THE UNIVERSITY OF THE
WITWATERSRAND

DOCTORAL THESIS

The Use of Semi-Supervised Machine Learning Techniques in the Search for New Bosons with the ATLAS Detector

---

Author:
Benjamin LIEBERMAN

Supervisor:
Prof. Bruce MELLADO

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

School of Physics

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Declaration of Authorship

I, Benjamin LIEBERMAN, declare that this thesis titled, "The Use of Semi-Supervised Machine Learning Techniques in the Search for New Bosons with the ATLAS Detector" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 24 June 2024

Abstract

Since the completion of the Standard Model, with the discovery of the Higgs Boson, there has been a significant shift in the exploration of new physics to explain deviations between model simulated data and that produced at the Large Hadron Collider. These investigations are greatly aided by the integration of advanced machine learning techniques.

Machine learning methods offer powerful solutions for complex collider physics challenges. However, these models often depend on simulations that might not fully align with actual physical phenomena. In order to remove this model dependency, semi-supervised classifiers can be used. This solution, however, is not without challenges.

In this thesis, an evaluation of the use and limitations of semi-supervised classifiers in particle physics is presented. This is achieved by using a well constrained di-lepton final state dataset to assess the efficacy and ability of the technique, compared to its supervised counterpart. Following this baseline study, the $S(150) \rightarrow Z\gamma$ final state, motivated by the multi-lepton anomalies at the LHC, is used to perform narrow resonance searches with semi-supervised Neural Network (NN) classifiers. This work details the methodology and outcomes of a frequentist study aimed at quantifying the extent of a trials factor introduced by semi-supervised NN classifiers. This involves an in-depth analysis of the rate of statistical fluctuations generated in the training of semi-supervised techniques. A secondary component of this study is the evaluation of machine learning based data generators, with emphasis on the Wasserstein Generative Adversarial Network (WGAN), to produce large quantities of realistic data for physics analyses.

The results of this investigation into semi-supervised techniques, firstly validates the efficacy and ability of these techniques to classify particle events. This is followed by the frequentist study results, which substantiation that the trials factor remains suitably managed with the use of semi-supervised NN classifiers. The insights derived from this research pave the way for enhancing the reliability of upcoming resonance searches, underscoring the potential of semi-supervised learning in searches for narrow resonances.

Acknowledgements

First and foremost, I extend my deepest gratitude to Professor Mellado, whose unwavering support and guidance have been the cornerstone of my research journey. Prof. Mellado, your profound expertise and insightful mentorship have not only shaped this thesis but have also greatly enriched my academic and personal growth. The opportunities you provided and the knowledge you imparted have been invaluable, and I am immensely thankful for your steadfast belief in my potential.

I would like to extend my heartfelt appreciation to my colleagues, especially Phuti Rapheeha, Salah-Eddine Dahbi, Finn Stevenson, and Edward Nkadimeng, for their invaluable contributions. The rich collaborative atmosphere and intellectual camaraderie that we nurtured together have been fundamental to both my personal growth and the success of this research. Each of you, through your unique perspectives, insightful critiques, and unwavering encouragement, has played an instrumental role in refining my work and expanding my academic horizons. I am deeply grateful for the myriad of stimulating discussions, our shared experiences in overcoming challenges, and the joy of celebrating our collective achievements. These interactions have not only enriched my research journey but have also left an indelible mark on my journey as a scholar.

Finally, I must acknowledge the unwavering support and love of my family. Your belief in my aspirations, your endless patience, and your comforting presence have been my constant source of strength and motivation. To my family, thank you for being my pillars of support and for the sacrifices you have made in supporting my academic pursuits. Your unwavering faith in me has been the driving force behind my endeavours, and this achievement is as much yours as it is mine.

Publications

Journal Articles

- Hernandez, Y., Kumar, M., Cornell, A.S., Dahbi, S.-E., Fang, Y., Lieberman, B., Mellado, B., Monnakgotla, K., Ruan, X., Xin, S., "The anomalous production of multi-leptons and its impact on the measurement of WH Production at the LHC". The European Physical Journal C, vol. 81, no. 4, p. 365 (2021). DOI: https://doi.org/10.1140/epjc/s10052-021-09137-1.

- Dahbi, S.-E., Choma, J., Mokgatitswane, G., Ruan, X., Lieberman, B., Mellado, B., Celik, T., "Machine Learning Approach for the search of resonances with topological features at the large hadron collider". International Journal of Modern Physics A, vol. 37, no. 3, p. 2150241 (2022). DOI: https://doi.org/10.1142/S0217751X21502419.

- Lieberman, B., Crivellin, A., Dahbi, S.-E., Stevenson, F., Tripathi N., Kumar M., Mellado, B., "Trials Factor for Semi-Supervised NN Classifiers in Searches for Narrow Resonances at the LHC". 2024. Under Review.

- Lieberman, B., Kong, J.D., Gusinow, R., Asgary, A., Bragazzi, N.L., Choma, J., Dahbi, S.-E., Hayashi, K., Kar, D., Kawonga, M., Mbada, M., Monnakgotla, K., Orbinski, J., Ruan, X., Stevenson, F., Wu, J., Mellado, B., "Big data- and artificial intelligence-based hot-spot analysis of covid-19: Gauteng, South Africa, as a case study". BMC Medical Informatics and Decision Making 23, 19 (2023). DOI: https://doi.org/10.1186/s12911-023-02098-3.

- Stevenson, F., Hayasi, K., Bragazzi, N., Kong, J.D., Asgary, A., Lieberman, B., Ruan, X., Mathaha, T., Dahbi, S.-E., Choma, J., Kawonga, M., Mbada, M., Tripathi, N., Orbinski, J., Mellado, B., and Wu, J., 2021. "Development of an early alert system for an additional wave of covid-19 cases using a recurrent neural network with long short-term memory". International Journal of Environmental Research and Public Health 18, 7376. https://doi.org/10.3390/ijerph18147376.

- Mellado, B., Wu, J., Kong, J.D., Bragazzi, N.L., Asgary, A., Kawonga, M., Choma, N., Hayasi, K., Lieberman, B., Mathaha, T., Mbada, M., Ruan, X., Stevenson, F., Orbinski, J., 2021. "Leveraging Artificial Intelligence and big data to optimize COVID-19 Clinical Public Health and vaccination roll-out strategies in Africa". International Journal of Environmental Research and Public Health 18, 7890. https://doi.org/10.3390/ijerph18157890.

- Alavinejad, M., Mellado, B., Asgary, A., Lieberman, B., Mbada, M., Mathaha, T., Stevenson, F., Tripathi, N., Swain, A.K., Orbinski, J., Wu, J., Kong, J.D., 2022. "Management of hospital beds and ventilators in the Gauteng province, South Africa, during the COVID-19 pandemic". PLOS Global Public Health. https://doi.org/10.1371/journal.pgph.0001113.

Conference Proceedings

- Lieberman, B., Choma, J., Dahbi, S.-E., Mellado, B., Ruan, X. "An investigation of overtraining within semi-supervised machine learning models in the search for heavy resonances at the LHC", in The Proceedings of SAIP2021, the 65th Annual Conference of the South African Institute of Physics, July 2021. ISBN: 978 − 0 − 620 − 97693 − 0. Available online at http://events.saip.org.za.

- Lieberman B., Dahbi, S.-E., Ruan, X., Stevenson, F., Mellado, B., "A frequentist study of the false signals generated in the training of semi-supervised neural network classifiers using a WGAN as a data generator", in The Proceedings of SAIP2022, the 66th Annual Conference of the South African Institute of Physics, July 2022. ISBN: 978 − 0 − 6397 − 4426 − 1. Available online at http://events.saip.org.za.

- Lieberman B., Dahbi S.-E., Mellado, B. "The use of Wasserstein Generative Adversarial Networks in searches for new resonances at the LHC". International Nuclear Physics Conference, September 2022. Proceedings: Journal of Physics Conference Series. September 2023. DOI: 10.1088/1742 − 6596/2586/1/012157. Available online at https://dx.doi.org/10.1088/1742-6596/2586/1/012157.

- Stevenson, F., Lieberman, B., Swain, A. and Mellado, B., "Evaluation and Optimisation of a Generative-Classification Hybrid Variational

Autoencoder in the Search for Resonances at the LHC". International Nuclear Physics Conference, September 2022. Proceedings: Journal of Physics Conference Series. September 2023. DOI: 10.1088/1742 − 6596/2586/1/012160. Available online at https://dx.doi.org/10.1088/1742-6596/2586/1/012160.

List of Abbreviations

2HDM	**T**wo-**H**iggs-**D**oublet **M**odel
2HDM+S	**T**wo-**H**iggs-**D**oublet **M**odel plus **S**inglet
Adam	**Ada**ptive **M**oment Estimation
ALICE	**A** **L**arge **I**on **C**ollider **E**xperiment
ANN	**A**rtificial **N**eural **N**etwork
ATLAS	**A** **T**oroidal **L**HC **A**pparatu**S**
AUC	**A**rea **U**nder the **C**urve
BCE	**B**inary **C**ross-**E**ntropy
BDT	**B**oosted **D**ecision **T**ree
BEH	**B**rout-**E**nglert-**H**iggs
BR	**B**ackground **R**ejection
BSM	**B**eyond the **S**tandard **M**odel
CDF	**C**umulative **D**istribution **F**unction
CERN	**E**uropean **O**rganization for **N**uclear **R**esearch
CMS	**C**ompact **M**uon **S**olenoid
CP	**C**harge **P**arity
CPU	**C**entral **P**rocessing **U**nit
CR	**C**ontrol **R**egion
DNN	**D**eep **N**eural **N**etwork
EM	**E**lectro**M**agnetic
EW	**E**lectro**W**eak
FCal	**F**orward **Cal**orimeter
FCNC	**F**lavour **C**hanging **N**eutral **C**urrents
GAN	**G**enerative **A**dversarial **N**etwork
GMM	**G**aussian **M**ixture **M**odels
GPU	**G**raphics **P**rocessing **U**nit
HEP	**H**igh **E**nergy **P**hysics
ID	**I**nner **D**etector
KDE	**K**ernel **D**ensity **E**stimation
KS	**K**olmogorov-**S**mirnov
LEE	**L**ook **e**lsewhere **E**ffect
LHC	**L**arge **H**adron **C**ollider
LHCb	**L**arge **H**adron **C**ollider **b**eauty
Linac 4	**L**inear **A**ccelerator 4
LAr	**L**iquid **A**rgon

MC	**M**onte **C**arlo
ML	**M**achine **L**earning
MLP	**M**ulti-**L**ayer **P**erceptron
MSE	**M**ean **S**quare **E**rror
NN	**N**eural **N**etwork
NPI	**N**on-**P**harmaceutical **I**nterventions
PDF	**P**robability **D**ensity **F**unction
PS	**P**roton **S**ynchrotron
PSB	**P**roton **S**ynchrotron **B**ooster
ReLU	**Re**ctified **L**inear **U**nit
RI	**R**isk **I**ndex
RMS	**R**oot **M**ean **S**quared
ROC	**R**eceiver **O**perating **C**haracteristic
SCT	**S**emi**C**onductor **T**racker
SGD	**S**tochastic **G**radient **D**escent
SIRD	**S**usceptible-**I**nfectious-**R**ecovered-**D**eath
SM	**S**tandard **M**odel
SPS	**S**uper **P**roton **S**ynchrotron
SSB	**S**ontaneous **S**ymmetry **B**reaking
TDAQ	**T**rigger and **D**ata Ac**Q**uisition system
TMVA	**T**oolkit for **M**ulti**V**ariate Data **A**nalysis
TRT	**T**ransition **R**adiation **T**racker
QCD	**Q**uantum **C**hromo**D**ynamics
QFT	**Q**uantum **F**ield **T**heory
VAE	**V**ariational **A**uto**E**ncoder
VAE+D	**V**ariational **A**uto**E**ncoder plus **D**iscriminator
vev	**v**acuum **e**xpectation **v**alue
WGAN	**W**asserstein **G**enerative **A**dversarial **N**etwork

Contents

List of Figures

List of Tables

In loving memory of my grandfather, Robert Lieberman, who asked for so little but gave so very much.

Chapter 1

Introduction to the Thesis

In the history of humanity, few endeavours have been as transformative as the pursuit of scientific discovery. Over the last century, advancements in science have fundamentally reshaped our understanding of the universe and catalysed societal progress on an unprecedented scale, from innovations in medicine and technology to a deeper understanding of the natural world. Among these scientific disciplines, High-Energy Physics (HEP) stands as a monumental testament to human curiosity and ingenuity. It peels back the layers of our universe to examine its most fundamental constituents, offering revolutionary insights that often blur the lines between the abstract and the empirically tangible. Within HEP, two methodological paradigms, bottom-up and top-down, serve as twin pillars in the harmonious relationship between theory and experiment. Historically, the field was predominantly conducted using a top-down approach. Theoretical constructs were erected first, often driven by mathematical elegance or overarching principles, and their empirical ramifications were then explored through meticulously designed experiments. This traditional methodology was instrumental in laying the groundwork for major breakthroughs such as the prediction and eventual discovery of elementary particles. However, the landscape of HEP is evolving. Today, the relationship between theory and experiment is increasingly bidirectional, incorporating elements of the bottom-up methodology as well. In this approach, researchers draw upon raw experimental data to refine existing theories or, in some cases, to spawn entirely new theoretical frameworks. The integration of these bottom-up

insights with traditional top-down theories creates a more holistic, agile scientific process, allowing for a richer, more nuanced understanding of the universe's fundamental makeup. In recent years, the advent of Machine Learning (ML) technologies has added a new dimension to this complex interplay. By automating data analysis and identifying intricate patterns beyond human capability, ML serves as both a mechanism for discovery and a bridge between theory and experiment, heralding a new era in our ongoing quest to decode the intricacies of the universe.

As humanities knowledge and understandings of high energy physics have grown, so too have the size and complexity of the experiments designed to probe its mysteries. This escalation is most conspicuously exemplified by the Large Hadron Collider (LHC) at CERN, the world's largest and most powerful particle accelerator. However, the scale of the LHC is not just an engineering feat; it's a necessity dictated by the increasingly intricate questions we seek to answer. With the capability to collide particles at near-light speed, the LHC generates an immense amount of data, petabytes per second, that holds the potential for groundbreaking discoveries. This deluge of data presents both an opportunity and a challenge; while the vast amount of information increases the odds of detecting rare events and novel phenomena, it also renders some traditional data analysis techniques insufficient. In this context, ML has emerged as an indispensable tool. Not only is it able to automate the arduous task of sifting through colossal data sets, but it also identifies complex, multidimensional patterns that are beyond the intuitive grasp of even the most experienced physicists. By doing so, ML technologies amplify the efficacy of our experimental apparatus, ensuring that theory and experiment continue to evolve in tandem as we deepen our pursuit to understand the universe at its most fundamental level.

New research conducted using ML methodologies to enhance traditional analyses are faced with new limitations. One such limitation, that is pivotal in particle physics discovery, is ML algorithms dependency on the simulated dataset on which it is trained. The particle physics models used to generate these datasets are structured and therefore limited by our current understandings of particles and their interactions. This means that the extent of ML algorithms ability to understand these particles are limited by the

model informed dataset. In order to discover new physics using ML, methods must be considered to reduce model dependency and therefore extract new phenomena which do not necessarily conform to current models of particle physics. One such method that has the potential to classify HEP data while reducing model dependency is the semi-supervised ML technique.

When conducting any resonance search, it is crucial to take into account the look-elsewhere effect to arrive at a statistically meaningful result, i.e. significance for an excess. This effect refers to the increased probability of finding apparent signals (that are actually statistical fluctuations) when searching over an extended parameter range, usually mass. When using semi-supervised classifiers, an additional look-elsewhere effect is introduced.
This is due to the fact that the signal sample, used for training, is not predefined like in the fully supervised method and a scan must be performed on the response distribution to determine the optimal threshold cut. This scan of the response distribution introduces an additional look-elsewhere effect that can manifest itself as "fake signals" which can significantly distort resonance patterns and influence their interpretations [1]. Hence, it is imperative to evaluate the probability of fake signals produced in the implementation of semi-supervised classifiers and expose it in the form of a trials factor.

The primary aim of this thesis is to provide a comprehensive investigation into the semi-supervised ML techniques to extract unknown particle signatures from known background processes. This includes an evaluation of the efficacy of semi-supervised techniques to classify data produced at the LHC, compared to more standardised, fully supervised, techniques. It also includes a comparison of the ability of different machine learning models to classify data using semi-supervised techniques. The thesis concludes with a novel analysis aimed at measuring the extent of an additional look-elsewhere effect, in the form of fake signals, introduced when using these techniques. This study focuses on exposing new physics using the 2HDM+S model.

In addition, this thesis explores the potential of ML based data generation

techniques, such as the Generative Adversarial Network, to generate realistic particle physics events for analyses. These techniques offer the potential of reducing time and CPU load when generating large dataset samples for HEP studies.

Finally, this thesis was conducted during a period of unprecedented challenges brought forth by the COVID-19 pandemic. As the severity of the crisis deepened, both provincial and national governments, especially within South Africa, urgently sought data-driven solutions to effectively combat the virus's rapid spread and its cascading impacts. To this end, extended research was applied to harness and adapt the intricate scientific methodologies, traditionally employed in subatomic particle physics, as a novel means to tackle the pandemic. By amalgamating these scientific techniques with epidemiological modelling and the power of machine learning, this research provided invaluable insights to inform policy decisions, thereby aiming to alleviate the immense strain on our healthcare systems and chart a data-informed course through the turbulent times.

The structure of this thesis is laid out as follows. Firstly, Chapter 2 lays the groundwork by introducing the Standard Model (SM) of particle physics and delving into Beyond the Standard Model (BSM) studies, thereby introducing and motivating the foundational theoretical model for the research. Chapter 3 takes readers on an exploration of the ATLAS experiment, the LHC, and CERN, which form the cornerstone of the experimental data utilised in this work. In Chapter 4, we pivot towards the domain of artificial intelligence, specifically introducing and elaborating on the machine learning techniques and models that underpin this research. In Chapter 5, the di-lepton dataset is introduced before being used to compare the fully supervised and semi-supervised learning techniques. This chapter offers a preliminary evaluation of semi-supervised classifiers by comparing different models with their fully supervised counterparts. The models include the Boosted Decision Trees (BDT), Multi-Layer Perceptrons (MLP), and Deep Neural Networks (DNN). Thereafter, Chapter 6 presents the selected $Z\gamma$ final state dataset before introducing a novel study methodology, and results that expose additional look elsewhere effects introduced when using

semi-supervised techniques to classify HEP data. Finally, Chapter 7 provides a conclusion, drawing threads from the preceding chapters to weave a cohesive narrative. In addition, two appendices enrich the discourse: Appendix A offers an extended evaluation of data generation methods anchored in machine learning, while Appendix B focuses on the application of scientific methods for the COVID-19 pandemic in South Africa.

Chapter 2

Particle Physics Introduction

2.1 The Standard Model of Particle Physics

Since the 1930s, the collaborative efforts and groundbreaking discoveries of numerous physicists have yielded profound insights into the nature and interactions of fundamental particles. These particles, which serve as the elemental constituents of matter, are subject to the influence of four fundamental forces that govern their behaviour. Emerging in the early 1970s, the Standard Model (SM) stands as a testament to human intellect's ability to synthesise intricate understandings of these particles and three of the fundamental forces that shape their interactions [2]. The SM is written in the language of quantum field theory (QFT), as matter at a fundamental level is made up of fields. The culmination of this intellectual journey was realised in 2012 with the pivotal detection of the Higgs boson, achieved through the concerted endeavours of the ATLAS [3] and CMS [4] collaborations. The SM was completed with the discovery of the Higgs boson, and can be described in the form of groupings of fundamental particles, as shown in Figure 2.1.

Every constituent particle encompassed within the prevailing SM has undergone rigorous scrutiny and validation across a wide spectrum of experiments and observations. This section will firstly outline the theoretical structure of the SM, followed by an introduction of BSM studies. The experimental apparatus used to observe these particles is described in Section 3.

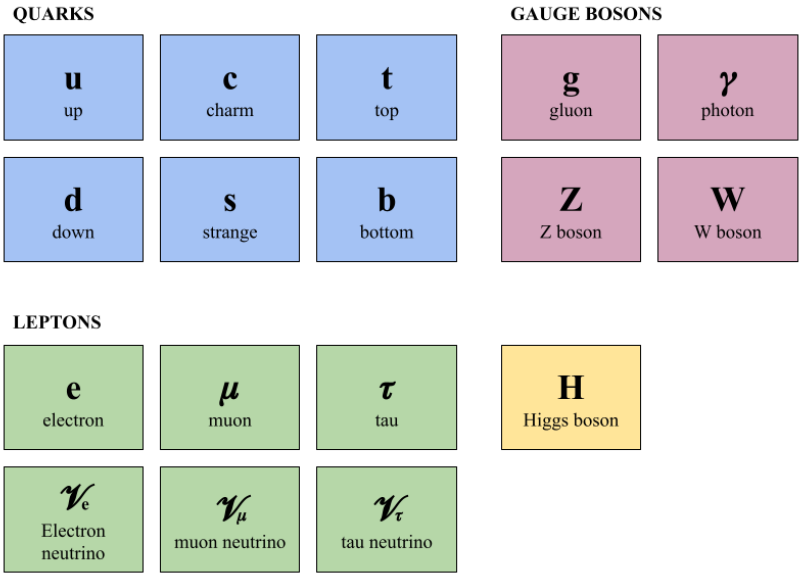


FIGURE 2.1: Overview of the Standard Model particles: Showcasing fermions (quarks and leptons), gauge bosons (force carriers), and the Higgs boson. This diagram encapsulates the three generations of matter and fundamental forces in particle physics.

2.1.1 Particles in the Standard Model

In the SM, particles are broadly classified into two main categories: fermions and bosons. Fermions, representing matter particles, and bosons, denoting force carriers, are distinguished not only by their spin but also by their interactions and how they organise into larger groups.

Matter particles

The matter that makes up the physical world consists of atoms. Each atom consists of a nucleus surrounded by one or more electron. The atomic nucleus consists of a combination of protons and neutrons. Both protons and neutrons fall under the category of baryons and are composed of quarks. More specifically, a proton is made up of two up quarks and a single down quark, while a neutron is composed of two down quarks and a single up quark. All matter can therefore be broken down into a combination of up

and down quarks, together with leptons in the form of electrons and cosmic neutrinos.

Fermions, often referred to as matter particles, manifest in two fundamental varieties: quarks and leptons. Characteristically, fermions possess half-integer spins ($\frac{1}{2}$, $\frac{3}{2}$, etc), adhering to the Pauli exclusion principle, and exhibit antisymmetric wave-functions in quantum mechanics. This implies that two fermionic particles cannot concurrently occupy the same quantum state.

Fermions can be divided into three generations which differ in mass and flavour quantum number, yet have identical electric and strong interactions. Early generation fermions are lighter and more stable, where later generations are progressively heavier and more unstable. Therefore, the unstable heavier particles, in the higher generation, decay into first generation particles to become stable, which makes up everyday matter (electrons orbiting bound states of quarks or more simply protons and neutrons).

Quarks obey the "confinement" principle and naturally group together to form hadrons (protons, neutrons, pions, etc.). The SM quarks are exposed in terms of their type, generation and associated mass, in Table 2.1. The *up-type* and *down-type* quarks have respective charges of $\frac{2}{3}$ and $-\frac{1}{3}$.

	Quarks		
	First generation	Second generation	Third generation
Up-type	Up quark (u) $2.3^{+0.7}_{-0.5}$ MeV	Charm quark (c) 1.275 ± 0.025 GeV	Top quark (t) 173.21 ± 0.87 GeV
Down-type	Down quark (d) $4.8^{+0.5}_{-0.3}$ MeV	Strange quark (s) 95 ± 5 MeV	Bottom quark (b) 4.66 ± 0.03 GeV

TABLE 2.1: The Standard Model quarks and their associated masses organised by generation. Quoted masses are expanded upon in Ref. [5].

