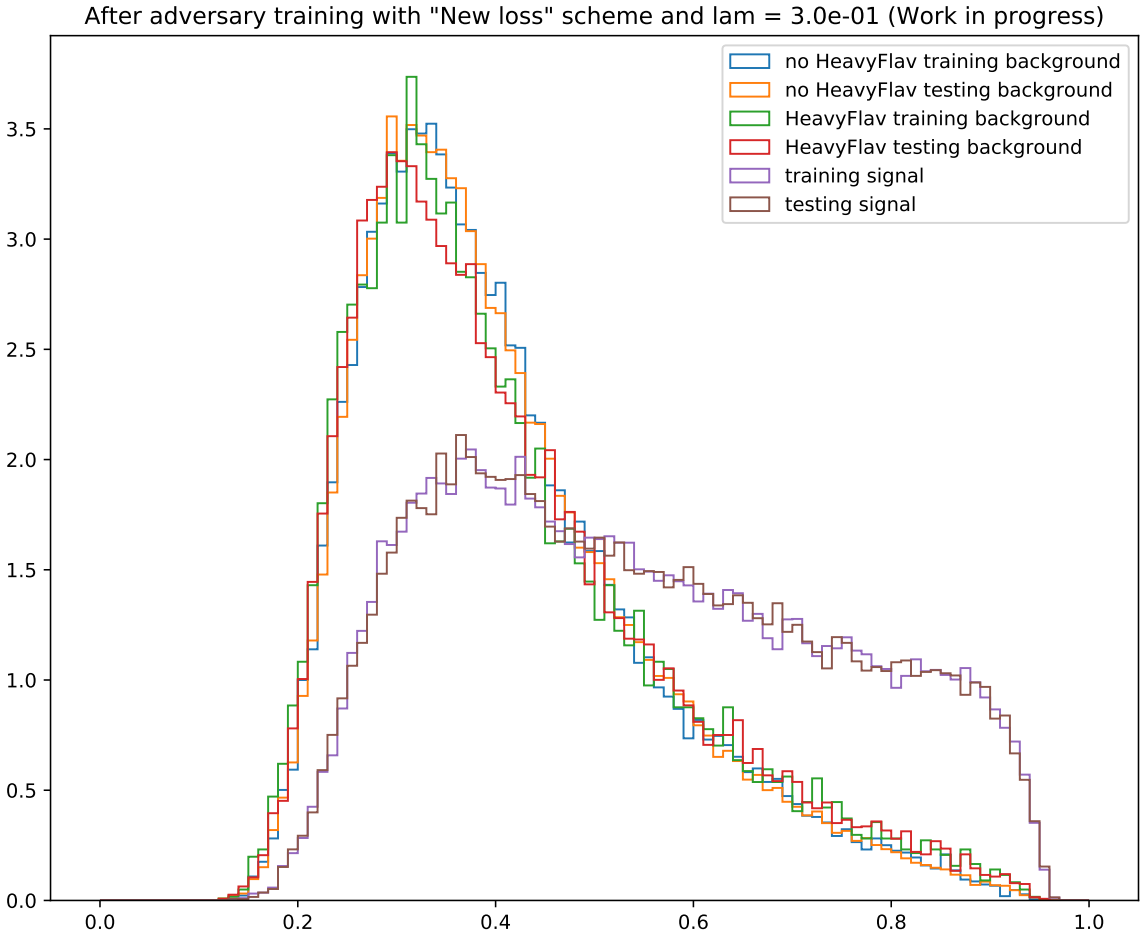


Figure 18: Neural network output distribution after adversary training with "New loss" scheme and $\lambda = 0.3$, normalised.

Table 8: Expected limit and significance of four top quark production cross section, at 35.8 fb^{-1}, calculated from the output from both variants of adversarial neural networks (ANN), compared to sensitivities of traditional neural network (NN) and boosted decision tree (BDT).

	For 35.8 fb^{-1} data			
	BDT	NN	ANN (Standard)	ANN (New loss)
Expected limit (fb)	90^{+42}_{-27}	78^{+37}_{-24}	79^{+37}_{-24}	83^{+39}_{-25}
Expected significance	0.21	0.25	0.24	0.23

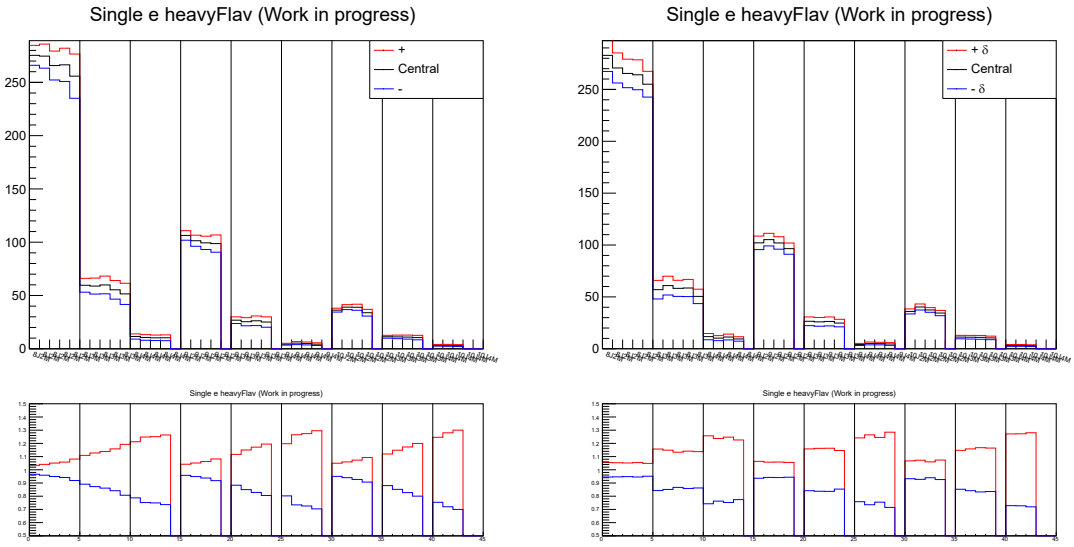
obtained via adversary training with the optimal value is shown in Figure 18. As shown in the figure, two distribution shapes from background events with and without HeavyFlav uncertainty are identical, while both of them are different from distribution shape from signal events. The distributions between training and testing datasets also do not show any signs of overtraining. With this perfect behaviour, we may use this neural network discriminator as the best possible network in the "New loss" variant.

6.8 Expected limit and significance calculated for 35.8 fb^{-1} data

Both of the best neural networks tuned with hyperparameter in each variant from Section 6.7 are then used to calculate the expected limit and significance in the same manner as described in Section 5.11. The expected limits of four top quark production cross section and expected significances calculated from both variants of adversary network training are shown in Table 8.

As shown in Table 8, the expected limit uncertainties from both variants of adversarial neural networks does not improve over the traditional neural network. Also, the expected significances of both variants also decreas slightly. As this is the first ever attempt for adversarial neural networks, these results are comparable to the results from the traditional neural network.

However, further investigations on HeavyFlav uncertainty have shown that the uncertainty distribution from adversary network in "New loss" variant becomes flatter when compared to those from the traditional neural network. Figures 19 and 20 show the difference between HeavyFlav uncertainty distributions for a neural network with and without adversary network training. Both figures represent the number of events from top-antitop production from each bin in each histogram generated by the procedure stated in Section 5.11, along with HeavyFlav uncertainty according to each bin. Both figures are also provided with their ratio plot below.