Leptons, which do not participate in the strong nuclear force, include particles such as electrons and neutrinos. The SM leptons are shown in terms of their type, generation and associated mass, in Table 2.2. The *charged* leptons have a charge of -1 while the *neutral* leptons have no charge.

	Leptons		
	First generation	Second generation	Third generation
Charge	Electron (e) 0.511 MeV	Muon (μ) 105.66 MeV	Tau (τ) 1776.86 ± 0.12 MeV
Neutral	Electron neutrino (v_e) < 2 eV	Muon neutrino (v_μ) < 2 eV	Tau neutrino (v_τ) < 2 eV

TABLE 2.2: The Standard Model leptons and their associated masses organised by generation. Quoted masses are expanded upon in Ref. [5].

Each particle introduced is paired with a corresponding anti-particle, which shares the same mass but has an opposite charge. Thus, for every fermion there exists an anti-fermion with identical mass and inverse charge.

Force-carrier particles

In our universe, there exist four cardinal forces that govern the interactions of matter and energy. These forces are:

1. **The gravitational force:** The force acting between any two objects with mass. It is the weakest of the forces but is able to act over an infinite range.
2. **The electromagnetic force:** This force governs the attraction of opposite electric charges and the repulsion of similar charges. Significantly stronger than gravity but weaker than the strong nuclear force, it underpins the chemical behaviour of matter and the properties of light.
3. **The strong nuclear force:** The fundamental force that occurs between quarks, binding them together to form protons, neutrons, and other hadronic particles. This is the strongest of the forces and acts upon a short-range, on the order of a nucleon's diameter. The residual effects extend further to bind protons and neutrons together, in the atomic nucleus, opposing the strong repulsion of positively charged protons.
4. **The weak nuclear force:** The fundamental force that is responsible for radioactive decay and participates in nuclear fission and fusion. It is

only stronger in magnitude than the gravitational force and has a very short range, smaller than the diameter of a typical nucleus.

These fundamental forces result from an exchange in energy of force-carrier particles, bosons. The transfer of discrete amounts of energy between particles, is achieved through the exchange of bosons. The SM bosons are therefore interaction mediators. Bosons, unlike fermions, have integer spin (0, 1, 2, etc.) and do not obey the Pauli exclusion principle. Therefore, bosons have quantum mechanical wave functions that are symmetrical and are able to form condensates.

For each of the fundamental forces there is a corresponding boson, carrier particle. These interaction mediators are in the form of spin 1 bosons, that are exposed through specific symmetries that must be obeyed by nature. These symmetries are known as gauge symmetries, leading to the interaction mediators to be referred to as gauge bosons. In the realm of fundamental forces, the electromagnetic and strong nuclear interactions are mediated by the massless photon, γ, and gluon, g, particles, respectively. While the weak nuclear interaction is transmitted through the massive $W^{\pm}$ and Z gauge bosons, which have measured masses of $80.385 \pm 0.015\,\text{GeV}$ and $91.1876 \pm 0.0021\,\text{GeV}$, respectively [5]. Gravity, as described by general relativity, is postulated to be conveyed by a spin-2 gauge boson termed the graviton. Nonetheless, due to gravity's weak magnitude on microscopic scales and its distinct theoretical description, integrating it into the quantum framework remains a challenge. To date, no experimental data has definitively confirmed the existence of the graviton. Table 2.3 provides an overview of the fundamental forces and their corresponding bosons.

The Higgs Boson

The SM is said to be complete with the addition of the massive Higgs boson particle. The Higgs boson is a spin-0 scalar particle with a mass of $125.09 \pm 0.24\,\text{GeV}$ [6]. Unlike gauge bosons, the Higgs interaction is not mediated by a gauge field. The existence of the Higgs boson (and the masses of all known massive elementary particles) was first suggested by Glashow, Weinberg, and Salam, in their theoretical descriptions of the electroweak

Gauge bosons

Force	Electromagnetic	Strong nuclear	Weak nuclear
Bosons	Photon (γ)	Gluon (g)	$W^{\pm}$ boson $80.385 \pm 0.015\,\text{GeV}$
			Z boson $91.1876 \pm 0.0021\,\text{GeV}$

TABLE 2.3: The Standard Model gauge bosons, the fundamental force they mediate and corresponding masses. Quoted masses are expanded upon in Ref. [5].

symmetry breaking [7–9]. These theories led to Peter Higgs theorising the Brout-Englert-Higgs Mechanism, that breaks electroweak symmetry in the SM, giving mass to particles through couplings, and therefore predicting the Higgs boson [10, 11]. The Higgs boson was experimentally discovered in 2012 by the ATLAS and CMS Collaboration [3, 4].

According to the SM, the Higgs boson, h, can decay into pairs of fermions or bosons. The mass of the Higgs boson is however not predicted by the SM, but the production cross-section and branching ratios can be calculated with high precision using the Higgs measurement. The Higgs mass, m_h, is therefore a free parameter within the SM. The probability of the Higgs boson decaying into a given particle, is expressed as a branching ratio. The branching ratios and total uncertainty of the Higgs boson are shown in Figure 2.2.

The branching ratios shown in Figure 2.2, demonstrate that the decay modes change in prominence depending on the Higgs mass, m_h. Therefore, the diagram exposes that the probability of the Higgs decaying into a certain particle or group of particles, is dependent on the mass of the Higgs. Current experimental measurements indicate that the quantum numbers and relativistic branching ratio to SM particles are consistent with SM predictions [13, 14].

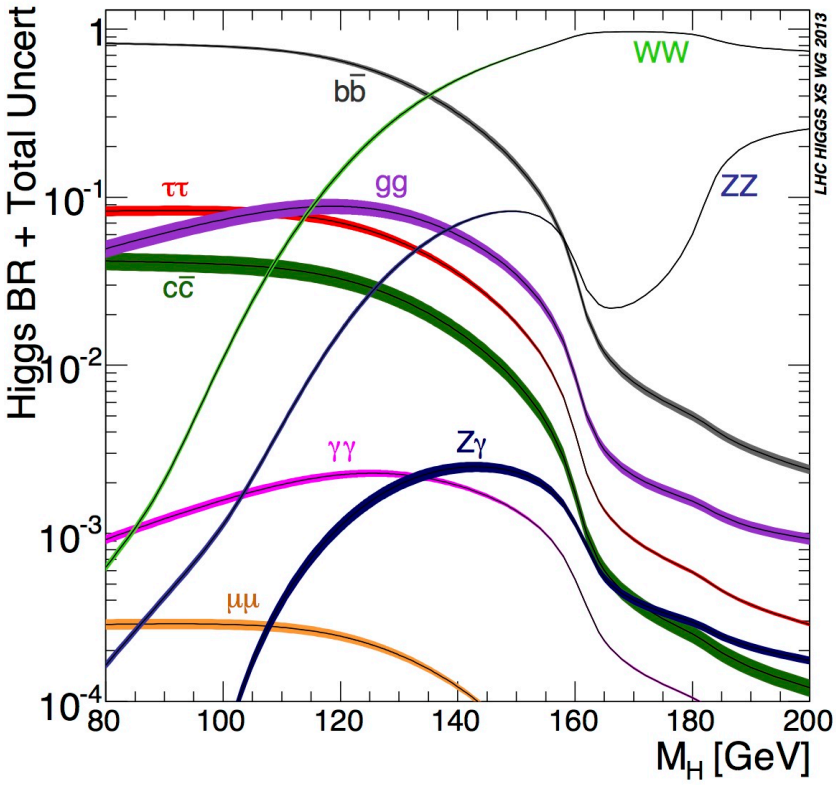


FIGURE 2.2: Standard Model Higgs boson decay branching ratios and total width [12].

2.1.2 The Standard Model Lagrangian

The SM can be described using group theory notation [15], encapsulating it as a gauge quantum theory underpinned by the internal symmetries of the unitary product groups. Specifically, the strong interactions are governed by the third-order special unitary group, SU(3), while the electroweak forces are described by the combined groups of SU(2) and U(1). Within this framework, U(n) represents an n^{th}-order unitary group, comprising all n by n unitary matrices, and its subset, SU(n), includes those unitary matrices with determinant one. Theories constructed around the SU(n) groups fall under the category of Yang-Mills theories [16].

The mathematical framework of the Standard Model (SM) is based on gauge quantum field theory, characterized by the internal symmetries of the unitary product group SU(3) x SU(2) x U(1). In order to build an equation to describe the SM, the equations of motion of particles can be described using

the Lagrangian of classical mechanics. The Lagrangian, L, describing the system dynamics of a particle in spatial dimension x, is defined as:

$$L = T - V, \tag{2.1}$$

where T is the kinetic energy of the particle and V is the potential energy that the particle is moving within. By integrating the Lagrangian of a particle with respect to time, the *action*, S, of the particle can be calculated as:

$$S = \int L\, dt. \tag{2.2}$$

The basic formulation of the Lagrangian, Equations 2.1 and 2.2, must be applied to fields. Considering the general field, $\Phi(x^\mu)$, as a function of space-time coordinates, $x^\mu = (x^1, x^2, x^3, x^4)$, the Lagrangian density, $\mathcal{L}$, is a function dependent on the fields and their derivatives. The Lagrangian density can be written as the sum of a polynomial, v:

$$\mathcal{L}(\Phi_i(x^\mu), \partial_\mu \Phi_i(x^\mu)) = \sum c_k v_k(\Phi_i(x^\mu), \partial_\mu \Phi_i(x^\mu)), \tag{2.3}$$

where $\Phi(x^\mu)$ and $\partial_\mu \Phi(x^\mu)$ are the general fields and their derivatives respectively for $i = 1, 2, 3, 4$. c_k are dimensional coefficients, $\frac{1}{\Lambda^D}$, where Λ is an energy scale and $D > 0$. The general fields, Φ, can be made up of a scalar field, ϕ, a vector field, ψ, or both. When considering the particles of the SM, a generalised version of the SM Lagrangian, $\mathcal{L}_{SM}$, can be written as:

$$\mathcal{L}_{SM} = \mathcal{L}_G + \mathcal{L}_F + \mathcal{L}_Y + \mathcal{L}_H. \tag{2.4}$$

The components of this Lagrangian are defined as:

$$\mathcal{L}_G = -\frac{1}{4} F_{\mu v} F^{\mu v}, \tag{2.5}$$

$$\mathcal{L}_F = i\bar{\psi} \not{D} \psi + h.c., \tag{2.6}$$

$$\mathcal{L}_Y = \bar{\psi}_i y_i \psi_j \phi + h.c., \tag{2.7}$$

$$\mathcal{L}_H = -|D_\mu \phi|^2 - V(\phi), \tag{2.8}$$

where the terms, $\mathcal{L}_G$ represents the couplings to gauge bosons, $\mathcal{L}_F$ represents the couplings to the fermions, $\mathcal{L}_Y$ reflects the Yukawa couplings and $\mathcal{L}_H$ represents the couplings to the Higgs boson. The *h.c.* stands for the Hermitian conjugate, ensuring that the Lagrangian remains real and thus physically meaningful. In fermionic couplings, the $\not{D}$ denotes the covariant derivative that incorporates the interactions with the gauge fields, and ψ represents the fermion fields. The $\mathcal{L}_Y$ term describes the interaction between the fermions and the Higgs field. The term represents the Yukawa coupling where ϕ is the Higgs field, and y_i are the Yukawa coupling constants. The Higgs field Lagrangian term describes the dynamics of the Higgs field, including its kinetic, $|D_\mu \phi|^2$, and potential, $V(\phi)$, energy.

Chirality

Chirality is a property of particles that refers to the "handedness" or orientation of their spin relative to their direction of motion. In the context of the Standard Model Lagrangian, chirality is important as it determines the type of interactions that particles can participate in. The Lagrangian includes terms that describe both left-handed (negative chirality) and right-handed (positive chirality) particles, with different interactions for each. The fermionic term of the Lagrangian, $\mathcal{L}_F$, includes fermionic fields, ψ, that can be decomposed into left-handed and right-handed components. These components are treated differently in the SM. For instance, in the weak interactions, only left-handed fermions and right-handed anti-fermions participate. This is a manifestation of the chiral nature of the weak force, which violates parity symmetry. The Yukawa interactions involve the Higgs field, ϕ, coupling to both left-handed and right-handed fermions. However, the coupling strengths (Yukawa couplings y_i) can be different for different chirality states. This difference is fundamental in giving masses to fermions once the Higgs field acquires a vacuum expectation value, a process intimately tied

to the chirality of the fermions. The gauge part of the Lagrangian, particularly for the weak interactions, also reflects chirality. The weak force carriers (W and Z bosons) couple only to left-handed particles and right-handed antiparticles. This is encoded in the covariant derivative, $\not{D}$, in the fermionic Lagrangian.

2.1.3 Quantum chromodynamics

The theory of quantum chromodynamics (QCD), describes the actions of the strong nuclear force. QCD therefore describes the strong interaction between quarks mediated by gluons. The strong force can be described using the $SU(3)_C$ gauge group, where C represents the colour charge. The gluons are the force carriers, and transmit the strong force between colour carrying particles, quarks. QCD has three salient properties, namely colour confinement, asymptotic freedom and chiral symmetry breaking. A comprehensive description of QCD can be found in Ref. [17].

Colour confinement

Colour charged particles cannot be directly detected as they cannot be isolated. This is explained by the first important QCD property, colour confinement, which occurs at low energies. This property exposes the fact that the force between a quark-antiquark pair remains constant when they are separated, while the potential energy grows linearly. The result of this is that the energy in the gluon field increases until a new quark-antiquark pair is spontaneously produced and hadrons are formed, shown in Figure 2.3. Hadrons do not have an associated colour charge, and are categorised into two main types, the mesons (one quark, one antiquark) and the baryons (three quarks).

Asymptotic freedom

A second important property of QCD, is asymptotic freedom. The asymptotic freedom occurs at high energies, and is the phenomenon of the quarks and gluons interacting very weakly, creating quark-gluon plasma. This is due to the fact that as the energy scale of the interaction is increased, the

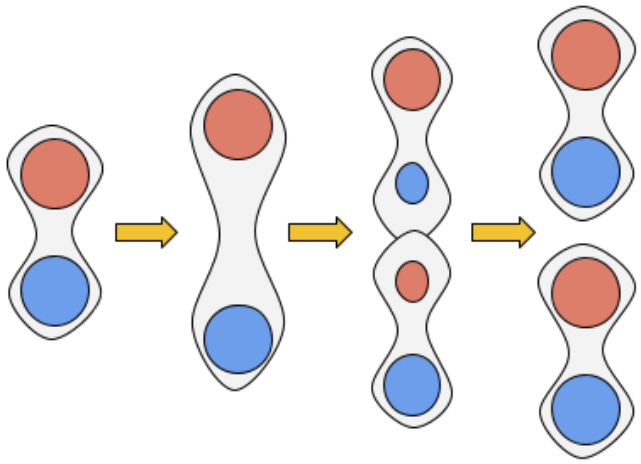


FIGURE 2.3: A simplified visual representation of quark colour confinement.

strength of the interactions between quarks and gluons decreases. When the energy scale is high enough, the particles no longer interact via the strong force and are said to have reached asymptotic freedom.

2.1.4 Electroweak theory

Electroweak (EW) theory was introduced by Glashow, Weinberg and Salam, in order to understand weak interactions [7–9]. The theory unified the descriptions of the electromagnetic and weak nuclear fundamental forces into a single Yang-Mills theory. In EW theory, left-handed fermions are grouped into weak iso-doublets, while right-handed fermions are treated as singlets. Studies on β decay have demonstrated that only left-handed fermions participate in charged current weak interactions [18]. This indicates that right-handed fermions remain invariant under the EW gauge transformations. The iso-doublets of the left-handed fermions can be expressed as:

$$q_L = \begin{pmatrix} u \\ d \end{pmatrix}_L, \; \ell_L = \begin{pmatrix} \nu_\ell \\ \ell^- \end{pmatrix}_L, \tag{2.9}$$

where the left-handed quark, q_L, and lepton, ℓ_L, have iso-doublets of up- and down-type quarks, and neutral and charged leptons respectively.

These formulations highlight that the EW interaction is described by the gauge group SU(2)$_L$ x U(1)$_Y$, where L denotes left-handed fields, and Y signifies hypercharge. Hypercharge, a conserved quantity articulated by the Gell-Mann-Nishijima equation [19, 20], is pivotal in the electroweak theory. It serves as the generator of the U(1) group and is defined by the equation:

$$Y = 2(Q - T_3), \tag{2.10}$$

where Q is the electric charge and T_3 is the third component of weak isospin of a fermion. The hypercharge is therefore introduced in order to unify the weak and electromagnetic forces.

Although the EW theory unifies the weak and electromagnetic force, when the mediators of the electromagnetic force (massless photon) and weak force (massive $W^{\pm}$ and Z bosons) are introduced into the SM Lagrangian, the SU(2)$_L$ x U(1)$_Y$ gauge invariance is destroyed. This is due to the gauge symmetry not allowing the inclusion of mass terms, for the gauge bosons, in the Lagrangian.

2.1.5 Spontaneous symmetry breaking

In order to provide a mechanism by which the fermions and gauge bosons acquire mass, the electroweak symmetry must be broken. Theorists Robert Brout, Francois Englert and Peter Higgs developed the Brout-Englert-Higgs (BEH) mechanism, which gives a mass to the W and Z bosons when they interact with an invisible field, now called the "Higgs field" [10, 11, 21, 22]. This interaction is called spontaneous symmetry breaking (SSB), and describes a system in a symmetric state spontaneously ending up in an asymmetric state. In terms of the SM Lagrangian, which obeys certain symmetries, it exposes that at lowest-energy vacuum solutions, these same symmetries are not manifested. Therefore, the BEH mechanism shows that a solution to a given problem does not necessarily have the same symmetry as the problem. This phenomenon can be explained using the "Sombrero Potential" shown in Figure 2.4.

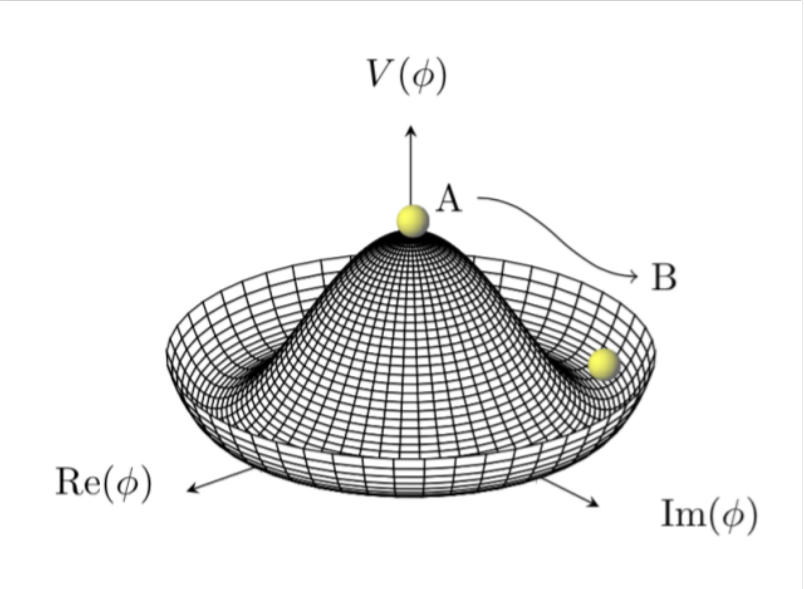


FIGURE 2.4: Demonstration of Jeffrey Goldstone's "Sombrero Potential" for potential function, $V(\phi)$.

In a simple idealised model, the SSB can be illustrated using scalar field theory. Considering the Lagrangian of a scalar field, ϕ, where the potential is given by $V(\phi)$,

$$\mathcal{L} = \frac{1}{2}\partial^{\mu}\phi\partial_{\mu}\phi - V(\phi). \tag{2.11}$$

The symmetry breaking is triggered in the term, $V(\phi)$, which can be expanded as:

$$V(\phi) = \frac{1}{2}\mu^2\phi^2 + \frac{1}{4}\lambda\phi^4. \tag{2.12}$$

If a positive mass term is considered, $\mu^2 > 0$, the minima of the potential will be zero. However, if a negative mass term is considered, $\mu^2 < 0$, the minima of the potential will lie within a circle of radius v, where $v = \sqrt{\frac{-\mu^2}{\lambda}}$. This demonstrates that the symmetry is spontaneously broken and a non-zero field, known as the vacuum expectation value (*vev*), is developed. Fermion and gauge fields couple with the *vev*, allowing mass terms to be included in the Lagrangian without destroying the gauge invariance. The

BEH mechanism therefore is able to provide fermions and gauge bosons with mass which is consistent with experimental observations. A consequence of the SSB in the SM is the expose and prediction of a massive scalar particle, the Higgs boson. The Higgs boson, introduced in Section 2.1.1, is given its mass from this generated *vev*.

2.1.6 Limitations of the Standard Model

The Standard Model of particle physics, while remarkably successful, does not encompass all known physical phenomena. Some significant examples that remain unexplained by the SM include:

- **Dark Matter and Dark Energy**: The SM does not account for dark matter, an invisible substance inferred from gravitational effects on visible matter, nor for dark energy, which is hypothesised to drive the universe's accelerated expansion. Together, these form the majority of the universe's mass-energy content [23].
- **Matter-Antimatter Asymmetry**: The SM cannot fully explain the observed dominance of matter over antimatter in the universe [24]. According to the "Big Bang" theory, matter and antimatter should have been created in equal amounts, yet we observe a universe largely composed of matter.
- **Neutrino Oscillations and Mass**: The phenomenon of neutrino oscillations implies that neutrinos have mass, contrary to the SM's original prediction of massless neutrinos [25]. This necessitates an extension to the SM to incorporate non-zero neutrino masses and understand their role in the cosmos.
- **Gravity and Quantum Gravity**: The SM does not include a quantum theory of gravity. Bridging the gap between quantum mechanics, as described by the SM, and general relativity, which explains gravity, remains a significant challenge in theoretical physics.
- **Hierarchy Problem**: The SM does not naturally explain the vast difference in scales between the weak force and gravity, nor why the Higgs

boson's mass is much lighter than the Planck scale, indicating the possibility of new physics beyond the SM.

- **Strong CP Problem and Axions**: The SM's inability to explain the absence of Charge Parity (CP) violation in strong interactions (the Strong CP Problem) has led to the hypothesis of new particles, such as axions, which could also potentially account for dark matter.

These unresolved phenomena represent some of the most profound mysteries in modern physics, driving the search for new theories and discoveries BSM.

2.2 Physics Beyond the Standard Model

The Standard Model of particle physics has been highly successful in explaining the behaviour of subatomic particles. However, it has limitations and cannot account for various phenomena such as dark matter and neutrino oscillations. In order to address these phenomena, physicists have proposed various theories and experiments to expand our understanding of the fundamental nature of the universe and go BSM. In this section, the phenomena unexplained by the current SM are expanded upon. This is followed by introducing the concepts of resonances, discovery significance and the look elsewhere effect which are fundamental in BSM studies.

In order to address the phenomena presented in Section 2.1.6, extensions and/or alternatives to the SM are explored. An important connection must be made to excesses found in run 1 and 2 data, produced at the LHC. These excesses demonstrate differences between experimental data and simulated SM data. Many BSM studies therefore explore potential models to explain these excesses together with these phenomena. In particular, the mechanism for electroweak symmetry breaking, often attributed to the Higgs sector in the Standard Model, still has many open questions. This makes it a fertile ground for exploring new theories that extend the Higgs sector and provide the opportunities for new bosons.

Resonance searches constitute a pivotal methodology in particle physics for detecting the existence of novel particles or phenomena. These searches

are characterised by identifying anomalous peaks within the invariant mass distribution of detected particle pairs. The concept of "resonance" in this context pertains to the mass and width of the postulated particle, which manifests as a distinct peak in the distribution analogous to a resonant frequency in other physical systems. Such resonance searches are instrumental in scrutinizing deviations or excesses observed in the data generated by the LHC. Incorporating BSM signals into our analytical framework enables a comprehensive evaluation of the extent to which these signals can account for the observed excesses within the SM paradigm. In order to extract these signal resonances of interest for further study, machine learning classifiers provide an ideal tool and will be expanded upon in Chapter 4.

The importance of a given resonance is measured using its statistical significance, commonly referred to as discovery significance. Discovery significance quantifies the confidence in distinguishing a true signal from background fluctuations. It is pivotal in affirming the detection of new particles or phenomena. Conventionally, a threshold of 5 standard deviations, corresponding to a p-value of $\pm 2.9 \cdot 10^{-7}$, is adopted. This criterion implies that the likelihood of attributing a signal of such magnitude to mere background noise is less than one in 2.9 million, thereby offering robust evidence for the presence of a novel entity.

Identifying new resonances within a specific mass range necessitates a comprehensive analysis that goes beyond merely assessing the statistical significance of localised spikes in observed events. An essential consideration is the probability of encountering such spikes anywhere within the examined mass range, a phenomenon termed the "**look elsewhere effect**" (LEE) [26]. LEE describes the increased likelihood of mistakenly attributing significance to a signal when conducting multiple searches within the same dataset. This misinterpretation arises because numerous searches elevate the chances of data fluctuations mimicking a significant signal, even if no genuine signal exists.

To mitigate this overestimation, the trials factor is employed. This factor adjusts the significance of an observed signal to reflect the number of conducted searches. In the context of new particle searches in particle physics,

the LEE and trials factor are critical. They substantially influence the interpretation of results and the confidence level in any discovery, ensuring a more accurate and reliable scientific conclusion.

2.3 Multi-Lepton Anomalies and Searches for Additional Bosons

The existence of additional scalar bosons can be included in the SM as long as their role in electroweak symmetry breaking is sufficiently small. This creates opportunities for the discovery of new bosons and an understanding of how they would affect the Higgs boson measurement. In order to investigate this possibility, the Two-Higgs-Doublet Model plus singlet (2HDM+S) can be used to explain the excesses in the LHC Run 1 and 2 data.

In the Two-Higgs-Doublet Model (2HDM), the Higgs sector of the SM, originally consisting of a single scalar doublet, is expanded by introducing an additional scalar doublet. This extension results in a richer Higgs spectrum post-electroweak symmetry breaking, featuring five distinct physical Higgs bosons: a pair of charged Higgs particles ($H^{\pm}$), two neutral CP-even scalars (commonly denoted as h and H, with h being the lighter), and one neutral CP-odd scalar (A). The 2HDMs are categorised into various types based on the unique coupling patterns of the doublets with up-type and down-type quarks. These include Type-I, Type-II, and other variants, each leading to different phenomenological implications, particularly in terms of flavour-changing neutral currents (FCNC) constraints. This diversity in coupling mechanisms allows 2HDM to provide a broad array of theoretical predictions, offering a fertile ground for exploring new physics beyond the SM.

The notation "$+S$" in 2HDM$+S$ model signifies the addition of a SU(2) singlet scalar field, denoted as S. This singlet scalar introduces a new layer of complexity by engaging in mixing interactions with the doublets of the 2HDM. Such interactions lead to notable modifications in the properties and interaction patterns of the Higgs bosons, potentially yielding new physics. The 2HDM$+S$ is just one of many possible theoretical models containing additional scalars [27, 28].

The 2HDM+S model framework, presented in Refs. [29, 30], introduces two real scalars, S and χ. In this model, the S scalar functions as a portal to dark matter through its interaction with χ, a dark matter candidate and possible source of missing transverse energy. A significant aspect of these studies is the investigation of the decay modes of the heavy scalar, H, within the model. It is observed that H predominantly decays into pairs of new singlet scalars (SS^* or SS'), where S' is a new neutral scalar with an approximate mass of 95 GeV, or into a combination of the singlet scalar and the lighter, Standard Model-like Higgs boson (Sh). This model's efficacy is demonstrated by its ability to account for observed deviations in the Higgs boson's transverse momentum distributions in LHC runs 1 and 2 data. The model, motivated by its success in explaining these excesses, was applied to the multi-lepton anomalies.

The multi-lepton anomalies observed at the LHC [29–37], describe specific processes characterised by final states containing two or more charged leptons (electrons and muons), with and without hadronic b-jets resulting in Higgs-like topologies. These occurrences are particularly noteworthy due to their statistically significant discrepancies from the predictions of the SM, as exposed using the prescribed theoretical model in Refs. [32, 34, 36, 37]. Furthermore, the model identifies a potential singlet candidate, S, as reported in Ref. [38]. Augmented with a Dark Matter candidate, as discussed in Ref. [39], the model offers insights into various astrophysical anomalies. Additionally, it has been extended to explain discrepancies in other experiments, such as the $g-2$ muon experiment (Fermilab reports in Refs. [40, 41]). A comprehensive review of the anomalies addressed by this model are available in Refs. [42, 43] and a summary of the multi-lepton excesses are exposed in Table 2.4.

The observed anomalies, particularly in multi-lepton events, align increasingly with the direct production of a scalar particle, H, with an approximate mass of 270 GeV. H predominantly decays into pairs of lighter scalars, $S(S')$, which exhibit characteristics similar to the SM Higgs boson. Notably, a subset of these anomalies features non-resonant, opposite-sign, different-flavour di-lepton final states, both with and without b-quark jets.

Final state	Characteristics	SM backgrounds	Significance
$\ell^+\ell^-$ + (b-jets) [34, 44, 45]	$m_{\ell\ell} < 100\,\text{GeV}$, $(1b, 2b)$	$t\bar{t}, Wt$	$> 5\sigma$
$\ell^+\ell^-$, (no jets) [32, 46]	$m_{\ell\ell} < 100\,\text{GeV}$	W^+W^-	$\approx 3\sigma$
$\ell^\pm\ell^\pm, 3\ell$ + (b-jets) [37, 47, 48]	Moderate H_T	$t\bar{t}W^\pm, t\bar{t}t\bar{t}$	$> 3\sigma$
$\ell^\pm\ell^\pm, 3\ell$, (no b-jet) [36, 49]	In association with h	$W^\pm h(125)$, WWW	$\gtrapprox 4\sigma$
$Z(\to \ell\ell)\ell$, (no b-jet) [34, 50]	$p_T^Z < 100\,\text{GeV}$	$ZW^\pm$	$> 3\sigma$

TABLE 2.4: Summary of the multi-lepton excesses [51].

The anomalies can be described by the production $H \to SS^*, SS', Sh$, where the first decays would be dominant. From the di-lepton component of the multi-lepton anomalies, and assuming $S \to W^+W^- \to \ell\ell\nu\nu, \ell = e, \mu$, the mass of S was predicted to be $m_S = 150 \pm 5\,\text{GeV}$ [32].

Prompted by these multi-lepton anomalies, a targeted search for the signatures of S within the specified mass range was conducted. This search, effectively a byproduct of the SM Higgs searches, focused on the kinematic sideband regions, explored by ATLAS and CMS, as detailed in Ref. [38]. The investigation combined the $\gamma\gamma$ and $Z\gamma$ channels, with associated leptons, di-jets, bottom quarks, and missing transverse energy. This comprehensive approach reported the first signs of a narrow resonance in the di-photon spectrum with local significance of 4.3σ and a global significance of 3.9σ for a mass $m_S = 151.5\,\text{GeV}$. Additionally, a narrow resonance has been reported in the di-lepton $t\bar{t}$ final state [44] and a broad resonance in the WW final state [52].

These findings bolster the hypothesis of a scalar resonance, S, decaying into photons, and to a lesser extent, $Z\gamma$, often in association with missing energy, jets, or leptons. Further corroboration with more data, as highlighted in Ref. [53], reveal significant indications of new scalar particles at approximately 95 GeV (S') and around 152 GeV (S). The existence of a 95 GeV scalar,

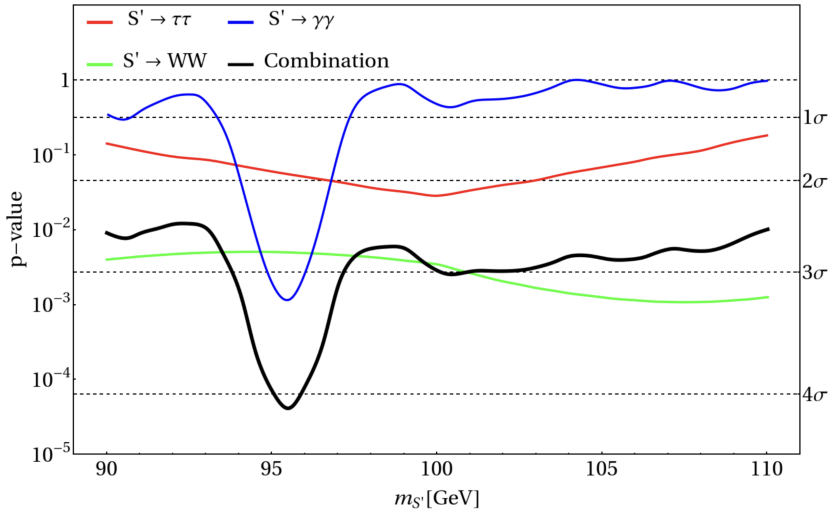


FIGURE 2.5: The p-values as a function of the mass of S' for the low-mass channels [53].

S', is particularly compelling, preferred over the SM framework with a confidence level of 3.8σ, as shown in Figure 2.5. In the context of the simplified model used here, the 95 GeV excess is a prime candidate to explain the presence of b-quarks in the multi-lepton anomalies. Moreover, interpreting the 152 GeV excesses within a simplified model that includes resonant pair production of S through the heavier scalar $H(270\,\text{GeV})$, results in a global significance of approximately 4.7σ, as shown in Figure 2.6.

While the production mechanism of S' remains an area of active investigation, the data robustly supports the associated production of S, possibly through the decay of a heavier boson, H, via the process $pp \to H \to SS^*$. An alternate or complementary decay chain could involve $H \to SS'$, where $S \to WW^*$ and S' would be responsible for generating the leptons and b-quarks, thereby explaining the multi-lepton anomalies observed in LHC data.

The compelling evidence presented by the scalar resonances S and S', and their potential roles in elucidating various anomalies observed in LHC data,

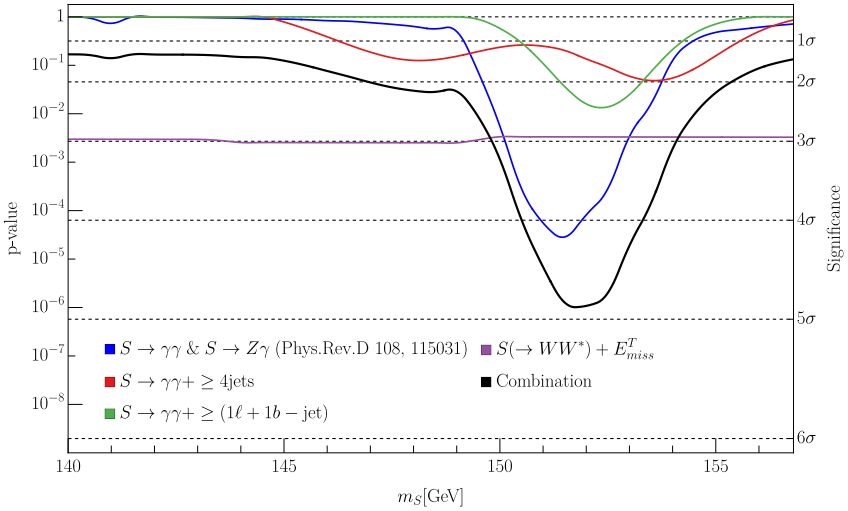


FIGURE 2.6: The p-values of the individual high mass channels as well as their combinations (including and excluding μe signal) [53].

set a strong foundation for further exploratory studies. The observed phenomena, especially in the context of multi-lepton events and their associated decay processes, provide a fertile ground for advanced analytical approaches. Motivated by the success and the findings of the model, this thesis seeks to delve deeper into these scientific inquiries. Specifically, it will explore the potential of semi-supervised machine learning classifiers as a mean to further understand and validate these results. This approach represents a novel frontier in particle physics research, leveraging the power of cutting-edge machine learning techniques to decipher complex data patterns and substantiate theoretical models in high energy physics.

Chapter 3

Particle Physics Experiments

Exploring the elusive realm of subatomic particles and their interactions, collider experiments stand at the forefront of experimental particle physics. In these sophisticated setups, particle beams are accelerated and precisely steered to collide, unlocking insights into the fundamental fabric of the universe. This methodology stems from a key principle in dimensional analysis: the inverse relationship between spatial scales and energy levels. Delving into the quantum world's minuscule distances necessitates the harnessing of extraordinarily high energies. To achieve this, state-of-the-art particle accelerators are employed, propelling particles to near-light speeds for high-energy collisions. Theoretical constructs, especially those derived from Quantum Field Theory (QFT), are put to the test through these experiments. They involve a meticulous comparison of measured collision cross-sections against theoretical predictions. These accelerators do not only facilitate collisions; they are intricate detectors, reconstructing and deciphering the remnants of the collisions. By analysing the nuances of collision events, researchers can calculate precise scattering cross-sections, offering invaluable insights into the forces and particles that constitute the very fabric of our universe.

In this chapter, an investigation of the Large Hadron Collider at CERN is presented, providing an exposé of the ATLAS detector and its subsystems. This sets the stage for the experimental methodologies and data analyses discussed in subsequent chapters.

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC), at the European Organization for Nuclear Research (CERN), is currently the largest and most powerful particle accelerator and collider in the world. The LHC is a circular accelerator situated beneath the France-Switzerland border [54]. Two parallel beam pipes circle around the LHC's circumference. These beams of particles, consisting of protons or heavy ions, are accelerated in opposite directions around the LHC. Once the two particle beams have reached optimal velocities, they are made to collide. These collisions occur at four points around the LHC, each contained by a different detector. At each detector, collision debris information is collected and processed as the fundamental data for the four major experiments. The LHC experiments, visualised in Figure 3.1, are:

1. A Toroidal LHC ApparatuS (ATLAS)
2. Compact Muon Solenoid (CMS)
3. A Large Ion Collider Experiment (ALICE)
4. Large Hadron Collider beauty (LHCb)

The ATLAS and CMS detectors at the LHC are versatile, general-purpose instruments engineered to investigate a broad spectrum of physics phenomena. Despite their parallel objectives in conducting similar studies, these experiments are distinguished by their unique design and technological innovations. This diversity in approach is crucial, as it enables cross-verification of scientific findings, ensuring robustness in the discovery process. On the other hand, the ALICE detector is specifically tailored for delving into heavy ion collisions, focusing on the study of the quark-gluon plasma, a state of matter believed to exist under extreme energy densities. Complementing these, the LHCb detector is dedicated to unravelling the mysteries of matter-antimatter asymmetry, with a particular emphasis on examining the properties and behaviours of the beauty quark. Each detector, with its distinct focus and methodology, plays a pivotal role in advancing our understanding of fundamental physics.

3.1.1 Design specifications

The LHC consists of a 26.7 km long concrete tunnel, constructed between 50 and 175 m underground [55]. This circular tunnel is used to accelerate two charged particle beams in opposite directions through two separate beam pipes. Each beam pipe maintains an ultra-high vacuum for the beams to travel through. A combination of thousands of superconducting electromagnets provide a magnetic field to guide and accelerate the beams. In order for the magnets to efficiently conduct electricity with negligible resistance or energy loss, the magnets, made of coils of superconducting electric cable, are cooled to -271.3°C. The magnet temperature is maintained using liquid helium.

The first series of magnets in the LHC consists of 1232 dipole magnets, each 15m long, which are used to bend and accelerate the beams. The second set of magnets, made up of 392 quadrupole magnets, each 5 to 7 m long, are used to focus the beam. Finally, just prior to collision, a combination of high order multipole magnets are used to "squeeze" the particles closer to the point of collision, increasing the probability of collisions.

In order to achieve proton-proton collisions, a bottled hydrogen gas is used as the source. Bunches of, on average, 120 billion protons are extracted by ionising the hydrogen. The accelerated proton beam consists of approximately 2808 bunches of protons. At the point of collision, the separation within the group of protons is such that collisions occur every 25 ns (at a frequency of 40 MHz) corresponding to 1 billion collisions per second. An important performance indicator for particle accelerators is the luminosity. The luminosity is proportional to the number of collisions occurring within a given area, cm^2, for a period of time, seconds. Therefore, a higher luminosity means more collisions occurring per area, per second and more data for the experiments to observe. The LHC was designed to achieve a beam with instantaneous luminosity of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, however the LHC has achieved instantaneous luminosities of more than double that.

The LHC is designed for proton beams to reach an energy of 7 TeV. This energy corresponds to a centre of mass energy (collision energy) of 14 TeV. Currently, the LHC has been run for three periods, none of which reached

14 TeV. From 2009 until 2013 the LHC was run for the first time (Run 1). During this period the maximum collision energy achieved was 7.8 TeV. The second period (Run 2), from 2015 to 2018, increased the collision energy to 13 TeV. Finally, during 2022 (Run 3), a collision energy of 13.6 TeV was achieved. Future upgrades will lead to further improvements in subsequent runs.

3.1.2 The CERN accelerator complex

The CERN accelerator complex, as shown in Figure 3.1, is made up of a series of particle accelerator machines. Each of the machines increases the energy of the particle beam before injecting it into the succeeding machine. The final accelerator in the sequence is the LHC, which increases the beam energy to record energies of 6.8 TeV per beam [56].

The current system for colliding proton beams begins with the Linear accelerator 4 (Linac4). In Linac 4, negative hydrogen ions are accelerated to 160 MeV. The ions are then both injected into the Proton Synchrotron Booster (PSB) and stripped of their two electrons. The beams, now composed of protons, are accelerated further in a series of accelerators including the proton Synchrotron (PS), Super Proton Synchrotron (SPS), and finally the LHC. The beam energies achieved at each machine in the series is summarised in Table 3.1.

Machine	Beam	Energy
Linac 4	negative hydrogen ions	160 MeV
PSB	protons	2 GeV
PS	protons	26 GeV
SPS	protons	450 GeV
LHC	protons	6.5 TeV

TABLE 3.1: Summary of beam type and energy of CERN accelerator complex machines involved in the proton beam acceleration.

When the beams are transferred to the LHC, they are injected into two separate pipes that are travelling in opposite directions, clockwise and anticlockwise. Once injected, it takes approximately 20 minutes for the beams

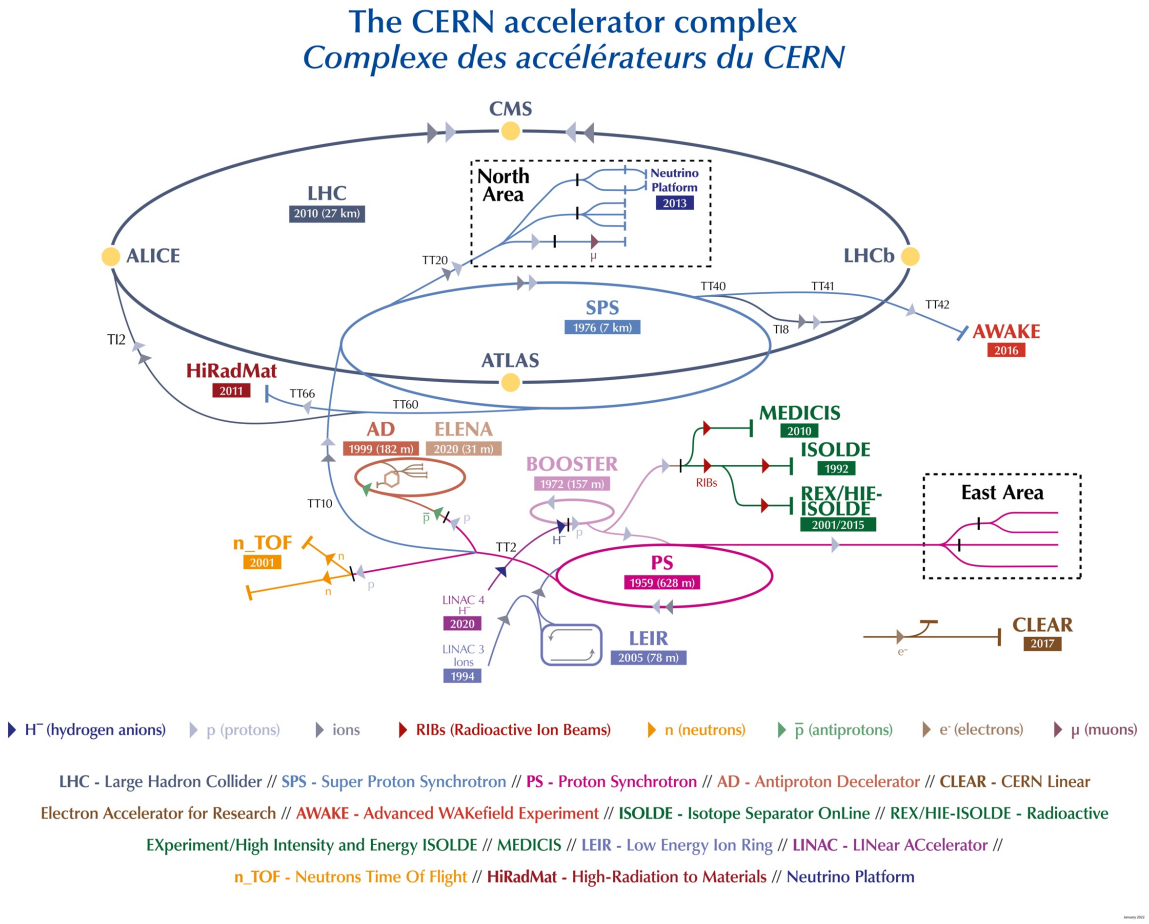


FIGURE 3.1: The CERN accelerator complex [57].

to be accelerated to their target energies and therefore be ready for collisions.

3.2 The ATLAS Detector

A Toroidal LHC ApparatuS (ATLAS), is the largest general-purpose detector at the LHC to date. It spans 46 meters in length, stands 25 meters high, and weighs 7000 tonnes [58]. As depicted in Figure 3.2, ATLAS is composed of various detection subsystems, all arranged around the point of collision.

The detector subsystems are able to measure the trajectory, momentum, and energy of particles. Individual particle debris can be identified, reconstructed and their properties recorded. The subsystems therefore track charged particles trajectories at high resolution and measure the energies

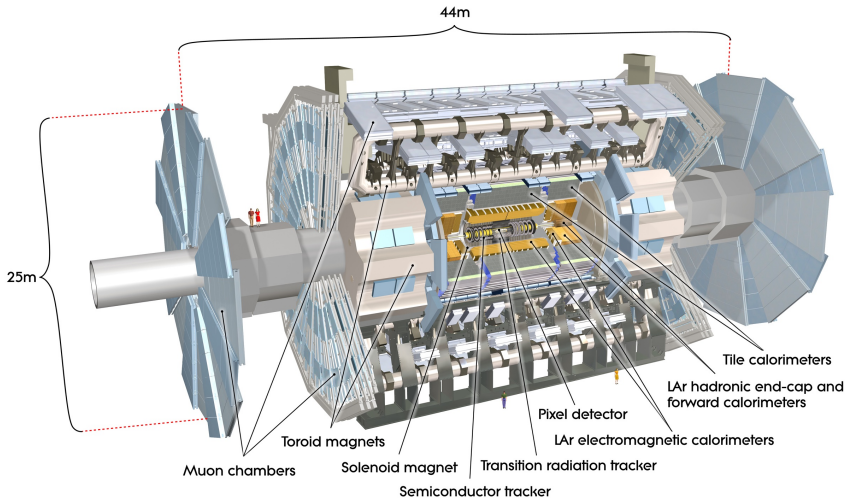


FIGURE 3.2: Generated image of the ATLAS detector [59].

of the visible final state particles. The non-visible missing transverse energy, E_T^{miss}, which can be produced from the production of invisible particles such as neutrinos, is also measured. The momenta of charged particles are measured using an extensive magnetic system which bends the paths of the particles, both measuring and directing them.