(a) with traditional NN

(b) with ANN (New loss variant)

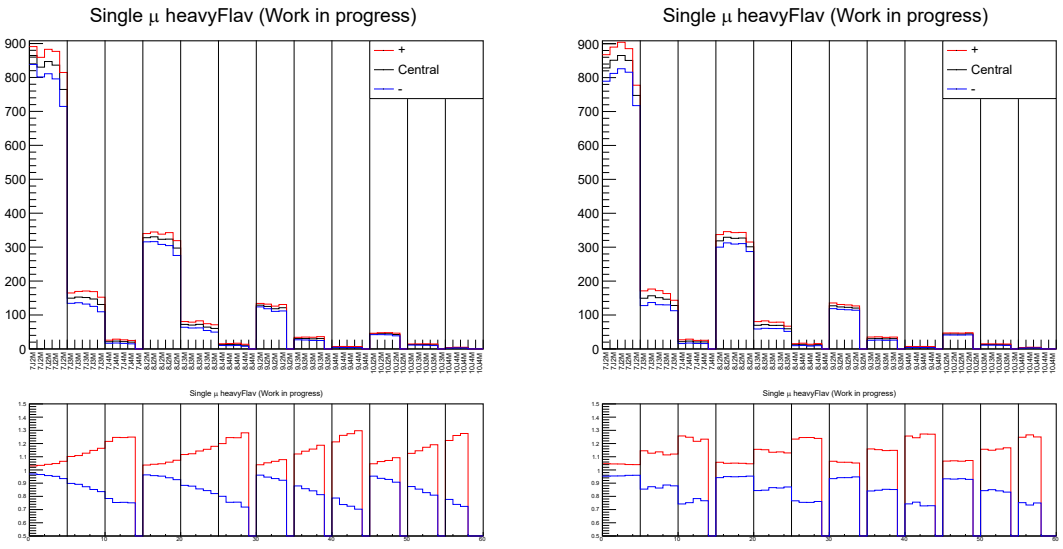
Figure 19: HeavyFlav uncertainty distributions from both types of neural network in single elctron channel.

The uncertainty distributions shown from both figures indicate that adversarial network training (with New loss variant) does its job: HeavyFlav uncertainty has been adjusted to be a flat uncertainty, almost having the same ratio, for each bin in a single histogram, leading to a reduction in the dependence of the shape of HeavyFlav uncertainty. With this modification introduced by adversarial network training, this may cause the data to be constrained more easily during the cross section expected limit calculation, as well as expected significance.

6.9 Expected limit and significance calculated for 200 fb^{-1} data

By applying the same speculation of data recording as described in Section 5.12, we may also calculate the expected limit and significance from both variants of adversary networks at 200 fb^{-1}.

Even with the uncertainty distribution modification introduced in the New loss variant of the adversary network, both variants of the adversarial neural network still do not improve the uncertainty of expected limits over the traditional neural network. The results shown in Table 9 and Figures 19 and 20 indicates one possibility: it is possible that reducing the dependence of the shape of HeavyFlav uncertainty alone does not help in reducing the uncertainty of expected limits and increasing expected significance. We will discuss possible studies that may be conducted in the future in the next chapter.



(a) with traditional NN

(b) with ANN (New loss variant)

Figure 20: HeavyFlav uncertainty distributions from both types of neural network in single muon channel.

Table 9: Expected limit and significance of four top quark production cross section, at 200 fb^{-1}, calculated from the output from both variants of adversarial neural networks (ANN), compared to sensitivities of traditional neural network (NN) and boosted decision tree (BDT).

	For 200.0 fb^{-1} data			
	BDT	NN	ANN (Standard)	ANN (New loss)
Expected limit (fb)	46^{+21}_{-14}	36^{+17}_{-11}	39^{+18}_{-12}	41^{+19}_{-12}
Expected significance	0.41	0.52	0.48	0.46

CHAPTER VII

CONCLUSION AND POSSIBLE FUTURE STUDIES

We have seen, throughout this thesis, that a traditional neural network alone, without any adversarial training, can already give better sensitivity than a boosted decision tree discriminator. This improvement by the traditional network without adversarial network training may be caused by the number of input variables used and the neural network's complexity. The results are more pronounced when we scale the amount of expected data up from the current integrated luminosity of 35.8 fb^{-1} to 200 fb^{-1}.