The detector systems are centred around the point of collision and move radially outwards. The cylindrical nature of the detector means that the coordinate system used to measure particles must also be cylindrical. Considering a Cartesian coordinate system, centred at the point of collision, the beam direction lies on the z-axis, tangentially to the LHC. The x-axis lies from the collision point towards the centre of the LHC and the y-axis is positioned vertically. The cylindrical coordinates can therefore be defined as:

$$R = \sqrt{x^2 + y^2}, \tag{3.1}$$

$$\phi = \tan^{-1}(\frac{y}{x}), \tag{3.2}$$

$$\eta = -\ln(\tan(\frac{\phi}{2})), \tag{3.3}$$

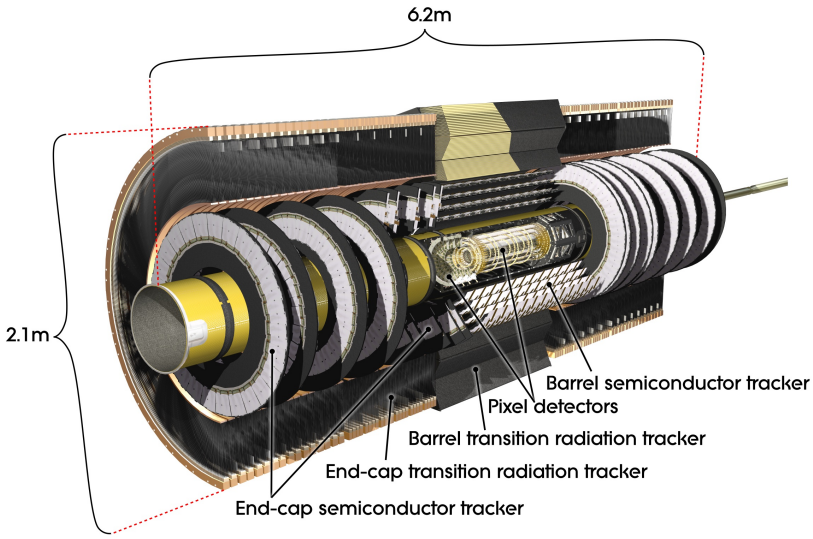


FIGURE 3.3: Generated Image of the ATLAS inner detector [60].

where R is the radial length, ϕ is the azimuthal angle and η is the pseudo-rapidity.

The modular subsystems that make up the ATLAS detector are the inner detector (ID), the calorimeters, muon spectrometers, the magnet system and the trigger and data acquisition systems. The following sections will expand upon each of these subsystems.

3.2.1 Inner detector

The inner detector is positioned directly around the point of collision, as shown in Figure 3.3. This very sensitive and compact detector consists of the pixel detectors, the semiconductor tracker (SCT), and the transition radiation tracker (TRT).

The first point of detection, only 3.3 cm from the beam line, is the Pixel Detectors. The Pixel Detectors consisting of four layers of silicon pixels with a total of more than 92 million pixels. The size of the pixels in the innermost layer are $50 \times 250\,\mu m^2$, while in the external layers they are $50 \times 400\,\mu m^2$. During collisions charged particles flying out from the point of collision

leave small deposits of energy in the Pixel Detector. These particle signatures, measured at a precision of almost 10 μm, are used to determine the momentum and the projected origin of each particle.

Surrounding the Pixel Detectors is the semiconductor tracker which is made up of four cylindrical barrel layers and 18 planar end-cap discs. These together contain more than 6 million implanted readout strips of silicon sensors (making up 6 million channels). These sensors are used to detect charged particles, produced in collisions, and reconstruct their tracks. The SCT measures the tracks of particles with a precision of up to 25 μm.

The final layer of the ID is the TRT, which is used to reconstruct particle tracks as well as to provide information on the particle type (electron or pion). The TRT consists of 300,000 straws (thin walled drift tubes) with diameters of only 4 mm. Each straw is filled with a gas mixture and has a gold-plated tungsten wire in its centre. Electric signals are therefore produced when the gas is ionised by charged particles passing through.

3.2.2 Calorimeters

The second set of detector subsystems at the ATLAS experiment are the calorimeters, as shown in Figure 3.4. The calorimeters measure the energy that particles lose as they move through the detector. The ATLAS calorimeters have layers of high-density materials interspersed with layers of an "active" medium. This high-density material is very absorbent and cause particles to deposit their energy, stopping them within the detector and measuring the particle's energy deposited.

Calorimeters are designed to halt and measure the energy of all known particles, except for muons and neutrinos. In the ATLAS experiment this includes the use of Electromagnetic (EM) calorimeters for electrons and photons, and Hadronic calorimeters for hadrons. The calorimeter system at ATLAS consists of the Liquid Argon Calorimeter (LAr) for EM measurements, and the Tile Hadronic Calorimeter for hadronic energy measurement.

The Liquid Argon Calorimeter is positioned around the ID. Its purpose is to measure the energy of electrons, photons and hadrons. The layers of

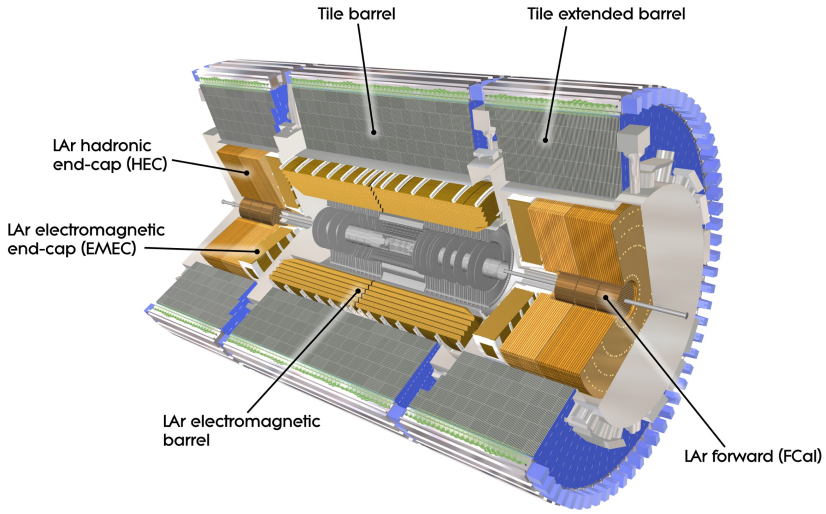


FIGURE 3.4: Generated image of the ATLAS calorimeter [61].

absorbent metal are made of either tungsten, copper or led. The interwoven layers of "active" liquid argon are therefore ionised by the particle energies and produces a current that can be measured. The LAr calorimeter is made up of the electromagnetic barrel and the end-caps. The end-caps include the electromagnetic end-cap, the hadronic end-cap and the forward calorimeter (FCal).

The second subsystem of the calorimeter is the Tile Hadronic Calorimeter, which surrounds the LAr calorimeter. Hadronic particles passing through the LAr calorimeter do not deposit all of their energy, the Tile Calorimeter therefore measures the energy of these hadronic particles. To this end the Tile calorimeter is made of layers of steel and plastic scintillating tiles. It consists of approximately 240,000 plastic scintillator tiles and is the heaviest component of the ATLAS experiment, weighing 2,900 tonnes.

3.2.3 Muon spectrometer

The muon spectrometer forms the outermost detector layers. It identifies and measures the momenta of the muons which have passed through the ID and calorimeters. The muon spectrometer contains 4,000 muon chambers and takes advantage of four different technologies. These are the Thin Gap

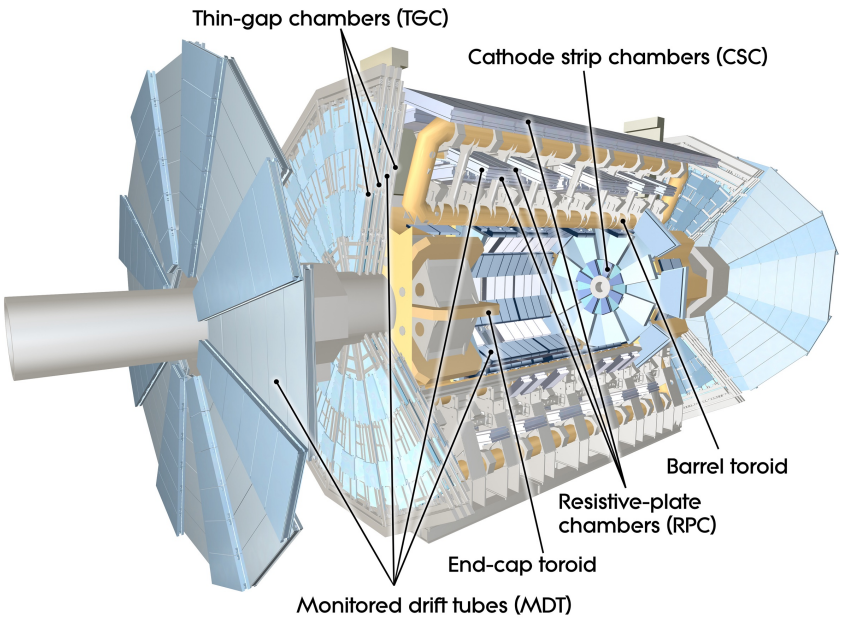


FIGURE 3.5: Generated image of the ATLAS muon spectrometer [62].

Chambers, the Resistive Plate Chambers, the Monitored Drift Tubes, and the Cathode Strip Chambers, which are shown in Figure 3.5.

The Thin Gap Chambers and Resistive Plate Chambers are used for triggering, and to measure the secondary coordinates at the ends and outer circumference regions of the detector, respectively. The Monitored Drift Tubes are used to measure the curvature of the tracks, with a tube resolution of 80 μm. The final component, the Cathode Strip Chambers, measure precision coordinates at the detector ends, with a resolution of 60 μm.

3.2.4 Magnet system

The last detector sub-system of the ATLAS detector is the magnet system. The magnet system is not directly used to detect particles but rather to bend the particle trajectories around the detector layers, enabling their charge and momenta to be measured. There are two types of superconducting magnets used, the solenoid and the toroidal, shown in Figure 3.6. The magnets are cooled to $-268^{\circ}C$, in order to produce a magnetic field of sufficient strength.

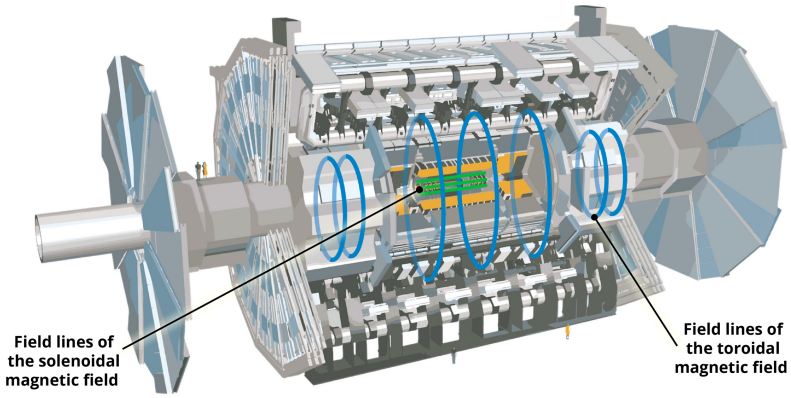


FIGURE 3.6: Generated Image of the ATLAS magnet system [63].

The solenoid magnet system is positioned around the ID, providing a 2 Tesla magnetic field. This system is 5.3 long, 4.5 cm thick and has a diameter of 2.4 m. The toroidal magnetic systems are used to measure the momentum of muons and produce magnetic fields of up to 3.5 Tesla. There are three toroidal magnet systems in ATLAS. The first is the barrel toroidal magnets, which are positioned around the centre of the detector. The second and third are the two end-cap toroidal magnets, which are positioned at either end of the experiment.

3.2.5 Trigger and data acquisition system

The trigger and data acquisition system (TDAQ) is the final piece to the ATLAS detector system, shown below in Figure 3.7. The TDAQ processes and filters the collision data in real time, selecting events with interesting characteristics that may lead to new discoveries, and discarding events that do not. The TDAQ system is necessary to deal with the immense amount of data produced by the ATLAS detector when observing collisions. When running, the ATLAS detector can observe up to 1.7 billion proton-proton collisions per second, corresponding to a total data volume of more than 60 terabytes per second. As this data flow is too large to effectively deal with, the ATLAS trigger systems selects and extracts the events of interest, reducing the volume and refining the quality of the data [64].

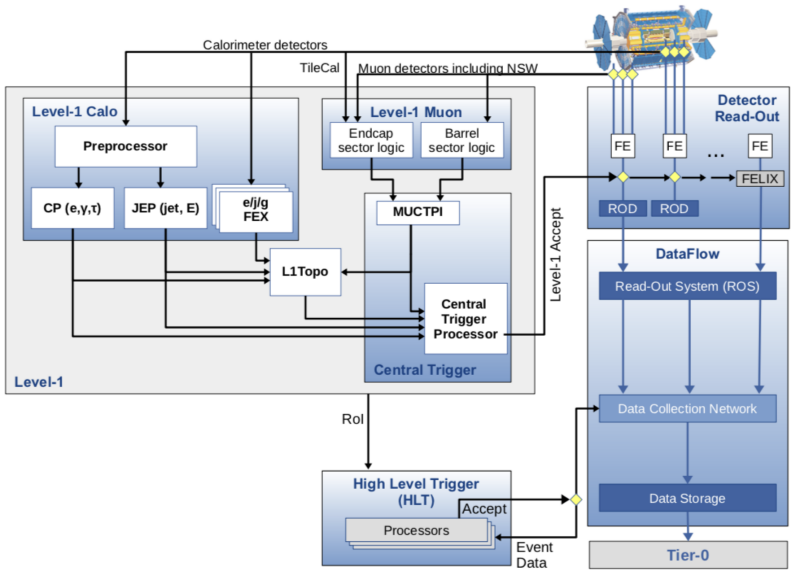


FIGURE 3.7: Diagram of the ATLAS trigger and data acquisition system [65].

The ATLAS trigger system is made up of two components, namely the hardware trigger (Level-1) and the software trigger (Level-2). The hardware trigger is located on the detector and consists of specially designed custom electronics. This hardware trigger works directly on the data from the calorimeters and Muon Spectrometer. As events are detected, they are passed to the Level-1 trigger, where they are stored in a storage buffer. Each event in the storage buffer is processed by the trigger within 2.5 μs of the collision event occurring. Events are selected by the Level-1 trigger and passed to the Level-2 trigger at a rate of approximately 100,000 events per second.

The software trigger, Level-2, takes events selected by the hardware trigger and performs a detailed analysis of each of the collision events. This analysis is conducted for each of the detector regions and takes only 0.2 seconds. In order to achieve this, the software trigger operates from a server farm of approximately 40,000 CPU cores. The Level-2 trigger selects around 1000 events per second which are sent to be stored in the data storage system for further analysis.

Chapter 4

Machine Learning in Particle Physics

4.1 Introduction to Machine Learning

Machine learning is a computational system that utilises algorithmic and statistical models in order to perform specific tasks. It does this by finding patterns and anomalies within a dataset. Machine learning techniques are currently being developed and applied in a vast array of applications and industries. Although the fundamental concepts underlying machine learning models are relatively consistent, the variety and complexity of techniques used to further improve specific use cases is expanding at a rapid rate. This growth is being driven by the increasing availability of big data, advancements in computing power, and the development of sophisticated machine learning algorithms. Some key industries and use cases of machine learning today include healthcare, where it improves the outcomes of patients through disease risk prediction, identification of at-risk patients, personalised treatments, and medical image analysis to enhance the ease and accuracy of diagnoses. In finance, machine learning detects and prevents fraud, predicts stock prices, makes investment recommendations, and extracts previously inaccessible patterns and trends in financial data. Marketing benefits from insights into customer behaviour, personalised recommendations, targeted advertisements, and identification of the most effective channels. Computer vision, applied to images, enhances security systems with real-time facial recognition, suspicious activity detection, and crime

prevention technologies. As a core technology in autonomous vehicles, machine learning enables decision-making based on sensor data and safe navigation in complex environments. Image generation through machine learning creates high-quality images of almost any conceivable content. Natural language processing systems, including chatbots such as CHAT-GPT [66], understand and respond to human language, making virtual and voice assistants easily accessible for a wide range of tasks, including online information access and complex problem-solving, such as to pass legal examinations.

In high energy physics, machine learning is widely researched and applied to many applications. These include classifiers to extract BSM particle signatures [67, 68], to label event and jet images [69, 70], and for particle track reconstruction [71]. Further applications include methodologies to generate realistic particle events, such as photon shower simulation [72]. A comprehensive list of machine learning related research in particle physics can be found in Ref. [73].

Machine learning is a data-driven technology, this means that relevant data is fundamental to any implementation. The success and ability of a model is therefore dependent on the accessibility, quality and scale of suitable data. As particle physics data produced at the LHC is arguably one of the leading areas of big-data and big-data analysis, it has remained at the forefront of machine learning research. With more recent large scale digitalization of information, and push for big-data research by companies such as Google, Facebook and Microsoft, it is important for a continuous exchange of research and technologies between industry and scientific research.

This chapter will firstly introduce machine learning classifiers together with their important fundamental concepts. This will be followed by a description of the different supervision techniques used in machine learning. The chapter will then introduce the Boosted Decision Tree and neural network models which will be implemented in the following chapters. Finally, relevant Monte Carlo and machine learning based data generators are introduced.

4.2 Machine Learning Classifiers

A machine learning classifier is a type of algorithm that employs computational algorithms to categorise input data according to its attributes. These classifiers are trained using labelled and/or unlabelled data to identify patterns that enable them to sort new, previously unseen data. The classifier's effectiveness hinges on several factors, including the volume and quality of the training data, the architecture and intricacy of the model, as well as domain-specific considerations.

Binary classifiers are a specialised form of classifier that distinguishes between two distinct classes. In the field of high-energy physics, binary classifiers are commonly used to isolate "signal," which corresponds to the particle events of interest, from "background," which encompasses unimportant particle events. To train a model for such discrimination, each of these categories is assigned a specific label, namely zero for background events and one for signal events.

In deploying classification models, key elements to consider include the loss function, model responsiveness, model architecture and hyperparameters, and the risk of overfitting. This section will elaborate on each of these foundational aspects of machine learning.

4.2.1 Loss function

In machine learning, a loss function serves as a gauge to quantify the discrepancy between a model's predicted output and the true output for a given data sample. The objective in training a machine learning model is to minimise this loss function across all training instances, thereby enhancing the model's predictive accuracy. The choice of loss function can differ based on the problem type, whether it is binary classification, regression, or other tasks. It can be designed to penalise various types of errors, such as false positives or false negatives, or to assess the magnitude of prediction errors.

Considering a binary classifier used to classify labelled data, where each event, $\vec{x}_i$, has a corresponding target/label, y_i. The value of the target, y_i, is

equal to "0" if $\vec{x}_i$ is a background event, and "1" if $\vec{x}_i$ is a signal event. For each given event, the model generates an output response, $\hat{y}_i \in (0,1)$. Model training therefore aims to minimise the difference between the targets and the responses of events, using a loss function.

The loss function most commonly used for binary classification problems, is the binary cross-entropy (BCE) function. Which quantifies the difference between the predicted probabilities and the actual binary labels in the dataset as:

$$\ell(y,\hat{y})_{\text{BCE}} = -y \cdot \log \hat{y} + (1-y) \cdot \log(1-\hat{y}). \tag{4.1}$$

An alternative loss function is the mean squared error (MSE). This function quantifies the average of the squares of the differences between the predicted, $\hat{y}$, and actual values, y, thereby providing a measure of the model's prediction accuracy. While MSE is predominantly utilised in regression problems due to its sensitivity to the magnitude of errors, it can also be adapted for binary classification tasks, albeit with certain considerations to address its characteristics in such contexts. The MSE for a set of predictions can be formally defined as:

$$\ell(y,\hat{y})_{\text{MSE}} = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2, \tag{4.2}$$

where y represents the vector of actual values, $\hat{y}$ denotes the vector of predicted values, N is the number of observations, and $\hat{y}_n - y_n$ calculates the prediction error for each observation.

The loss function is therefore pivotal in the machine learning model training process. Employed by an optimization algorithm, it guides the adjustments made to the model's internal parameters, and serves as a gauge for assessing the effectiveness of those adjustments. The training cycle involves iteratively fine-tuning these internal parameters with the aim of minimizing the loss function. Once the loss function ceases to improve, the model is considered to have reached convergence, delivering its optimal performance within the constraints of its configuration and architecture.

4.2.2 Model response

The response of a machine learning model typically refers to the output or prediction that the model produces on a given sample. When evaluating a model's response, it's common to analyse the response distribution. This approach helps in understanding how frequently different responses occur in relation to each other across a set of events. Considering a binary classifier used to separate signal from background, the desired output is a response distribution where signal events tend to 1 and background events tend to 0. If all signal events have response values greater than 0.5 and all background events have response values smaller than 0.5, a cut at 0.5 would with 100% success allow the classification of signal from background.

The response of a model can be evaluated using metrics such as accuracy, precision, recall, F1 score, and the Area Under the Curve (AUC), depending on the specific task and domain of the model. These metrics consider if the model responses are true positives (Tp), true negative (Tn), false positive (Fp), or false negative (Fn). True positives and true negatives are the number of events correctly labelled as positive and negative respectively. False positives are the number of negative events incorrectly labelled as positive, and false negatives are the positive events incorrectly labelled as negatives. If we consider a binary classifier used to classify signal events from background, signal events are defined as positives and background events as negatives.

The accuracy of machine learning classifier is a measurement of the proportion of correctly labelled cases (both positive and negative). The accuracy of a model's response can be calculated as:

$$\text{accuracy} = \frac{Tp + Tn}{Tp + Fp + Tn + Fn}. \tag{4.3}$$

A model with high accuracy is therefore a model that is able to correctly identify both positive, signal, and negative, background, cases. Precision measures the proportion of predicted positive cases that are actually positive. The precision of a model's response can therefore be calculated as:

$$\text{precision} = \frac{Tp}{Tp + Fp}. \tag{4.4}$$

A model with high precision is therefore a model that is successful in identifying positive cases and not misclassifying negative cases as positive. The recall of a binary classifier measures the proportion of actual positive cases that the model correctly labelled as positive. The recall of a model's response can be calculated as:

$$\text{recall} = \frac{Tp}{Tp + Fn}. \tag{4.5}$$

A model with high recall therefore is able to correctly identify the positive cases. This is important in applications where missing positive cases can have serious consequences. The F1 score is a harmonic mean of the precision and recall of a binary classifier, providing a single score that balances both scores. The F1 score can be calculated as:

$$F1_{\text{score}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{4.6}$$

The F1 score takes into account both the proportion of positive cases that are correctly identified (recall) and the proportion of predicted positive cases that are actually positive (precision). A high F1 score reflects both a high recall and precision, while a low F1 score indicates that the model is insufficient in one or both metrics. An F1 score is therefore especially useful when classifying a dataset with an uneven distribution of positive and negative cases.

The Receiver Operating characteristic (ROC) curve is a plot that graphically illustrates a binary classifier's performance at different classification thresholds. The ROC curve compares the rate of true positives, y-axis, to the rate of false positives, x-axis, as shown in Figure 4.1.

The AUC, of the ROC curve, is a scalar value that represents the overall performance of a model. The value of the AUC ranges from 0 to 1. An AUC value of 0.5 indicates that the model is randomly guessing outputs, while

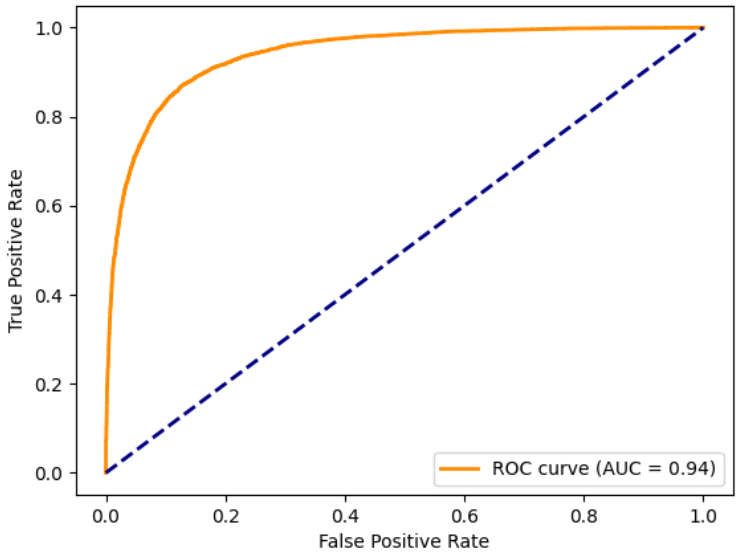


FIGURE 4.1: Receiver operating characteristic curve plot example.

a value of 1 reflects that the model is able to classify all events with 100% accuracy. The ROC curve shown in Figure 4.1, has an AUC of 0.94 reflecting that it is able to successfully classify events 94% of the time. A ROC curve therefore exposes how successfully a model is able to distinguish between positive and negative classes, and the AUC quantifies the overall ability of the model to make correct predictions.

4.2.3 Model architecture and hyperparameters

In the realm of machine learning, model architecture refers to the structured design and layout of algorithms, encapsulating elements like layers, nodes, and activation functions. Essentially, it serves as the blueprint that dictates how information flows through the model. Different architectures lend themselves to varying types of problems such as classification, regression, or more intricate tasks like sequence prediction. The architecture governs both the computational complexity and the performance capabilities of the model. Its importance cannot be overstated; a well-designed architecture not only enables effective training but also impacts how well the model generalises to new, unseen data.

Hyperparameters, on the other hand, play a crucial role in dictating the learning process and performance of models. Unlike parameters, which are learned from the training data, hyperparameters are set prior to the learning process and are not derived from the data itself. They include variables such as learning rate, batch sizes, regularisation coefficients, and architecture-related decisions for neural networks. The appropriate selection and tuning of these hyperparameters is vital, as it profoundly influences the model's ability to effectively learn from data and generalise to unseen scenarios.

4.2.4 Overtraining

Overtraining, also known as overfitting, is a critical challenge in machine learning where a model excessively learns from the training data, including its noise and random fluctuations. This occurs when a model becomes overly complex and starts fitting to the idiosyncrasies of the training data rather than learning the underlying general patterns. As a consequence, while the model might exhibit high accuracy on the training dataset, its performance significantly drops when exposed to new, unseen data, a phenomenon marked by high prediction variance. To counter overfitting, practitioners implement strategies like regularisation, which penalises the complexity of the model, and early stopping, which halts the training when the model's performance on a validation set no longer improves. Additionally, using a validation set for continuous performance monitoring during training, helps in balancing the model's ability to learn from the training data while retaining its ability to generalise on new data. These methods collectively work towards maintaining an optimal balance between learning detailed patterns from the training data, and being able to apply the same level of success to new, unseen data.

4.3 Machine Learning Supervision

Machine learning has two distinct phases, training and inference. The training phase consists of "teaching" a model to be able to best achieve a function on a training dataset. The secondary phase, inference, refers to when a trained model is used to extract information from new data as it was trained

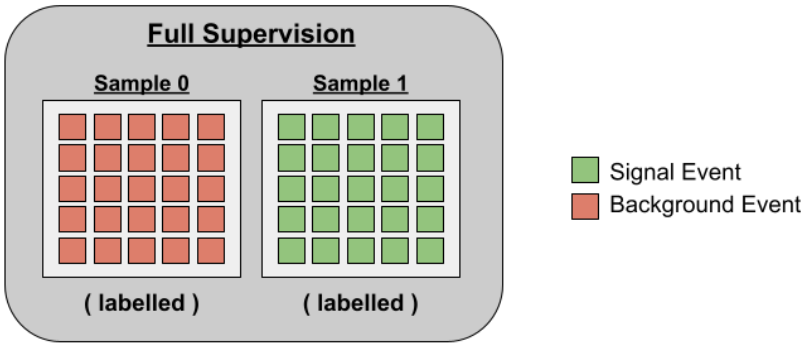


FIGURE 4.2: Fully supervised classifier event labelling.

to do. During the training phase the term "supervision" refers to the extent to which a model is guided or labelled. The three main categories of machine learning supervision expanded upon in this section are fully supervised, unsupervised, and semi-supervised.

4.3.1 Fully supervised learning

Fully supervised machine learning models are trained using completely labelled data. The model is provided with a dataset with distinct categories, each with a unique label/target, usually in the form of whole numbers. If we consider a binary classification example where signal is being classified from background, the background and signal are labelled "0" and "1" respectively. The model can therefore be trained on these two distinct samples, as shown in Figure 4.2.

The label is the desired output of the model and is therefore integral in training the model on what it is trying to achieve. The objective of the model is to output a response tending to "1" for any given signal event and a response tending to "0" for a given background event. In turn the training of the model involves repeatedly minimizing the difference between the true, y_i, and predicted, $\hat{y}_i$ label until the model no longer improves. This can be mathematically expressed as the minimising of the loss function, $\ell(y_i, \hat{y}_i)$, calculated as:

$$f_{full} = argmin_{f:\mathbb{R}^n \to [0,1]} \sum_{i=1}^{N} \ell(y_i, \hat{y}_i). \tag{4.7}$$

Fully supervised machine learning classifiers can therefore achieve very high accuracy and are easy to evaluate because the model is trained with a clear relationship between the inputs and outputs. It can be applied to a wide range of problems as long as the data is correctly labelled. Fully supervised methods however have limited generalization and are prone to overfitting, due to becoming too specialised to the training dataset. This leads to the potential of fully supervised models having difficulty handling unseen data.

4.3.2 Unsupervised learning

Unsupervised learning is a style of learning where a model is trained on unlabelled data. This means that no pre-existing labels are attached to the data presented to the model while it is being trained. It is therefore a type of self-organised Hebbian learning that determines patterns in a dataset that are unlikely to have been seen before. It is advantageous that unsupervised learning does not need labelled data, which is often difficult to obtain and evaluate. Unsupervised learning is an excellent tool to reduce the dimensions of a dataset by reducing the number of features while retaining the most relevant information.

However, unsupervised learning is difficult to evaluate and interpret due to there being no clear objective function, as in full supervision. The results are often more difficult to interpret as underlying patterns are less likely to be immediately understood. Finally, unsupervised methods are more susceptible to be influenced by noise in the data as they are not guided to avoid it.

Unsupervised learning, distinct from supervised methods, processes data without pre-existing expectations or knowledge, free from the constraints of prior labelling. This approach allows the algorithm to independently identify patterns, providing insights without the influence of the trainer's biases or perceptions. An exemplary application of unsupervised techniques is detailed in Appendix B.3, where Gaussian mixture models are employed to cluster geolocation data.

4.3.3 Semi-supervised (weak) learning

Semi- or weak supervised techniques utilise aspects of both supervised and unsupervised learning styles. This technique makes use of a combination of labelled and unlabelled data during training. The model is trained with a well constrained sample of labelled data together with an unlabelled sample. The machine can therefore learn to accurately discern events that match the labelled data, as well as find patterns within the unlabelled events which aren't restricted by prior expectations. This takes advantage of both the successes of supervised and unsupervised learning.

In particle physics, the fully-supervised approach can falter if the manifestation of new physics diverge from the model used as a reference. This means that if the simulated data that these models are trained on is flawed, due to insufficient truth-level information, the model will learn these artefacts of the simulation. In order to reduce this model dependency, weak supervised methods can be used [74–76].

The classification without labels (CWoLa) study, presented by the ATLAS collaboration, introduces weak learning techniques to reduce the model dependency on simulated data [77]. The study uses mixed samples (combining signal and background events in each sample) to train binary classification models, and expose that weak supervised models are able to classify signal events with success comparable to fully-supervised methods [78–81]. Variations of the weak supervised method have been successfully used for resonance anomaly detection [68, 82–84]. A further advantage of the weak supervised method is shown to be its ability to be trained on impure samples [85], or directly on data, reducing the need for simulation.

Research into the use of weak supervision in particle physics, has been extended to include multi-model classifiers [86], graph neural networks [87], and with the inclusion of transfer learning [88]. Finally, the viability of the methodologies is exposed for narrow resonance searches in Refs. [68, 82] with a review of weak supervision in Ref. [89].

This weak methodology can be adapted to use partial labelling, with a labelled background sample and unlabelled mixed (signal and background)

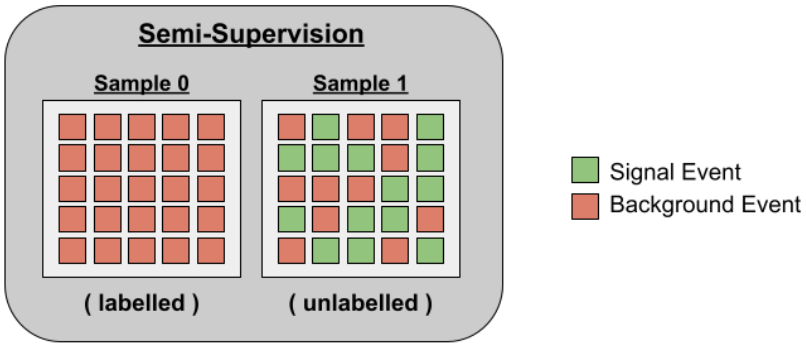


FIGURE 4.3: Semi-supervised classifier event labelling.

sample. This semi-supervised technique, implemented via a side-band analysis, is able to reduce model dependency and detect new physics in the form of topological variations between the possible signal and the side-band regions [90]. The semi-supervised methodology implemented in this thesis, will therefore use the informed knowledge of the background processes, sample 0, to determine patterns within the unlabelled sample, sample 1, as shown in Figure 4.3.

When applying these semi-supervised learning techniques, the minimisation of the loss function is described mathematically as:

$$f_{semi} = argmin_{f:\mathbb{R}^n \to [0,1]} \sum_k \ell \left(\frac{1}{|k|} \sum_{i \in k} \hat{y}_i, y_k \right), \tag{4.8}$$

where k denotes the batches of training data and y_k is the signal ratio in each batch. Particle signal processes can thus be extracted without having to conform to prior definitions. This in turn encourages the potential of experimentally finding discrepancies between the current theoretical expected results, and those experimentally determined. This may therefore lead to an increased chance of discovering new physics.

4.4 The Boosted Decision Tree

When considering binary classifiers, one of the most common and successfully implemented techniques in high energy particle physics is the Boosted

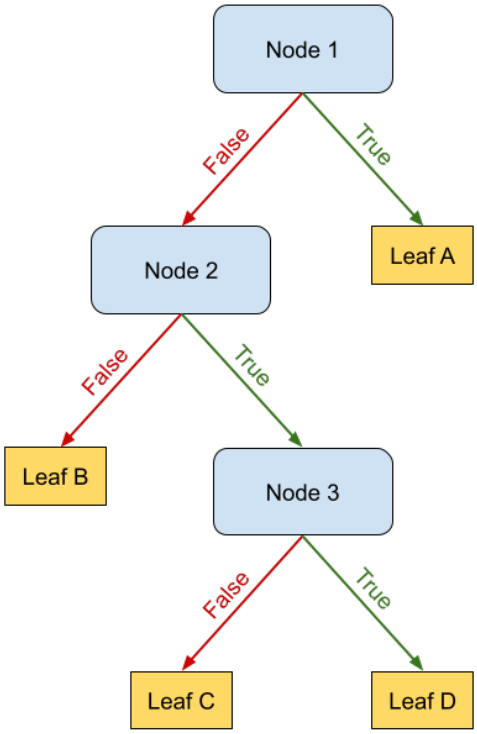


FIGURE 4.4: Illustrative decision tree structure with conditional branches and terminal nodes.

Decision Tree (BDT) [91–93]. BDTs are powerful tools for analysis due to their ability to handle many input variables, have a high discrimination power, and are relatively easy to interpret.

4.4.1 The decision tree

A decision tree takes a set of input features and separates the input data recursively based on those features. The structure of a decision tree, shown in Figure 4.4, consists of nodes and leaves. The first node in the decision tree is called the root node. Each node represents a binary decision that is based on a "cut criteria" of one of the input features. The leaves of the decision tree represent a class label or probability. In the case of binary classification of particles, the leaf nodes represent the class labels, signal and background.

The training of a decision tree consists of defining the "cut criteria" for each node. The training process starts by defining the feature and corresponding

cut value, that provides the best separation between signal and background events for the entire training sample. The training sample is split using the decision criteria at the root node, and two new nodes are created for each sub-sample. The creation of each node repeats the process of defining the root node on the given subset of data. This process of creating new nodes is stopped once a given criterion is met, such as the lowest node reaches a minimum number of events or a maximum signal purity.

When training a BDT, the value connected to the decision at each node is selected to minimise the entropy. Information gain is calculated as the difference in entropy before and after the split and is therefore maximised in training. The entropy is at a minimum for a $1/0$ split, and at maximum for a $50/50$ split. Each decision node in the BDT is created recursively on the given features.

4.4.2 Boosting

To enhance the decision trees ability to classify data, a composite model known as the Boosted Decision Tree (BDT) can be employed. This model synergises the strengths of multiple decision trees through a method called "boosting". Boosting iteratively constructs a series of trees where each tree is designed to address the misclassifications of its predecessors. In this process, each decision tree, $h_t(x)$, is assigned a weight, w_t, which reflects its accuracy, with more accurate trees receiving higher weights. The ensemble's final prediction, $\hat{y}$, is a weighted sum of the outputs from all individual trees:

$$\hat{y} = \sum_t w_t h_t(x), \tag{4.9}$$

where t indexes the decision trees in the BDT. The optimization of a BDT involves minimizing an objective function, $O(x)$, which is defined as:

$$O(x) = \sum_i l(\hat{y}_i, y_i) + \sum_t \Omega(f_t), \tag{4.10}$$

where $l(\hat{y}_i, y_i)$ denotes the loss function that measures the discrepancy between the predicted value, $\hat{y}_i$, and the true value, y_i, for the i^{th} data point. The term $\Omega(f_t)$ represents a regularisation term that penalises the complexity of the t^{th} tree, preventing overfitting. Boosting methods include Adaptive Boosting (AdaBoost), Gradient Boosting, and Extreme Gradient Boosting (XGBoost), each iteratively adding trees to the model to reduce the overall loss function.

The **adaptive boosting** (AdaBoost) method is considered one of the pioneering forms of boosting [94]. In AdaBoost, once training events are distributed into their respective final leaf nodes, the re-weighting mechanism takes effect. Signal events misclassified as background, as well as background events misclassified as signal, are assigned with significantly larger weights compared to those correctly classified. These re-weighted training events are then used to construct a new decision tree. This process, repeated often several hundred times, culminates in a collection of decision trees termed a forest.

The **gradient boosting** technique constructs new trees using the principles of gradient descent [95]. It operates by training a succession of models incrementally, treating each tree as an additive component within a function expansion framework. The parameters for each tree are optimised by minimizing a differentiable loss function. For classification tasks, this function typically assumes the form of a binomial log-likelihood. Consequently, gradient boosting refines the model by sequentially introducing learners that are optimally aligned with the negative gradient of the loss function for the entire ensemble.

Lastly, the **extreme gradient boosting** (XGBoost) approach is a more sophisticated and regularised version of gradient boosting [96]. Unlike its predecessor, XGBoost builds trees in parallel rather than sequentially. It evaluates the performance of each decision node by scanning the gradient values, utilizing the partial sums at each level of the trees. XGBoost leverages second-order derivatives of the loss function, which not only indicate the direction of the gradients but also how to minimise the loss function

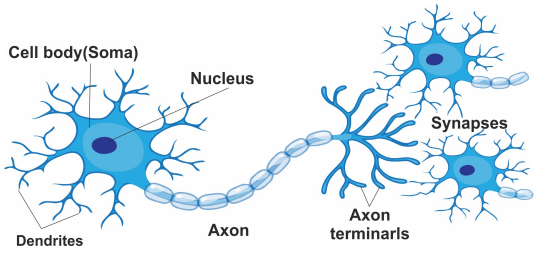


FIGURE 4.5: The biological neuron [98].

most effectively. This, paired with advanced regularisation techniques, significantly enhances the model's ability to generalise.

4.5 The Neural Network

Artificial neural networks are based on the network of biological neurons in the brain. The biological neuron, shown in Figure 4.5, is the building blocks of the human nervous system [97]. Neurons are distinguished from other cells in the human body by their unique ability to process and transmit information, facilitating communication across various bodily systems. The dendrites of a biological neuron are specialised to receive information from adjacent neurons and relay it to the soma, or cell body, which houses the cell nucleus and processes the incoming information. The soma then converts the processed information into chemical signals, which travel along the axon to the synapses. The chemicals cause the synapsis to send neurotransmitters to the dendrites of adjacent neurons. This intricate network of neurons possesses the remarkable capacity to store, process, and transform information, underpinning complex problem-solving and decision-making capabilities. Hence, the biological neuron is fundamental to the architecture of human intelligence.

The human brain has more than 86 billion neurons [99], forming a neural network that processes and integrates sensory information, controls bodily functions and movements, regulates emotions and behaviour, forms and retrieves memories, and enables thinking and problem-solving.

In the late 1940s and early 1950s, the first artificial neural networks, ANN,

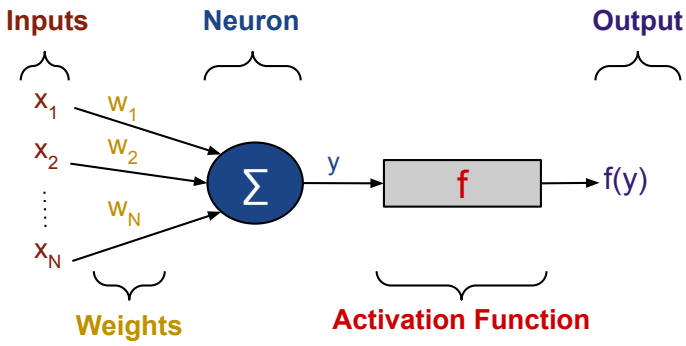


FIGURE 4.6: The artificial neuron.

were developed to mimic the way biological neural networks learn and process information. Since these early developments advancements in computational power and algorithms has led to ANNs being widely used today. The artificial neuron functions very similarly to the biological one, receiving and processing a signal before sending the processed signal to the neurons connected to it. The connections between artificial neurons are called edges. The "signal" at each edge is a real number, and is processed, at the output of the neuron, using a non-linear function of the sum of its outputs. Weights at each neuron edge, are adjusted during training to "teach" the neural network to understand patterns in the training data. The artificial neuron mechanism, can be described in more detail using the diagram in Figure 4.6.

The artificial neuron model, as illustrated in Figure 4.6, receives N inputs labelled $x_1, x_2, \ldots x_N$. These inputs are analogous to the dendrites in a biological neuron and are each multiplied by their respective synaptic weights $w_1, w_2, \ldots w_N$. The artificial neuron computes the weighted sum of these inputs, adding a bias term, b, to account for threshold effects, as captured by the equation:

$$y = \sum_{i}^{N} x_i * w_i + b, \tag{4.11}$$

where y is the neuron output before activation. Subsequently, this output

is passed through an activation function, f, which introduces non-linearity and determines the final output of the neuron. This output can then be used as an input to subsequent neurons in the network.

In an ANN, neurons are aggregated into layers. Each layer may perform different transformations on their inputs. The input features of the model are processed by the first layer of neurons, known as the input layer. The signal outputs of each layer are sent to the proceeding layer until they reach the last layer, the output layer, where the final output of the ANN is produced.

4.5.1 Neuron activation

The neuron activation or activation functions, are a fundamental component of neural networks. Activation functions are non-linear functions applied to the neuron output, y, that introduce non-linearities in the neuron outputs. This is essential in allowing neural networks to learn and model complex relationships between the input features and model output. There are many types of activation functions used in ANNs, each with its own advantages and disadvantages. Some of the most relevant activation functions are Sigmoid, ReLU, Tanh, and Softmax which are expanded upon in this section.

The **Sigmoid function** is a smooth, S-shaped function that is nonlinear, monotonic and continuously differentiable [100]. The sigmoid function, calculated as:

$$f(y) = \frac{1}{1 + e^{-y}}, \tag{4.12}$$

maps inputs to a value between 0 and 1. This makes it an ideal output function for neural networks used for binary classification problems.

The primary limitation of the sigmoid activation function arises from its characteristic saturation at the extreme ends of its output range. When the input values are significantly high or low, the sigmoid function flattens out, causing the gradients to approach zero. This phenomenon, referred to as the vanishing gradient problem, can severely hamper the optimization process during training. As the gradients become negligibly small, the weight updates during backpropagation are minimal, leading to a situation where the

learning stalls, and the network cannot effectively learn or further improve its performance.

The **Rectified Linear Unit, ReLU**, function is a piecewise linear function that outputs zero if the input value is negative, and the input value itself if it is positive [101], calculated as:

$$f(y) = \begin{cases} 0 & \text{if } y \leq 0, \\ y & \text{if } y > 0. \end{cases} \tag{4.13}$$

ReLU is computationally efficient compared to other activation functions and introduces non-linearity into the network allowing it to learn complex relationships between inputs and outputs. It introduces sparse representations into the network, meaning that only a subset of the networks neurons are activated for a given input. Sparsity provides a form of regularisation that reduces overtraining of the network. ReLU also reduces the problem of a vanishing gradient, which can occur in deep learning models, by retaining a constant gradient for positive inputs. This improves gradient flow during backpropagation and therefore allows for more efficient model training. The ReLU activation function is therefore very successful as it provides an excellent balance between sparsity, non-linearity, efficiency and gradient flow.

The **Hyperbolic Tangent, Tanh**, function is a non-linear function similar to the Sigmoid, as it is also a smooth S-shaped function. The Tanh function however, maps input values to a range between -1 and 1, meaning output values are negative for negative input values and positive for positive input values. The mathematical expression of the function is shown as:

$$f(y) = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}}. \tag{4.14}$$

The Tanh function is very effective at normalising the output of a layer and preventing values from growing too small or too large. This has led to the function being frequently used in the hidden layers of neural networks. It is also used for binary classification problems when the output threshold is at

zero. The Tanh function however does not solve the problem of vanishing gradients.

The **softmax** function is a generalization of the sigmoid function. It is a non-linear, unbounded function that maps input values to a set of output values, between 0 and 1, that sum to 1. In other words the softmax function takes a vector of inputs and normalises them into a probability distribution over the given categories. Each element of the output vector represents the probability of the corresponding category and the sum of all elements in the output vector is equal to 1. The function can be expressed mathematically as:

$$f(\vec{z}_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}, \tag{4.15}$$

where $\vec{z}$ is the input vector, K is the number of classes, e^{z_i} is the standard exponential function for input vector and e^{z_j} is the standard exponential function for output vector. The softmax function is very successful when used to train neural networks on classification problems with multiple categories. One potential problem with the function is that it can produce very small probabilities for some classes when the input values are very large or very small. This can lead to numerical stability issues during model training.

4.5.2 Backpropagation and optimisers

During the training phase of machine learning models, backpropagation and optimiser algorithms are fundamental in minimising the loss function. Backpropagation is used to compute the gradient of the error with respect to the weights of the neural network, while the optimiser is used to update the weights of the network to minimise the error. The training process is an iterative one, with the optimiser updating the model's internal parameters after each iteration until the loss function reaches a minimum.