We may compare the expected cross section limit obtained with both traditional neural networks, with and without adversarial training, to the latest four top quark production search conducted by the ATLAS collaboration [30], as well as analyses in the past conducted by the CMS collaboration [12] [24] with $\sqrt{s} = 13$ TeV. From Table 10, we can see the expected limit on cross section derived from different integrated luminosity and different channels. As the data becomes more abundant, the uncertainty range for the expected limit becomes narrower. The inclusion of different final state channels, such as the dilepton channel, also reduces the uncertainty range significantly. (Unfortunately, the ATLAS 2018 search does not mention the uncertainty of its expected limit.) This thesis, however, only focuses on the single lepton channel, and the inclusion of dilepton channels to neural-network-based analysis, both same-sign and opposite-sign, is an interesting approach that can be done in the future.

We may also compare our speculation of expected sensitivity of the search for four top quark production to the projection of expected sensitivity achievable by High Luminosity LHC (HL-LHC) and High Energy LHC (HE-LHC) upgrades within the CMS detector. The search conducted in the HL-LHC projection [31] does not fully utilise any machine learning techniques in the search, and the expected significance is calculated with an integrated luminosity of 300 fb^{-1}, with $\sqrt{s} = 14$ TeV. The projection also takes three scenarios of systematic uncertainties evolution into account. While this comparison seems unfair, due to the centre-of-mass energy and more integrated luminosity, we may apply traditional neural networks for the search, at the same amount of centre-of-mass energy and integrated luminosity, to achieve better expected significance than this speculation.

Throughout this thesis, only the results from *expected* limit on cross section and *expected* significance are shown. The results are supposed to be accurate *if SM predictions*

Table 10: Expected limit on the cross section of four top quark production, compared with previous literature.

	Expected limit on cross section
This thesis 35.8 fb^{-1} data, $\sqrt{s} = 13$ TeV	
Single lepton NN	78^{+37}_{-24} fb
Single lepton ANN (Standard)	79^{+37}_{-24} fb
Single lepton ANN (New loss)	83^{+39}_{-25} fb
CMS 2015 search [24] 2.6 fb^{-1} data, $\sqrt{s} = 13$ TeV	
Single lepton BDT	151^{+90}_{-52} fb
Opposite-sign dilepton BDT	227^{+154}_{-84} fb
Single lepton + Opposite-sign dilepton BDT	118^{+76}_{-41} fb
CMS 2016 search [12] 35.8 fb^{-1} data, $\sqrt{s} = 13$ TeV	
Single lepton BDT	86^{+40}_{-26} fb
Dilepton BDT	67^{+41}_{-23} fb
Single lepton + Dilepton BDT	52^{+26}_{-17} fb
ATLAS 2018 search [30] 36.1 fb^{-1} data, $\sqrt{s} = 13$ TeV	
Single lepton + Dilepton	33 fb

Table 11: Expected significance projection for four top quark production search.

	Expected significance
This thesis 200 fb^{-1} data, $\sqrt{s} = 13$ TeV	
Boosted decision tree	0.41
Traditional neural network	0.52
Adversarial neural network (Standard)	0.48
Adversarial neural network (New loss)	0.46
HL-LHC 300 fb^{-1} data, $\sqrt{s} = 14$ TeV	
Systematic uncertainties unchanged	2.71
Best case scenario	2.93

are correct. To determine whether or not SM predictions are correct, however, we need to compare the expected limit to the *observed* cross section limit, and we can obtain the observed cross section limit only by unblinding the analysis, i.e. use the recorded data from the detector itself. Unblinding the analysis is also another interesting work that can also be done.

With the introduction of adversarial neural network training, the sensitivity slightly drops from the case where the traditional neural network is used. The results for adversarial network approach are also comparable for 200 fb^{-1} amount of data. Even with this slight drop, there is one variant of the adversarial network that has successfully modified the shape of only one systematic uncertainty distribution in a way that its shape may be ignored. With this modification and the sensitivity results shown, we may conclude that modifying the shape of one systematic uncertainty is not enough to lower the sensitivity impacts altogether, and focusing on only one systematic uncertainty may not give us better sensitivity in terms of expected limits and significances.