The backpropagation algorithm is based on the gradient descent optimisation technique. There are two phases involved in the backpropagation

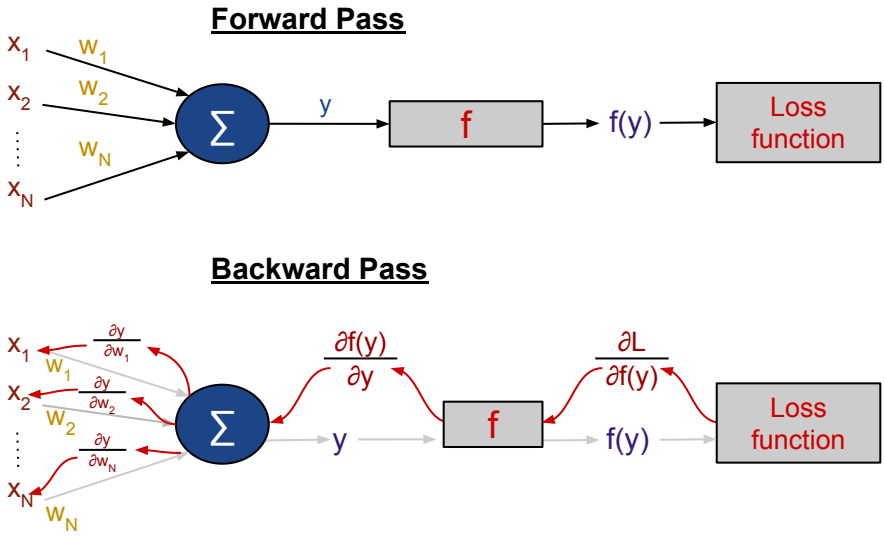


FIGURE 4.7: Diagram of the forward and backward pass of a single node neural network.

algorithm: the forward pass and the backward pass.

The **forward pass** refers to the process of inputs being propagated through the neural network layer by layer, generating an output. The output(s) of each neuron in each layer is calculated by applying an activation function to the weighted sum of inputs, as shown in the forward pass of Figure 4.7 and Equation 4.11. The final component of the forward pass is the loss function which calculates the error by comparing the predicted output of the final layer to the target output.

Considering an example neural network with three layers (input layer, hidden layer and output layer), each containing two neurons, as shown in Figure 4.8. The outputs to the forward pass, $O^{out}_{0|1}$, are calculated iteratively starting from the inputs to the input layer, x_1 and x_2.

The output of each neuron can be calculated using Equation 4.11. The outputs of the input neurons, $I^{out}_{0|1}$, can therefore be calculated as:

$$I^{out}_{0|1} = f_{input}(I^{out}_{0|1}) = f_{input}(x_{0|1}),$$

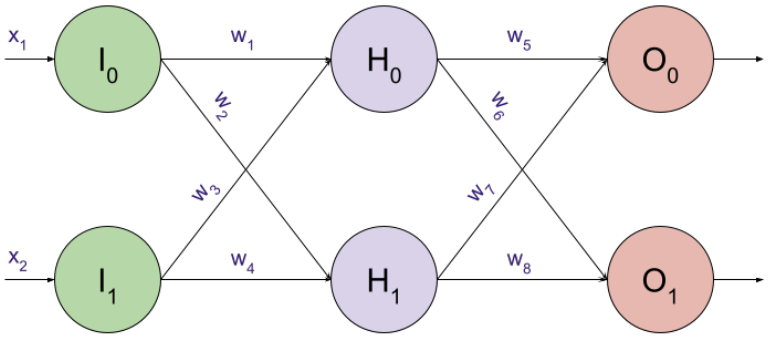


FIGURE 4.8: Forward pass of a neural network with two input nodes, hidden nodes and output nodes.

where f_{input} is the activation function used in the input layer. The input to hidden layer nodes can then be calculated as:

$$H_{0|1}^{out} = f_{hidden}(H_{0|1}^{in}) = f_{hidden}(I_0^{out} \cdot w_{1|2} + I_1^{out} \cdot w_{3|4} + b_0),$$

where f_{hidden} is the hidden layer's activation function and b_0 is the bias connected to the hidden layer neurons. Finally, the outputs of the output layer's neurons can be calculated as:

$$O_{0|1}^{out} = f_{output}(O_{0|1}^{in}) = f_{output}(H_0^{out} \cdot w_{5|6} + H_1^{out} \cdot w_{7|8} + b_1),$$

where f_{output} is the output layer's activation function and b_1 is the bias connected to the neurons in the output layer. Using a loss equation with the given outputs and targets for each output neuron, we can therefore calculate the total error, E_{total}, as the sum of loss calculated for each output, $O_{0|1}^{out}$.

The **backward pass** of the backpropagation algorithm is the mechanism of using the loss from the forward pass to inform changes to neuron weights in order to reduce the loss and therefore improve the accuracy. This is achieved by propagating the loss backwards through the network and calculating the gradient of the loss with respect to the weights of the network, as shown in Figure 4.7. A neural network can be understood as a composite function. The derivative of composite functions, in this case individual gradients, can

be calculated using the Chain Rule. This means that the Chain Rule can be used to allocate the total loss to the various layers of the neural network.

Considering the example shown in Figure 4.8, the backwards pass calculates the gradients with respect to each weight, working backwards from the loss function. Starting with the weights preceding the output layer, $w_{5|6|7|8}$, the extent to which each weight effects the total loss is calculated. The value of w_5 or w_7 is calculated as:

$$\frac{\partial E_{Total}}{\partial w_{5|7}} = \frac{\partial E_{Total}}{\partial O_0^{out}} \cdot \frac{\partial O_0^{out}}{\partial O_0^{in}} \cdot \frac{\partial O_0^{in}}{\partial w_{5|7}}. \tag{4.16}$$

The process can be repeated using the second output neuron, O_1, to calculate the gradients for w_6 and w_8. Calculating the gradients of the weights preceding the hidden layer ($w_{1|2|3|4}$), follows the same process as for $w_{5|6|7|8}$ but must take into account of the fact that the output of each hidden layer neuron contributes to the output (and therefore error) of multiple output neurons. Therefore, considering the weights connected to H_0 the calculation is:

$$\frac{\partial E_{Total}}{\partial w_{1|3}} = \frac{\partial E_{Total}}{\partial H_0^{out}} \cdot \frac{\partial H_0^{out}}{\partial H_0^{in}} \cdot \frac{\partial H_0^{in}}{\partial w_{1|3}}, \tag{4.17}$$

where $\frac{\partial E_{Total}}{\partial H_0^{out}}$ must take into consideration both output neurons. This can be written as:

$$\frac{\partial E_{Total}}{\partial H_0^{out}} = \frac{\partial E_{O_0}}{\partial H_0^{out}} + \frac{\partial E_{O_1}}{\partial H_0^{out}}. \tag{4.18}$$

In order to calculate the gradients of weights preceding H_1, Equation 4.17 can be written as:

$$\frac{\partial E_{Total}}{\partial w_{2|4}} = (\frac{\partial E_{O_0}}{\partial H_1^{out}} + \frac{\partial E_{O_1}}{\partial H_1^{out}}) \cdot \frac{\partial H_1^{out}}{\partial H_1^{in}} \cdot \frac{\partial H_1^{in}}{\partial w_{2|4}}. \tag{4.19}$$

Optimisers take the gradients of each weight with respect to the loss and use it to update the value of each weight in order to improve the accuracy (reduce the loss) of the model. There are different types of optimisers, including gradient descent, stochastic gradient descent, RMSprop and Adam, each with their own unique strengths and weaknesses. Selecting the right optimiser can have a significant impact on the performance of a model.

The **gradient descent** method can be considered the base optimiser utilised. The updated weights are calculated using gradient descent as:

$$w_{t+1}^n = w_t^n - \eta \cdot \frac{\partial E_{Total}}{\partial w_t^n}, \tag{4.20}$$

where w_t^n is the current value of weight n for iteration t, η is the learning rate, and E_{Total} is the average loss for a batch of training events. The learning rate is therefore an important hyperparameter as it determines the "step-size" of the optimisation. An alternative version of this optimiser is the **stochastic gradient descent**, SGD, method. The SGD in contrast performs an update to the weights for each training event. This causes the SGD method to fluctuate more than the standard gradient descent before converging.

Other optimisers consider the momentum, in the form of previous updates to each weight, in their calculation. This provides a form of adaptive learning rates corresponding to each internal parameter. The **RMSprop** method demonstrates this by dividing the gradient by a running average of its recent magnitude. The update using RMSprop is calculated as:

$$w_{t+1}^n = w_t^n - \frac{\eta}{\sqrt{E[g^2]_t^n + \epsilon}} g_t^n, \tag{4.21}$$

where $\sqrt{E[g^2]_t^n + \epsilon}$ is the root mean squared, RMS, error of the gradient, g_t^n, and ϵ is a smoothing term. The running average of the gradient, $E[g^2]_t^n$, is calculated as:

$$E[g^2]_t^n = \gamma \cdot E[g^2]_{t-1}^n + (1-\gamma) \cdot g_t^2, \tag{4.22}$$

where γ is an update factor, usually equal to 0.9.

The **Adaptive Moment Estimation, Adam**, is an optimiser that computes adaptive learning rates for each weight. This method uses the decaying average of the past gradients m_t, similar to momentum, together with the decaying average of the past squared gradients, v_t, similar to RMSprop. The Adam optimisation step is calculated as:

$$w_{t+1}^n = w_t^n - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \cdot \hat{m}_t, \tag{4.23}$$

where the $\hat{m}_t$ and $\hat{v}_t$ components are calculated as:

$$\hat{m}_t = \frac{\beta_1 \cdot m_{t-1} + (1-\beta_1) g_t}{1-\beta_1} \quad \text{and} \quad \hat{v}_t = \frac{\beta_2 \cdot v_{t-1} + (1-\beta_2) g_t^2}{1-\beta_2}, \tag{4.24}$$

where β_1 and β_2 are hyperparameter constants, usually equal to 0.9 and 0.999 respectively.

4.5.3 The Multilayer perceptron

The Multilayer Perceptron (MLP) is an extended version of the perceptron model, which was developed by Frank Rosenblatt in 1957 [102]. The perceptron is a linear classifier capable of classifying data into two categories. The model uses a learning algorithm to adjust the weights at each neuron edge based on misclassification, allowing the model to learn from data. This model therefore became the foundation of the majority of neural network models developed today. Currently, more than 50% of neural networks used today are still MLPs.

The multilayer perceptron is a feed-forward artificial neural network, meaning that connections between nodes do not form a cycle. Therefore, data being input into the model, for training or inference, is processed linearly, starting at the input layer and sequentially moving through each of the hidden layers to the output layer. An example feed forward structure of the MLP is shown in Figure 4.9.

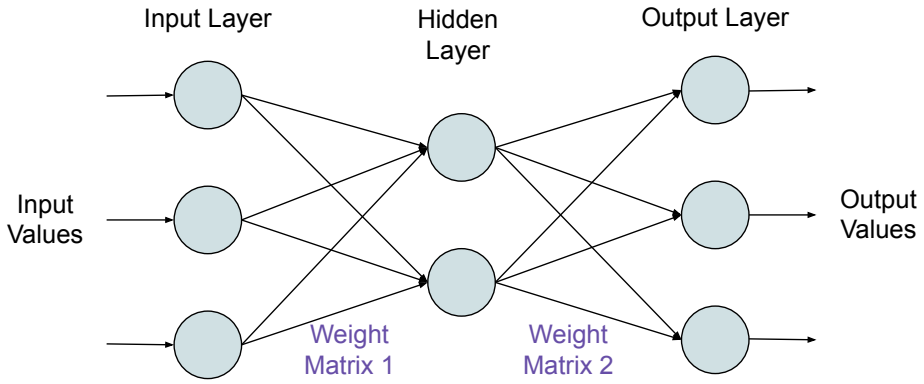


FIGURE 4.9: Diagrammatic example of multilayer perceptron architecture.

The MLP structure consists of at least three layers of nodes: an input layer, hidden layer(s) and an output layer. The input layer consists of inactive neurons and is often used to normalise the inputs. The hidden layers have active neurons which are usually activated using the sigmoid function. The neurons in the output layer are linear and use the sigmoid activation forcing the model output to be a value in the range $(0, 1)$. Each neuron in a layer following the input layer, computes a linear combination of the outputs of the previous layer. The output of the neuron is thus a function of that combination with specified activation function for the given layer. As a linear combination of sigmoids can approximate any continuous function, and the probability of training data conforming to each given category is known, the MLP is able to learn to classify data with excellent success.

The MLP model is trained using the backpropagation algorithm, to modify its weights and minimise the error of its predictions. This process is repeated for many iterations (or epochs) until the network's predictions are as accurate as possible.

There are a wide range of applications for MLPs in particle physics. They are particularly useful for pattern recognition tasks, such as identifying particles or reconstructing events. This includes the application of interest, where an MLP is trained to recognise specific patterns of data that correspond to different types of particles.

Despite their power and versatility, MLPs do have certain limitations. They can suffer from overfitting, a problem where the network learns the training data too well and performs poorly on unseen data. Techniques like regularisation, dropout, and early stopping are used to help mitigate this. MLPs have difficulty dealing with temporal data, as they lack an inherent way to deal with time-series data. Finally, MLPs require a large amount of labelled training data, which can be expensive and time-consuming to produce. Despite these challenges, the multilayer perceptron remains a cornerstone of machine learning and a powerful tool for classifying data.

4.5.4 Deep neural networks

Deep Neural Networks, DNNs, represent the evolution of MLPs to handle more complex problems with higher levels of abstraction [103]. The term "deep" refers to the number of layers in the network, typically implying three or more hidden layers. Deep learning models, as opposed to MLPs, therefore have more complex structures with an unbounded number of layers of bounded size. This allows the model to retain theoretical universality, under given conditions, while encouraging optimisation and practical implementation [104].

The hidden layers of a DNN allow the network to learn features from the data in a hierarchical manner. Early layers learn to detect simple patterns, while subsequent layers combine these simple patterns to detect more complex patterns. These layers can be heterogeneous and can therefore differ from biologically informed models. Deep learning models can have tens or even thousands of hidden layers and thus contain a very large number of internal parameters, allowing the complex assemblage of layers to perform complicated high dimensional tasks. A generalised DNN model structure is shown in Figure 4.10.

The DNN model, like the MLP, uses non-linear activation functions. The activation functions used in DNNs have a consequential impact on the network's performance and ability to converge during training. In order to help mitigate issues, such as vanishing gradients during the training of the network, the ReLU activation function and its variants are commonly used.

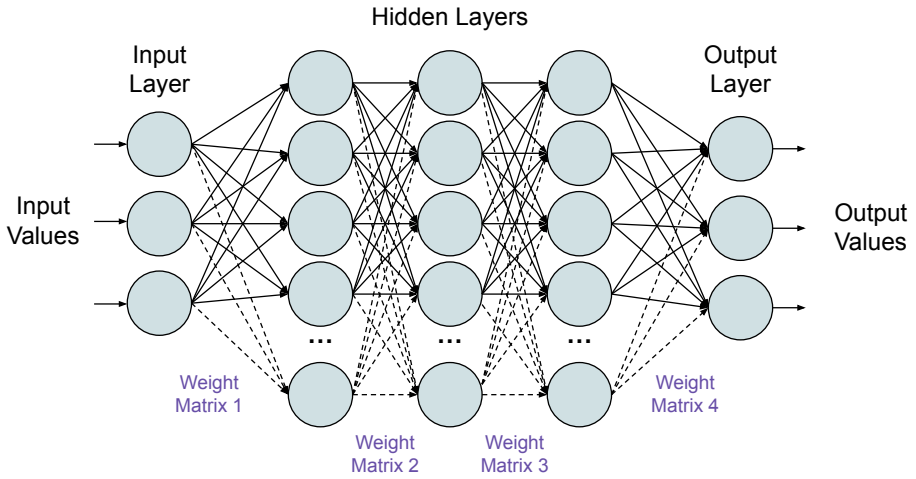


FIGURE 4.10: Diagrammatic example of deep neural network architecture with three hidden layers.

The training of a DNN involves the same backpropagation algorithm as other neural networks. However, training DNNs can be challenging when confronted with issues such as the vanishing gradient problem. In order to overcome this, more sophisticated optimization algorithms like Adam and RMSProp are commonly used.

DNNs have demonstrated remarkable success in a wide range of particle physics applications. They are especially useful for tasks that involve high-dimensional data or require the extraction of complex features. The primary use of DNNs in this research, and one of its most successful use cases, is as a classifier. DNNs can be trained to classify collision events based on their properties as well as identify different types of particles based on their interaction within the detector. These applications can be extended to improving the precision of measurements and sensitivity of analysis by learning complex relationships between variables and reducing systematic uncertainties.

Despite their abilities, DNNs have certain limitations. These include the fact that they require large quantities of data and computational resources to train. The "black box" nature of DNNs, created by the large number of internal weights, can make interpreting the models decision-making process difficult. They are also susceptible to over-fitting and must be carefully

trained and evaluated. In particle physics applications, an import requirement is that the model obeys the laws of physics. This means that the model must be invariant to certain transformations and symmetries inherent in the data. Various techniques are used to address these limitations such as regularisation between layers, early stopping and dropout.

Although there are limitations to DNN models, they provide a powerful tool for complex pattern recognition and data analysis problems in particle physics. Due to ongoing research, the DNNs capabilities are being expanded, and its limitations are being addressed.

4.6 Monte Carlo and Generative Models

Data generation is an essential part of particle physics research and machine learning. The LHC, introduced in Chapter 3, generates an immense amount of data with each particle collision event. However, it is not feasible to store all of this data due to technological and financial constraints. Thus, there is a need to selectively record and analyse only the most promising events. Simulated data comes into play, aiding in the design of the triggers and filters that decide which events to keep for further analysis. Moreover, accurate and efficient simulation of these complex physical processes is crucial for interpreting the results of the LHC experiments. Comparisons between experimental data, produced at the LHC, and data simulated using theoretical models, allows for validation and/or exposes the limitations of the given theoretical model. As a prime example, the discovery of the Higgs boson relied heavily on the comparison of the observed data with high-precision Monte Carlo simulations.

However, these detailed simulations are computationally expensive. More recent machine learning based techniques promise to accelerate this process dramatically while preserving the necessary details of the data. They also have the potential to tackle the challenges posed by the planned upgrades to the LHC, which will increase the collision rate and, consequently, the complexity and volume of data [105].

Data generation can play a crucial role in machine learning. It is used in the training, validation, testing, and ultimately in the development of robust models that can predict, classify, and recognise patterns effectively.

In this section, we will delve into the methods and techniques employed for data generation. Firstly focusing on Monte Carlo event generation methods. Before exploring the more recent Kernel Density Estimation (KDE), Generative Adversarial Networks (GANs), and Wasserstein GANs (WGANs) machine learning based data generation models.

4.6.1 Monte Carlo event generation

The Monte Carlo, MC, method, named after the renowned casino due to its inherent randomness, is a broadly used technique in physics and has gained traction in the machine learning community [106]. This approach excels at simulating complex systems and generating data, and has been crucial in the domain of particle physics where it is used to simulate the interactions of particles [107].

These simulations are vital for interpreting the results of the LHC experiments, as they allow researchers to compare observed data with theoretical predictions. MC simulations generate events representing possible outcomes of particle collisions. These events are built up from the initial hard scattering process, where the high-energy collision occurs, followed by the parton shower, hadronisation, and finally, the decay of unstable particles. Each of these stages involves complex physical processes, and MC methods are used to model these stochastic processes in a probabilistic way. Several general-purpose event generators employ MC methods to simulate high-energy physics events at the LHC, including Pythia, Herwig, Sherpa, and MadGraph:

Pythia is one of the most widely used event generators. It provides a broad range of processes for event generation, with a strong focus on the description of high-energy collisions involving heavy quarks and jets [108].

Herwig, on the other hand, emphasises the soft and non-perturbative aspects of event generation, including multiple parton interactions, underlying events, and beam remnants [109].

Sherpa is a versatile event generator designed to handle a wide variety of processes and experimental setups. It features a flexible matrix element generator and a comprehensive set of parton shower algorithms [110].

MadGraph is an essential tool in particle physics simulations that complement the use of MC methods. MadGraph is a program for generating high-energy collision events according to theoretical models. It uses Feynman diagrams, which represent the mathematical expressions describing the behaviour of subatomic particles, to generate matrix elements for high-energy physics processes [111]. The generated matrix elements can then be interfaced with event generators like Pythia and Herwig, which add the effects of parton showering, hadronisation, and multiple parton interactions to produce a more realistic simulation of collision events. MadGraph is particularly useful for new physics searches, as it allows users to implement their theoretical models and generate events according to these models. This capability has become increasingly important with ongoing searches for new physics beyond the Standard Model at the LHC.

4.6.2 Kernel density estimation

Kernel Density Estimation (KDE) is a non-parametric approach for estimating the probability density function of a random variable [112, 113]. It excels in creating smooth approximations of data distributions, a valuable asset for complex, high-dimensional datasets like those in particle physics. As an unsupervised machine learning technique, KDE learns patterns and structures directly from training data, without predefined models. Once trained, it can generate new, large-scale event data, mirroring the learned patterns from its training set, offering a distinct advantage over traditional Monte Carlo (MC) methods.

KDE model training and data generation consists of three parts; fitting, density estimation and data sampling:

1. **Fitting**: The first process consists of fitting a kernel, $K(x;h)$, to each data point of each feature of the training data, x. A kernel is a positive function which is controlled by the bandwidth parameter, h. The kernel function used is usually Gaussian but can be exponential, linear, cosine or even a tophat function. The bandwidth parameter controls the width of these kernels and therefore significantly affects the resulting estimate.

2. **Density Estimation**: The second process combines all the individual kernels together to form a single smooth estimate of the data's probability density function. This function describes the underlying distribution of the data. Given a kernel of the form $K(x;h)$, the density estimate, ρ_K, for a single point, y, within a group of points, x_i, is calculated as:
$$\rho_K(y) = \sum_{i=1}^{N} K(y - x_i; h). \tag{4.25}$$

3. **Data Sampling**: To generate new data, also known as sampling, the estimated density function is utilised. This function, when sampled randomly, produces data points more frequently in areas of higher density (where data points are densely packed) and less so in areas of lower density (where data points are sparse). Consequently, by sampling numerous events, a synthetic dataset encompassing thousands of entries can be created.

In the case where the training dataset is multivariate, a multivariate kernel is used. This is essentially a higher-dimensional version of the univariate kernel. For example, a Gaussian kernel in two-dimensional data is bell-shaped. The same kernel in three-dimensions becomes a bell-shaped volume. Similarly, the bandwidth parameter in the multivariate case becomes a bandwidth matrix rather than a single value. The diagonal elements of this matrix control the width of the kernel in each dimension, and the off-diagonal elements control the correlation between dimensions. This increased complexity in higher dimensional data, increases the computational intensity of the problem.

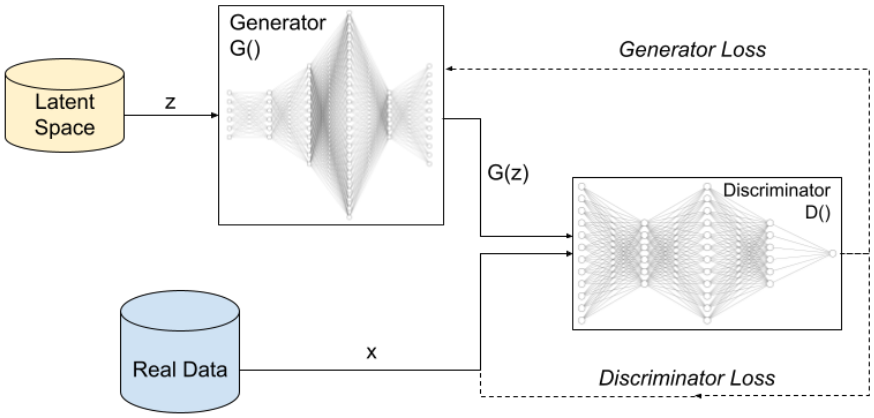


FIGURE 4.11: Systematic diagram of a generative adversarial network.