As stated in the introduction of this thesis, this is the first ever attempt to incorporate adversarial neural networks in a real-life LHC analysis. Due to these findings, I would like to propose three possible approaches to further tackle the sensitivity impacts problem using the same adversary network approach.

1. **Determine other systematic uncertainties that are also correlated** The work in this thesis tackles only one systematic uncertainty, with no regard to other systematic uncertainties that may also be correlated. One variant of the adversarial networks have successfully weakened one systematic uncertainty, but it is also possible that there might be other systematic uncertainties correlated to the one the adversarial network is trained against. Determining other uncertainties and training the adversary network against these systematic uncertainties can lower the impact altogether, which leads to the second approach.

2. **Design the adversary network to train on multiple uncertainties altogether.** We may create an adversarial neural network and train it to be a classifier, which classifies whether certain systematic uncertainties are present. We may *also* create an adversarial network to classify *multiple systematic uncertainties* at the same time, i.e. whether an input event contains any systematic uncertainties at all. For an optimistic approach, we may also train the adversarial network to be a *regressor* to assess the factor caused by several systematic uncertainties.

3. **Design the adversary network so that the uncertainty of discriminator output distribution decreases at signal-rich values (close to 1).** We have seen that we may design an adversarial network in such a way that a distribution from one systematic uncertainty becomes almost flat. To constrain the data further during expected limits calculation, we may cause the uncertainty distribution to become smaller at signal-rich values. With this way, a histogram bin containing high numbers of signal-like events will have smaller systematic uncertainty, causing the bin to be more constrained.

The results presented throughout this thesis is just the beginning, as this is the first attempt to use the adversarial neural network to pivot neural network discriminators in real-life LHC analysis. With two possible approaches proposed above, the outlook for this approach to be used in LHC analyses can be said to be fruitful, possibly leading us to more and more precise measurements within the LHC in the future.

Bibliography

[1] S. Chatrchyan et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. Phys. Lett. B, 716:30–61, 2012.

[2] F. Abe et al. Observation of top quark production in $\overline{p}p$ collisions with the collider detector at fermilab. Phys. Rev. Lett., 74:2626–2631, 14, Apr. 1995.

[3] M. Tanabashi et al. Review of particle physics. Phys. Rev. D, 98:030001, 3, Aug. 2018.

[4] M. Kobayashi and T. Maskawa. Cp-violation in the renormalizable theory of weak interaction. Prog. Theor. Phys., 49:652–657, 2, Feb. 1973.

[5] A. Quadt. Top quark physics at hadron colliders. Eur. Phys. J. C, 48(3):835–1000, Dec. 2006.

[6] A. M. Sirunyan et al. Cross section measurement of t-channel single top quark production in pp collisions at $\sqrt{s} = 13$ TeV. Phys. Lett. B, 772:752–776, 2017.

[7] A. M. Sirunyan et al. Measurement of the production cross section for single top quarks in association with W bosons in proton-proton collisions at $\sqrt{s} = 13$ TeV. J. High Energy Phys., 10:117, 2018.

[8] B. Yang and N. Liu. One-loop QCD correction to top pair production in the littlest Higgs model with T-parity at the LHC. EPL, 111(1):11001, 2015.

[9] S. Chatrchyan et al. Measurement of the $t\bar{t}$ production cross section and the top quark mass in the dilepton channel in pp collisions at $\sqrt{s} = 7$ TeV. J. High Energy Phys., 07:049, 2011.

[10] A. M. Sirunyan et al. Measurement of the $t\bar{t}$ production cross section using events with one lepton and at least one jet in pp collisions at $\sqrt{s} = 13$ TeV. J. High Energy Phys., 09:051, 2017.

[11] A. M. Sirunyan et al. Measurement of tt normalised multi-differential cross sections in pp collisions at $\sqrt{s} = 13$ TeV, and simultaneous determination of the strong coupling strength, top quark pole mass, and parton distribution functions. Submitted to: Eur. Phys. J. C, 2019.

[12] A. M. Sirunyan et al. Search for the production of four top quarks in the single-lepton and opposite-sign dilepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV. Submitted to: J. High Energy Phys., 2019.