4.6.3 Generative adversarial network

The Generative Adversarial Network, GAN, is a neural network based data generator that was first introduced in 2014 [114]. Since its inception the GAN has emerged as a powerful class of generative models in machine learning. The GAN is a deep-learning-based model that is trained to understand complex patterns in a dataset and generate new realistic samples [115–117]. GAN architecture consists of two neural networks. Namely, the generator, G, and discriminator, D, networks which are engaged in a game-theoretic framework. The generator network produces synthetic data instances by mapping a source of noise, z, to the input space. The discriminator evaluates the authenticity of both real, x, and synthetic, $G(z)$, data instances and learns to distinguish between the two. The goal of the generator is to produce data instances that the discriminator cannot distinguish from real instances, and the goal of the discriminator is to correctly classify real and synthetic instances. The GAN architecture and interactions are shown in Figure 4.11.

The loss function of the GAN can thus be defined as follows. Firstly to understand the generator's distribution, p_g, over the data, x, a prior on the input noise variable can be defined as $p_z(z)$. The generator's mapping to the data space can then be represented as $G(z;\theta_g)$, where θ_g is the internal parameters of the generator neural network. Similarly, the discriminator

model can be represented as $D(x;\theta_d)$, where θ_d is the internal parameters of the discriminator neural network. The output of the discriminator, $D(x)$ or $D(G(z))$, is a single scalar value representing the probability that x or $G(z)$ is from the real data. The training of the GAN is therefore the simultaneous maximising of the discriminator output, to correctly label training and generated samples, and the minimising the generator output with respect to the discriminator's ability to differentiate it. This can be expressed as:

$$\min_G \max_D L(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))], \tag{4.26}$$

where $\mathbb{E}$ is the expected values for multiple data points. This expression is derived directly from the binary cross-entropy loss, Equation 4.1, which is the loss function used by both the generator and discriminator. The discriminator network is trained as a fully-supervised model. During its training, the generator's weights and biases are fixed. The discriminator network is therefore penalised when it incorrectly predicts that a generated sample is real or real sample is generated. The discriminator loss function, L_D, can therefore be expressed as:

$$L_D = \max\left[\mathbb{E}_{x\sim p_{data}(x)}[\log(D(x))] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]\right]. \tag{4.27}$$

On the other hand, the generator is an unsupervised deep neural network. During the generator training iterations, the discriminator network's internal weights and biases are kept fixed. The first term in Equation 4.26, becomes zero and the generator's loss function, L_G, can therefore be written as:

$$L_G = \min\ \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))]. \tag{4.28}$$

The training of the GAN is an iterative process. During each training iteration, or epoch, updates are made to both the generator and discriminator

networks. A common strategy is to update each network once per iteration. However, as the system is limited by the discriminator's ability to classify samples, it has become standard practice to update the discriminator network multiple times for every generator update. This allows the GAN to be trained more quickly and provides more meaningful gradients for the generator to learn from.

It is important to note that the balance between generator and discriminator training is delicate. If the discriminator becomes too powerful, it may fail to provide useful feedback to the generator, causing the generator's learning to stagnate. If the generator becomes too powerful, it might consistently fool the discriminator, which also results in suboptimal learning.

Common limitations that must be considered when training a GAN include the issue of vanishing gradients, which can lead to the generator training to fail. Additionally, the issue of modal collapse, where the generator produces samples with limited diversity or even the same sample, regardless of the generator inputs. This means that the generator collapses to a few modes of the training distribution and ignores the rest. Another issue is the potential failure to converge due to the two networks falling into a cycle where they continuously try to outsmart each other, leading to oscillation or divergence of the model parameters. In order to address some of these limitations, more advanced method such as using the Wasserstein loss mechanism and applying penalties to the gradients can be used.

4.6.4 Wasserstein generative adversarial network with gradient penalty

The Wasserstein Generative Adversarial Network with gradient penalty, WGAN-gp, is a version of the GAN which adopts a Wasserstein distance mechanism, in its loss function, and a gradient penalty, to improve the model's ability to converge, as well as addressing vanishing gradients [118]. The WGAN-gp model replaces the discriminator network of the GAN with a critic network, C, which uses a gradient penalty, GP, to control its gradients. The model's training mechanism therefore involves the playoff between the generator trying to produce samples that the critic is unable to

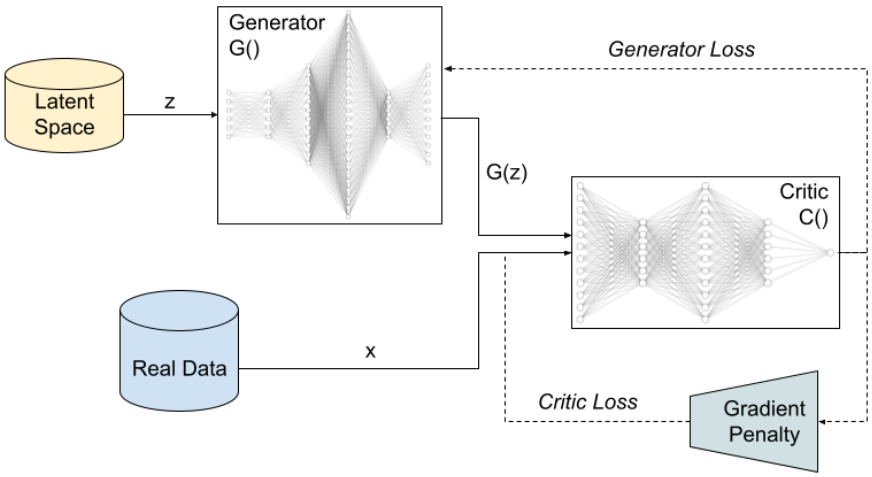


FIGURE 4.12: Systematic diagram of Wasserstein generative adversarial network with gradient penalty.

differentiate from the real data, while the critic is being trained to discriminate between real and generated events. An overview of the WGAN with gradient penalty is shown in Figure 4.12.

Wasserstein GANs adopt the Wasserstein distance, $W(q, p)$, which can be defined as the minimum cost of transporting mass in order to transform the distribution q into the distribution p. The discriminator is replaced in the improved model with a critic, C, and the gradients are controlled using a gradient penalty, GP, which penalises the norm of the critic gradients with respect to the input. The generator loss, L_g, function is defined as:

$$L_g = \min \mathop{\mathbb{E}}_{\tilde{x} \sim \mathbb{P}_g} [C(\tilde{x})], \tag{4.29}$$

where $\tilde{x} = G(z)$ and z is the latent space noise and $\mathbb{P}_g$ is the generator model distribution, $\tilde{x}$. The critic loss, L_c, with gradient penalty is defined as:

$$L_c = \max \mathop{\mathbb{E}}_{x \sim \mathbb{P}_r} [C(x)] - \mathop{\mathbb{E}}_{\tilde{x} \sim \mathbb{P}_g} [C(\tilde{x})] + \lambda * GP, \tag{4.30}$$

where λ is the gradient penalty coefficient, $\mathbb{P}_r$ is the MC data distribution, and the gradient penalty, GP, is defined as:

$$GP = \mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_{\tilde{x}}}[(||\nabla_{\tilde{x}}C(\tilde{x})||_2 - 1)^2]. \tag{4.31}$$

This WGAN model structure takes advantage of neural networks to teach itself the intricacies of a dataset and generate realistic, high quality events. The ability of these machine learning based data generators has much potential in HEP studies and must be further investigated.

Chapter 5

Di-Lepton Final State Classifier Supervision Study

5.1 Introduction

Semi-supervised machine learning classifiers provide an innovative method to reduce model dependency in the classification of particle signatures at the LHC. These state-of-the-art techniques, explored in Section 4.3.3, and elaborated upon in Ref. [90], introduce their capabilities in the realm of high-energy physics. However, to evaluate the full potential of these techniques, further investigation is necessary. To delve deeper into the effectiveness of semi-supervised methods, this chapter embarks on a comparative analysis with fully supervised classifiers. Focusing on three leading classification models in particle physics, the Boosted Decision Tree (BDT), Multi-Layer Perceptron (MLP), and Deep Neural Network (DNN), the research contrasts their performance under full and semi-supervised training regimes. This comparison aims to shed light on the practical advantages and limitations of semi-supervised learning in this advanced field.

Several studies from the ATLAS and CMS experiments have revealed notable deviations in multi-lepton final state data compared to SM expectations, as detailed in Refs. [29–32, 36]. These anomalies, suggestive of potential BSM physics, align well with interpretations offered by the $2HDM + S$ model, as detailed in Section 2.3. The observed discrepancies, described by this model, underscore the importance of investigating the $H \to Sh, SS^*, SS'$

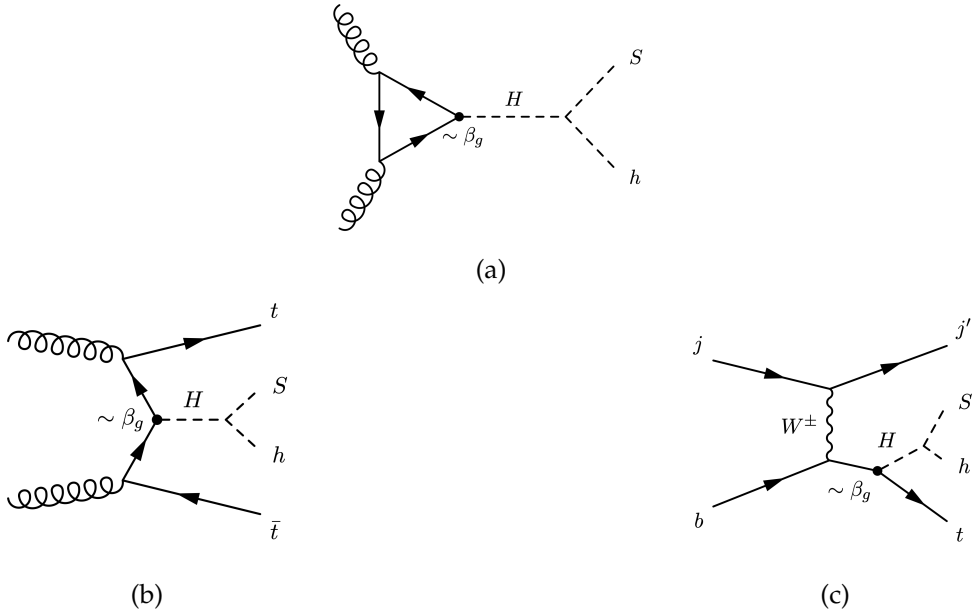


FIGURE 5.1: Feynman diagrams showing leading order production modes for H decaying into Sh for (a) gluon gluon fusion, ggF, (b) top pair associated production, ttH, and (c) single top associated production, tH [34].

decay processes, particularly in a final state comprising leptons of differing flavours and opposite charges: $e^{\pm}\mu^{\mp}$. This specific final state is predominantly influenced by the $t\bar{t}$ process, known for its significant cross-section, thus providing a well constrained data sample rich in statistics conducive for a rigorous analysis. Such a high-statistics environment is ideal for assessing the efficacy of advanced machine learning techniques, especially in distinguishing signal from background processes in $H \rightarrow Sh, SS^*, SS'$ searches. Consequently, the di-lepton dataset is shown to provide well constrained signal and background samples, ideal for the evaluation of the semi-supervised machine learning technique.

The study concentrates on the $H \rightarrow Sh$ final states, with production modes illustrated in Figure 5.1. These include the three main production mechanisms: gluon fusion, top pair associated production and single top associated production.

In this Chapter, a di-lepton study together with a base analysis of semi-supervised techniques are presented. Firstly, the di-lepton MC dataset used

in this study is introduced, the signal and background regions of interest are defined, and the selected kinematics are presented. Following this, the methodology and resultant comparison of the semi- and fully supervised techniques, using the BDT, MLP and DNN classifier models, are presented. This analysis forms a baseline study for the efficacy of the semi-supervised technique in classifying high energy particle signatures.

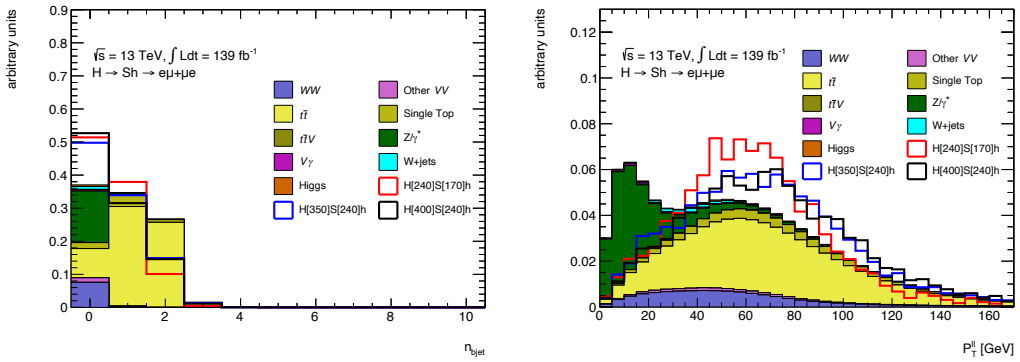
5.2 Di-Lepton Final State Dataset

5.2.1 Monte Carlo Dataset

The di-lepton final state Monte Carlo dataset is made up of an Electron, e^-, and a Muon, μ^+. This lepton pair has a transverse momentum greater that 10 GeV, opposite charge and different flavour ($e\mu + \mu e$). The study was conducted in context of three different processes being examined that produce the di-lepton signal dataset, $H[240] \rightarrow S[170]h$, $H[350] \rightarrow S[240]h$ and $H[400] \rightarrow S[240]h$. Each provides a signal dataset of the di-lepton system. The background samples are a combination of processes that include Higgs, VV, $V\gamma$, $\bar{t}t$, tV/ttV and $W + Jets$. This provides a good representation of processes occurring simultaneously to that of the di-leptons and therefore are viable background processes when classifying the di-lepton system.

The MC data is generated with PYTHIA8 [108] using the NNPDF 2.3 leading order [119] for parton showering, with the A14 tune [120], and without considering detector effects. While the parameters of the model are fixed, the kinematics of the final state are presented with $m_H = 250\,\text{GeV}$ and $m_H = 260\,\text{GeV}$, in addition to the nominal value. The $H \rightarrow Sh$ samples are generated including WW, ZZ, $\tau\tau$ and $\gamma\gamma$ decay modes for the S and h bosons to obtain the relevant final states with leptons, photons and jets for this study. Finally, the SM Vh events are generated for each Higgs boson decay mode of interest separately.

The dataset includes variables of transverse energy, E_T, missing transverse energy, E_T^{miss}, pseudo-rapidity, η, transverse mass, m_T, di-lepton invariant



(a) (b)

FIGURE 5.2: Distributions of the number of b-tagged jets (a) and di-lepton transverse momentum after at most one b-tagged jet requirement (b) for the SM background processes and several $H \to Sh$ signals normalised to unity.

mass, $m_{\ell\ell}$, transverse momentum, p_T, and weights. The data provides variables for both leading and sub-leading leptons as well as leading and sub-leading jets.

5.2.2 Signal and Background Definitions

The $H \to Sh$ search strategy relies on the number of b-tagged jets to define both the signal and background (a top enriched region) regions. This is exposed in Figure 5.2a which compares the b-tagged jet multiplicity in $e^{\pm}\mu^{\mp}$ events for signal and SM background processes from MC simulation. Events with at most one b-tagged jet include approximately 90% of the total $H \to Sh$ signal, and thus are selected to form the signal region. In addition, candidate events are required to satisfy a lower threshold on the di-lepton transverse momentum: $p_{\mathrm{T}}^{\ell\ell} > 30\,\mathrm{GeV}$ to reject the Drell-Yan process, a mechanism where a quark from one proton and an antiquark from another proton annihilate to produce a lepton pair, as shown in Figure 5.2b. On the other hand, events with exactly two b-tagged jets are dominated by $t\bar{t}$ process and have a small signal contamination. This feature is therefore used to define a top enriched region to validate the MC simulation and from which to extrapolate the $t\bar{t}$ contribution to the signal region.

The extent of separation for several kinematic distributions, between the

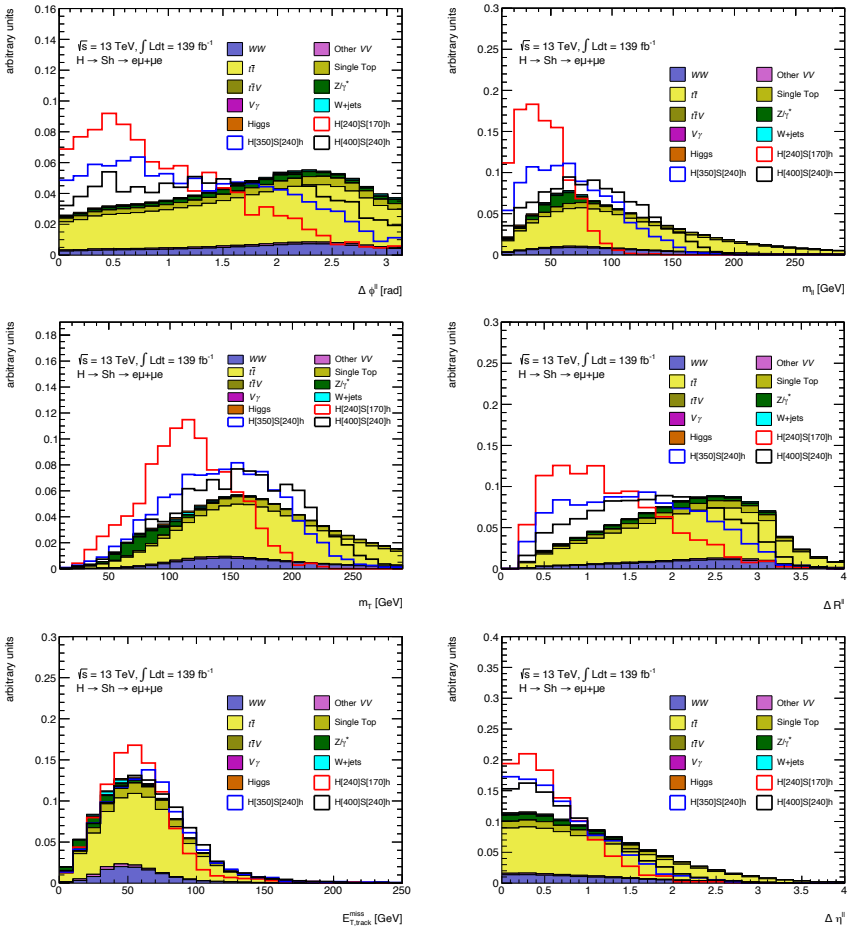


FIGURE 5.3: Kinematic distributions in the signal region from MC simulation for several $H \rightarrow Sh$ signal and SM background processes normalised to unity.

signal and SM background processes for simulated events in the signal region, is illustrated in Figure 5.3.

It can be extrapolated from Figure 5.3 that the $H[240] \rightarrow S[170]h$ signal sample has the largest separation power and therefore is selected as the signal sample for this study.

The background sample is defined as a top control region (CR), a signal free region enriched in top processes. This region can be defined using the invariant mass of the two b-tagged jets (m_{bb}) for the $H \rightarrow Sh$ signal and the SM processes from simulation, as shown in Figure 5.4. By selecting top CR

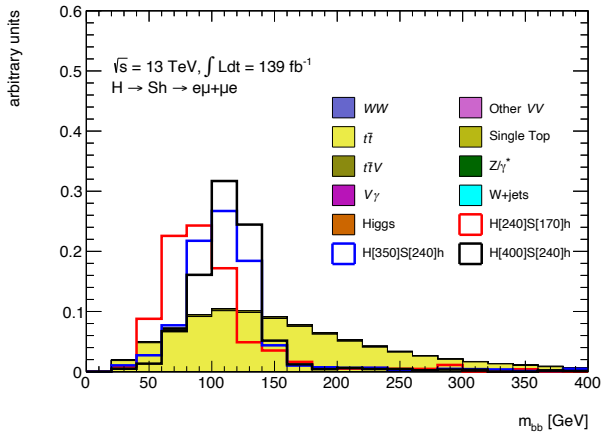


FIGURE 5.4: Distribution of the invariant mass of the two b-tagged jets for the SM background processes and several $H \to Sh$ signals normalised to unity.

events with at least two b-tagged jets and $m_{bb} > 150\,\text{GeV}$, orthogonality between the top CR and the signal region is ensured. The latter threshold on m_{bb} further rejects the remaining $H \to Sh$ signal contamination in events with two b-tagged jets. This provides the ideal background sample to evaluate machine learning classifiers.

Bound-state effects on top quark production

Several top-quark measurements by the ATLAS and CMS experiments result in an enhanced top-quark cross-section production [121]. In particular several top-quark kinematic distributions, including the di-lepton invariant mass, $m_{\ell\ell}$, and the angular separation between the leptons, $\Delta\phi_{\ell\ell}$, present shape differences at low $m_{\ell\ell}$ and $\Delta\phi_{\ell\ell}$ regions. These discrepancies could be interpreted as bound-state effects on top-quark production [122]. In light of this, it is crucial to evaluate the possible implications on the $t\bar{t}$ bound-state effects in the $H \to Sh$ search.

Figure 5.5 shows the $t\bar{t}$ invariant mass, $m_{t\bar{t}}$, distribution and compares different $m_{\ell\ell}$ spectra in four $m_{t\bar{t}}$ regions for the $e^{\pm}\mu^{\mp}$ final state using $t\bar{t}$ process from MC simulation. For the latter, there is a clear correlation between the mean of the $m_{\ell\ell}$ distributions and the range of the $m_{t\bar{t}}$ regions.

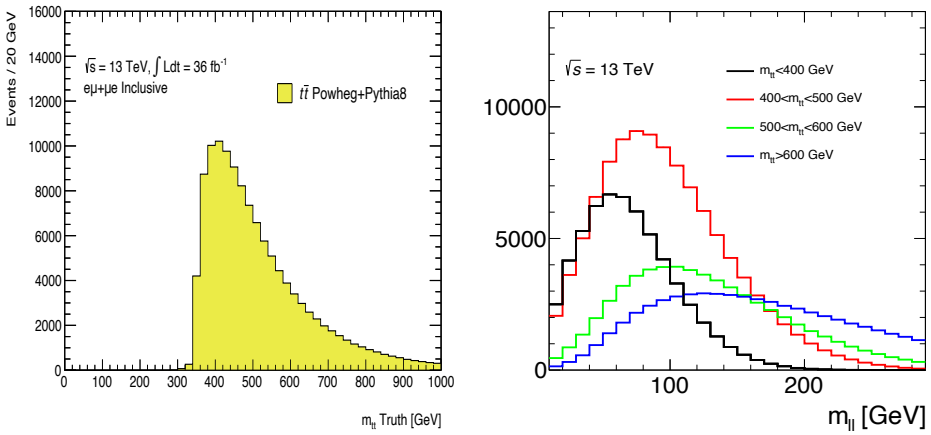


FIGURE 5.5: Distributions of the $t\bar{t}$ invariant mass (left) and di-lepton invariant mass for different $m_{t\bar{t}}$ regions (right) from $t\bar{t}$ MC simulation.

Figure 5.6 compares the number of b-tagged jets in different $m_{t\bar{t}}$ regions for the $t\bar{t}$ process from MC simulation. Three thresholds for the b-tagged jet transverse momentum, p_{T}^{b-jet}, have been considered: 20 GeV, 25 GeV and 30 GeV. The lower pad in these distributions shows the ratio in each b-tagged jet bin for the region with $m_{t\bar{t}} < 400$ GeV over the other $m_{t\bar{t}}$ regions. The highest discrepancy is observed for events with no b-tagged jets. The observed differences are below 20%. The impact on the $t\bar{t}$ bound-state effects in the $H \to Sh$ search is expected to only be around a few percent since the $t\bar{t}$ contribution in the 0 b-tagged jet bin represents around 10% of the total background, as shown in Figure 5.4.

5.2.3 Kinematic Feature Selection and Cut Summary

Before the di-lepton dataset can be used for analysis, the Top control region must be analysed as the base background sample. The cuts required to achieve this, discussed above, can be summarised as:

1. $N_{bj} \geq 2$,
2. $m_{bb} > 150\,GeV$.