[13] L. Beck, F. Blekman, and J. Goldstein. The Search for the Standard Model Production of Four Top Quarks, 2016.

[14] ATLAS Collaboration. Search for production of vector-like quark pairs and of four top quarks in the lepton plus jets final state in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. Technical report ATLAS-CONF-2015-012, CERN, Geneva, Mar. 2015.

[15] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, M. Backes, T. Carli, O. Cohen, and A. Christov. TMVA - Toolkit for Multivariate Data Analysis. arXiv e-prints:physics/0703039, physics/0703039, Mar. 2007.

[16] G. Louppe, M. Kagan, and K. Cranmer. Learning to pivot with adversarial networks. Adv. Neural Inf. Process Syst., 30:981–990, 2017. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors.

[17] Y. Jin, J. Zhang, M. Li, Y. Tian, H. Zhu, and Z. Fang. Towards the Automatic Anime Characters Creation with Generative Adversarial Networks. arXiv e-prints, arXiv:1708.05509, Aug. 2017.

[18] C. Shimmin, P. Sadowski, P. Baldi, E. Weik, D. Whiteson, E. Goul, and A. Søgaard. Decorrelated Jet Substructure Tagging using Adversarial Neural Networks. Phys. Rev. D, 96(7):074034, 2017.

[19] J.-L. Caron. CERN Aerial view.. Vue aérienne du CERN. AC Collection. Legacy of AC. Pictures from 1992 to 2002., June 1986.

[20] D. Barney. CMS Detector Slice. CMS Collection., Jan. 2016.

[21] A. Sirunyan et al. Particle-flow reconstruction and global event description with the CMS detector. J. Instrum., 12(10):P10003–P10003, Oct. 2017.

[22] CMS Collaboration. The CMS experiment at the CERN LHC. J. Instrum., 3(08):S08004–S08004, Aug. 2008.

[23] V. Khachatryan et al. Search for Standard Model Production of Four Top Quarks in the Lepton + Jets Channel in pp Collisions at $\sqrt{s} = 8$ TeV. J. High Energy Phys., 11:154, 2014.

[24] A. M. Sirunyan et al. Search for standard model production of four top quarks in proton-proton collisions at $\sqrt{s} = 13$ TeV. Phys. Lett. B, 772:336–358, 2017.

[25] M. Czakon, P. Fiedler, and A. Mitov. Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_S^4)$. Phys. Rev. Lett., 110:252004, 2013.

[26] G. Cowan, K. Cranmer, E. Gross, and O. Vitells. Asymptotic formulae for likelihood-based tests of new physics. Eur. Phys. J. C, 71:1554, 2011. [Erratum: Eur. Phys. J. C 73,2501(2013)].

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, and G. Louppe. Scikit-learn: Machine Learning in Python. arXiv e-prints, arXiv:1201.0490, Jan. 2012.

[28] F. Chollet et al. Keras, 2015.

[29] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv e-prints, arXiv:1502.03167, Feb. 2015.

[30] M. Aaboud et al. Search for four-top-quark production in the single-lepton and opposite-sign dilepton final states in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Phys. Rev. D, 99(5):052009, 2019.

[31] CMS Collaboration. Projections of sensitivities for tttt production at HL-LHC and HE-LHC. Technical report CMS-PAS-FTR-18-031, CERN, Geneva, 1900.

Biography

Vichayanun Wachirapusitanand was born in Bangkok, Thailand, on 17 July 1994. He first graduated from Mahidol Wittayanusorn School in 2012, and received BSc in Physics from Chulalongkorn University in 2017. His undergraduate thesis have been supervised by Assist Prof Dr Burin Asavapibhop. In 2017, he has been selected as a CERN Summer Student, under the supervision of Dr Martijn Mulders. His current interest is incorporating Machine Learning techniques into High Energy Physics experiments.

Publications:

1. V. Wachirapusitanand, N. Suwonjandee, B. Asavapibhop and N. Srimanobhas J. Phys. Conf. Ser. 1144:012031, 2018.