It is shown in Figure 5.7, that by making these cuts we are able to extract a sample dominated by the $t\bar{t}$ background process as desired.

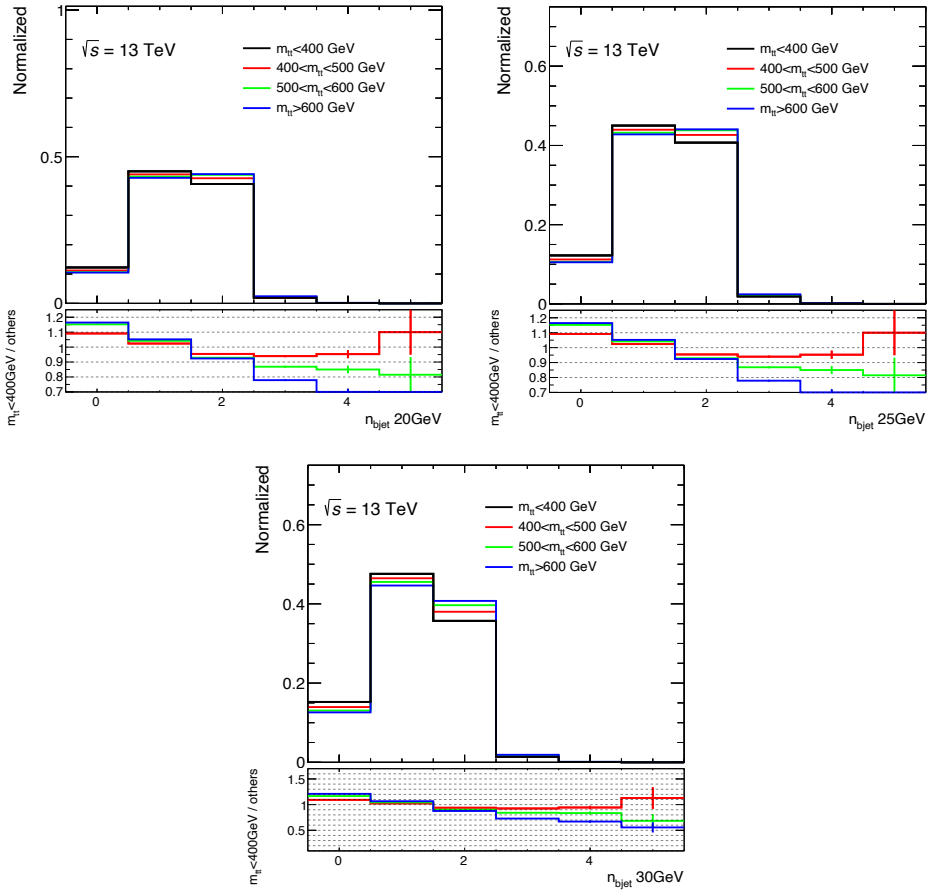
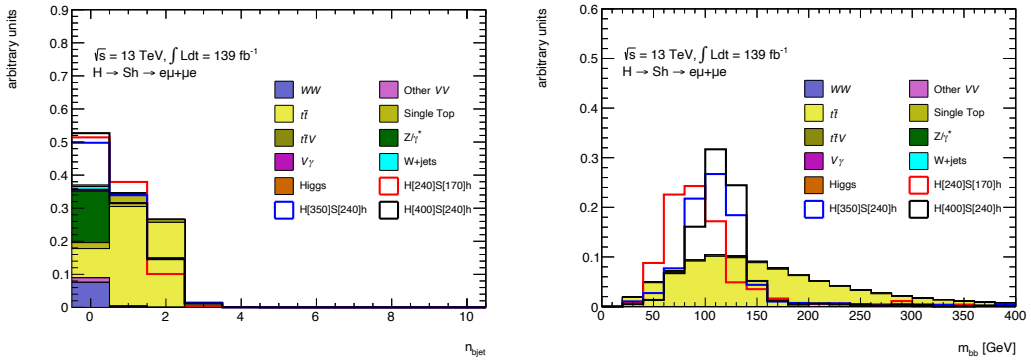


FIGURE 5.6: Distributions of the number of b-tagged jets in different $m_{t\bar{t}}$ bins from $t\bar{t}$ MC process using different b-tagged jet p_{T} thresholds: 20 GeV (top left), 25 GeV (top right) and 30 GeV (bottom). Distributions are normalised to unity for easier comparison.

Optimal kinematic variables for the di-lepton system are selected to be used for machine learning training. These include the invariant mass, $m_{\ell\ell}$, transverse mass, m_T, transverse momentum of the leading and sub-leading leptons, $p_{T_{\ell 0}}$ and $p_{T_{\ell 1}}$, transverse momentum of the di-lepton system, p_T, missing transverse energy, E_T^{miss}, the difference in the azimuthal angle between the two leptons, $\Delta\phi_{\ell\ell}$, the difference in the pseudo-rapidity between the two leptons, $\Delta\eta_{\ell\ell}$, and the distance in the $\eta - \phi$ space between the two leptons, $\Delta R_{\ell\ell}$.



(a) (b)

FIGURE 5.7: Di-lepton top control region cuts. Distribution of the number of b-tagged jets (a) before at least two b-tagged jet requirement and the distribution of invariant mass of the two b-tagged jets (b) before applying requirement of m_{bb} being greater than 150 GeV.

5.3 Methodology

To assess the efficacy of semi-supervised techniques relative to fully supervised methods, the following methodology is employed. Initially, this involves the processing of the data and selection of kinematics to establish a well-defined and constrained analytical framework. Subsequently, it involves the implementation, training, and optimisation of both the semi-supervised and fully supervised models.

5.3.1 Data processing and kinematic selection

In the initial phase of the study, a comprehensive data processing and evaluation of the di-lepton dataset is undertaken. This crucial step involves a detailed analysis of the dataset to ascertain the size and characteristics of both signal and background samples, essential for accurate model comparison.

The investigation begins with a thorough analysis of the di-lepton dataset, focusing on the ratio and distribution of signal to background events. This examination uncovers the fundamental structure of the data, which is essential for the subsequent phases of classifier model implementation.

Following this, the selection and assessment of kinematic variables are carried out. This process involves a systematic analysis of various factors to identify their relevance and impact on the study. The selection of variables is driven by the objective to maximise the discriminative power of the classifier models, choosing those that effectively differentiate between signal and background events.

Upon identifying key variables, kinematic features are calculated. This step refines the dataset, ensuring that the classifiers have access to the most relevant and informative data. These features are rigorously evaluated to confirm their effectiveness in enhancing classifier performance.

Finally, event weighting is applied to both signal and background events, a crucial step in ensuring the classifier models accurately represent real-world physics phenomena. This weighting corrects potential imbalances or biases in the dataset, ensuring that the results are not just statistically sound, but also physically significant.

In summary, the process of data processing and kinematic selection is a multifaceted and meticulous procedure, optimising the dataset for the effective implementation and comparison of various classifier models. This foundation is vital for reliable and insightful analysis.

5.3.2 Model implementation, training, and optimisation

The efficacy of semi-supervised techniques in this context is assessed by first establishing a robust baseline with fully supervised learning models. The initial stage involves the meticulous development of supervised machine learning algorithms, specifically tailored for the BDT, MLP, and DNN models. A critical aspect of this stage is the rigorous training, testing, and optimisation of each model. The primary objective here is to achieve high accuracy in distinguishing between labelled di-lepton signals and background processes, which is fundamental for valid comparative analysis.

Upon successfully establishing the fully supervised baseline, a parallel and equally thorough process is undertaken with semi-supervised techniques. This involves adapting each of the BDT, MLP, and DNN models to operate

in a semi-supervised framework. The unique challenge here is to train and optimise these models to classify di-lepton signal samples within a mixture of unlabelled signal and background events, using a set of labelled background samples as a guide. This step is crucial in demonstrating the viability of semi-supervised methods in scenarios with incomplete labelling or directly on LHC data.

The optimisation of each model is a meticulous process, involving not just the adjustment of model architectures but also the careful tuning of hyperparameters to ensure optimal performance. This step is pivotal in enhancing the models' ability to discern subtle patterns and distinctions in the data, a task that is central to the success of classifiers.

For the implementation of all models, the study leverages the Toolkit for Multivariate Data Analysis with ROOT (TMVA) [123]. TMVA offers a comprehensive suite of machine learning tools specifically developed for particle physics research, particularly suited to analyses involving LHC data. Its sophisticated features and particle physics-centric design make it the optimal tool for this study, ensuring that the models are not only technically sound but also aligned with the unique requirements of high-energy physics research.

The final step involves a critical comparison of the optimised models, focusing on both their classification algorithms and supervision techniques. To accomplish this, each model's ability is quantitatively measured using the resultant AUC score. Furthermore, the Kolmogorov-Smirnov test is employed to assess the extent of overtraining in each model. This dual approach ensures a comprehensive evaluation, allowing for a nuanced understanding of the performance and reliability of the models under different supervision regimes.

In conclusion, the model implementation, training, and optimisation process is designed to rigorously test the capabilities of semi-supervised learning in comparison with fully supervised methods, within the specific context of high-energy physics and particle classification. The approach used here, from model development to optimisation, underscores the commitment to achieving accurate, reliable results that can significantly contribute

to the understanding of semi-supervised learning in this field.

5.4 Model and Supervision Results

5.4.1 Setup and data preprocessing

This section details the setup and data preprocessing steps undertaken for the analysis of the di-lepton final state dataset. The focus is on defining and processing the signal and background regions, as well as the event samples, to ensure a well-constrained dataset for the analysis.

Initially, necessary cuts are applied to delineate the signal and background regions. This process is instrumental in refining the dataset, resulting in a clearer segregation of signal and background events. The effectiveness of these cuts is demonstrated by the quantitative breakdown of the signal and background Monte Carlo events, as shown in Table 5.1. This table presents the raw number of background events for each process, as well as the final number of raw and weighted signal and background events after cuts, offering a transparent view of how these cuts impact the dataset.

Following the event selection process, the study advances to the training phase of the classifier models. A crucial step in this phase is the selection of an optimal and well-constrained set of kinematic features. This selection process is not arbitrary; it involves a careful analysis of the kinematic feature distribution separation as well as feature ranking during the model implementation phase. The feature ranking, in particular, plays a vital role as it reflects the extent to which the value of a feature influences a machine learning model's ability to classify data. The chosen features are those that exhibit the most significant impact on the discrimination ability of the classification models, thereby maximizing their potential to accurately segregate signal from background events.

In summary, the setup and data preprocessing steps lay a solid foundation for the subsequent analysis. By defining the signal and background

Number of Signal and Background Events			
Signal			
Process	**Raw Events**		**Weighted**
X[240]→S[170]h	1282		6410
Background			
Process	**Raw Events**	$N_{bj} \geq 2$	$m_{bb} > 150\,\mathbf{GeV}$
Higgs	985203	10894	1046
WW	1444022	5430	1151
VV	1348938	9473	2643
Vγ	9533	31	5
$t\bar{t}$	7553172	2929354	1326976
Top	620381	169148	83544
Zγ	1120628	5495	1850
W Jets	447	0	0
Total	13082324	3129825	1417215
Weighted Total	438396	114145	51825

TABLE 5.1: Number of signal and background events before and after selection cuts for $\sqrt{s} = 13\,\text{TeV}$.

regions, applying targeted selection cuts, and choosing the most influential kinematic features, the dataset is optimally prepared for the advanced classification tasks ahead.

5.4.2 Model implementation and development

In the model implementation and development phase of this study, the focus is on optimising the architecture and hyperparameters of each of the machine learning methods: BDT, MLP, and DNN. This optimisation is critical for enhancing the models' ability to effectively discriminate between signal and background events in the di-lepton dataset.

To ensure a transparent and comprehensive understanding of this process, a summary of the evaluated and final optimised hyperparameter values for each model is provided in Tables 5.2, 5.3, and 5.4. These tables present a detailed overview of the parameters considered for each model, the range of values evaluated, and the optimal selections made based on the analysis.

Boosted Decision Tree		
Hyper-Parameter	**Evaluation Range**	**Optimised Selection**
Boost Type	–	Gradient Boosting
Number of Trees	$[10, 1000]$	100
Max Depth	$[1, 10]$	2
Min Node Size	$[0.01, 0.1]$	0.02
number of Cuts	$[100, 500]$	200
Shrinkage	$[0.01, 0.1]$	0.04
Bagged Sample Fraction	$[0.1, 1]$	0.25

TABLE 5.2: BDT model hyperparameter optimisation.

Multilayer Perceptron		
Hyper-Parameter	**Evaluation Range**	**Optimised Selection**
Learning Rate	$[1 \cdot 10^{-1}, 1 \cdot 10^{-3}]$	$1 \cdot 10^{-2}$
Epochs	$[10, 1000]$	600
Neuron Activation	[tanh, sigmoid]	tanh
Hidden Layers	$[1, 5]$	2
Neurons per Hidden Layer	$[5, 100]$	80
Training Method	[MLP, MLPBFGS, MLPBNN]	MLPBNN

TABLE 5.3: MLP model hyperparameter optimisation.

Deep Neural Network		
Hyper-Parameter	**Evaluation Range**	**Optimised Selection**
Learning Rate	$[1 \cdot 10^{-1}, 1 \cdot 10^{-3}]$	$1 \cdot 10^{-2}$
Batch Size	$[1, 500]$	256
Momentum	–	0.9
Weight Decay	$[1 \cdot 10^{-2}, 1 \cdot 10^{-5}]$	$1 \cdot 10^{-4}$
Epochs	$[10, 2000]$	800
Neuron Activation	[tanh, ReLu]	tanh
Hidden Layers	$[1, 10]$	5
Neurons per Hidden Layer	$[10, 200]$	128

TABLE 5.4: DNN model hyperparameter optimisation.

The optimisation of hyperparameters for each algorithm is guided by two primary objectives: to maximise the area under the ROC curve (a measure of background rejection versus signal efficiency), and to minimise the extent of

overtraining. Overtraining can significantly reduce a model's ability to generalise to new data, hence its mitigation is a key aspect of this optimisation process.

To evaluate the extent of model overtraining, the Kolmogorov-Smirnov (KS) test is used. This statistical test compares the output response distributions of the training and validation datasets, providing a measure of their similarity. Conducted concurrently with the generation of these distributions, the KS test facilitates real-time monitoring and adjustment, ensuring maximal generalisability of each model.

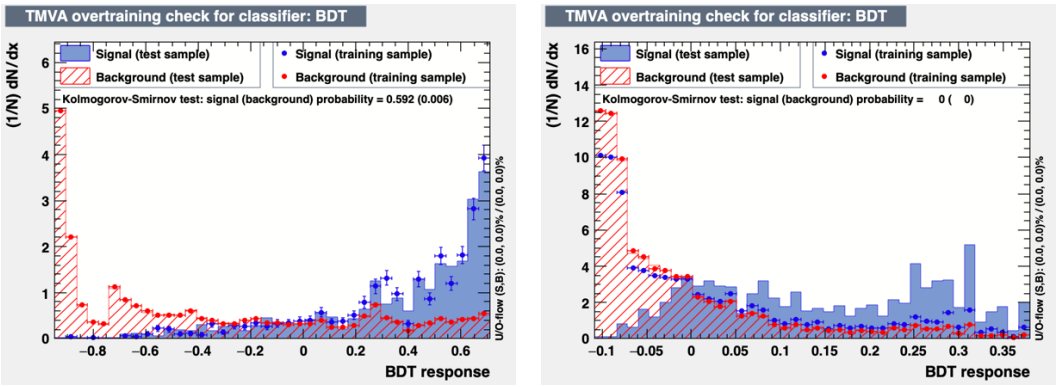
Overall, the model implementation and optimisation is crucial for ensuring that each machine learning model is not only technically sound in terms of its architecture and parameters but also effective and reliable in its performance. The optimisation process undertaken here sets a strong foundation for the final evaluation and discussion of the results in the subsequent section.

5.4.3 Final results and discussion

The analysis of the resultant response distributions for each model, trained under both fully-supervised and semi-supervised techniques, offers valuable insights into their respective abilities to differentiate between signal and background events in the di-lepton dataset. This section delves into the performance of the BDT, MLP and DNN models, scrutinizing their response distributions and comparing their efficacy through ROC curves.

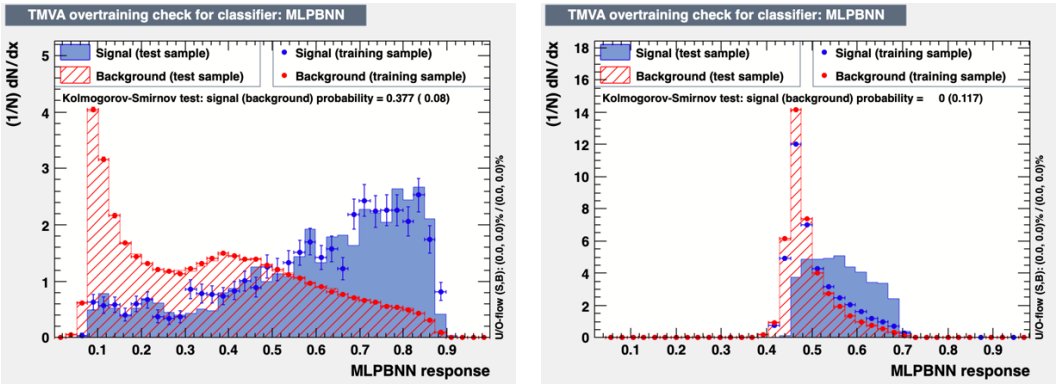
Starting with the BDT model, the response distributions for the fully supervised (Figure 5.8a) and semi-supervised (Figure 5.8b) approaches demonstrate effective discrimination between signal and background events. Although the separation is not optimal, the tendency of background events to be labelled with response values tending towards zero in both methods underscores their effectiveness.

Turning to the MLP model, its response distributions for both training techniques are shown in Figure 5.9. The MLP responses, when compared to the BDT, exhibit smoother distributions but with a lesser degree of separation.



(a) Fully-Supervised. (b) Semi-Supervised.

FIGURE 5.8: Response distributions of final optimised BDT models.

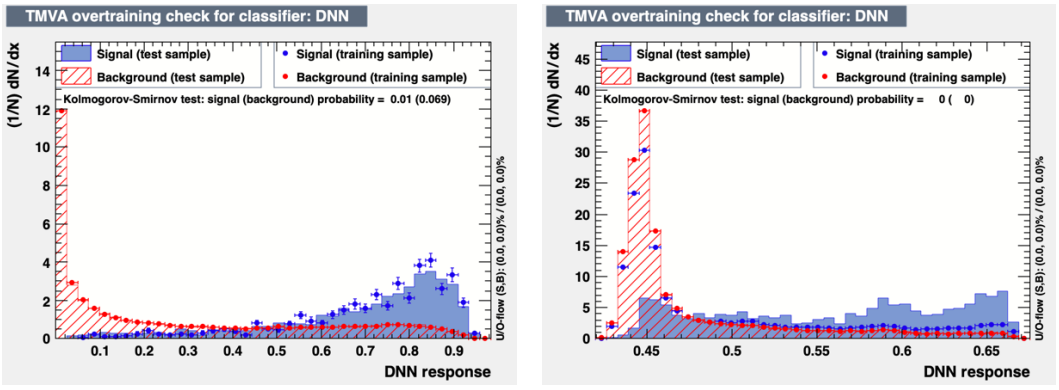


(a) Fully-Supervised. (b) Semi-Supervised.

FIGURE 5.9: Response distributions of final optimised MLP models.

Despite this, the MLP successfully distinguishes between signal and background events, although with a more significant overlap in responses than that of the BDT.

The DNN model's response distributions, presented in Figure 5.10, demonstrate an exceptional ability to classify background events for both supervision techniques. The fully-supervised DNN model shows notable success in classifying signal events, with the semi-supervised model's separation only slightly inferior. The semi-supervised response distributions illustrate that, in semi-supervised training, the signal events more closely follow the background dynamics then in the fully supervised method.



(a) Fully-Supervised. (b) Semi-Supervised.

FIGURE 5.10: Response distributions of final optimised DNN models.

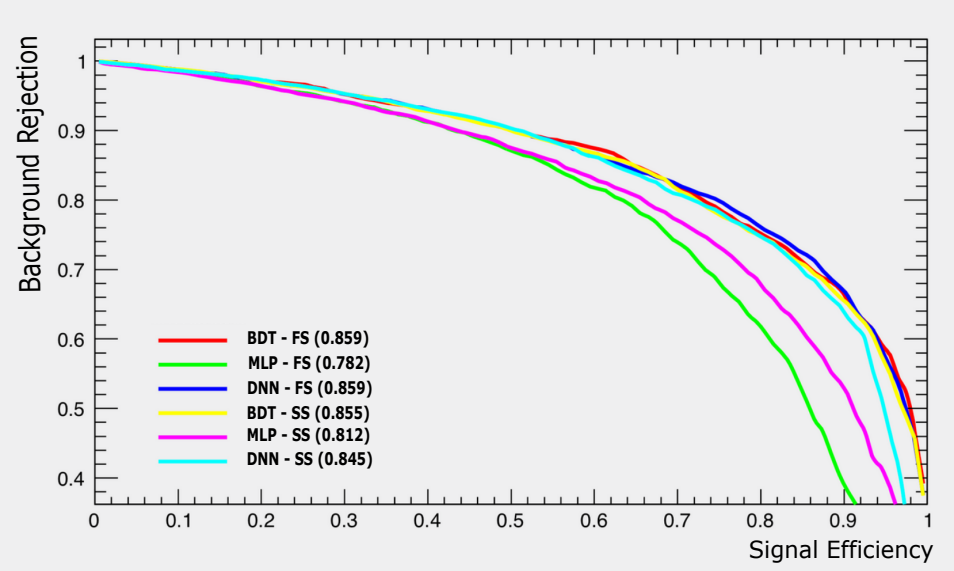


FIGURE 5.11: ROC curve comparison of BDT, MLP and DNN models using fully-supervised (FS) and semi-supervised (SS) techniques.

For a direct comparison of the models and the effectiveness of the supervision techniques, the ROC curves, illustrated in Figure 5.11, are used. These curves facilitate the computation of the area under the curve (AUC), presented in the legend of Figure 5.11, providing a quantifiable measure of each classifier's performance.

The AUC analysis reveals that the best-performing models are the fully supervised BDT and DNN, with AUC scores of 85.9%. Their semi-supervised

counterparts also exhibit strong performance, with the BDT and DNN scoring 85.5% and 84.5%, respectively. The MLP model produced the least desirable separation with AUC scores for fully supervised and semi-supervised of 0.783 and 0.812 respectively. The fact that the semi-supervised MLP produced better results than it's fully supervised counterpart is likely an indication that the model was not successfully optimised.

Notably, the fully supervised technique outperforms the semi-supervised method by a marginal difference of less than 3%. This finding underscores the semi-supervised technique's capacity to reduce model dependency, and still maintain excellent classification efficiency for particle signatures.

Additionally, when assessing the extent of over-training, the DNN model emerges as the most robust, followed by the BDT, with the MLP showing the highest susceptibility to over-training. This hierarchy in over-training resistance further influences the decision to prioritise the DNN model for subsequent studies, given its superior balance of classification accuracy and generalisability.

Chapter 6

Measuring a Trials Factor in Semi-Supervised $Z\gamma$ Resonance Searches

6.1 Introduction

Leveraging the powerful potential of machine learning classifiers for resonance searches at the LHC, opens up new possibilities for exploring the fundamental constituents of the universe. These classifiers can learn intricate patterns and nonlinear relationships from large and high-dimensional datasets, providing a significant advantage when dealing with the LHC's complex and voluminous data. The utility of machine learning models in this context is not limited to merely sifting through data but extends to identifying and classifying potential resonance signals, often buried deep within background noise. By enabling a more precise and efficient resonance search process, classifiers can accelerate the discovery of new particles or exotic phenomena, pushing the boundaries of our current understanding of particle physics. Semi-supervised classifiers introduce a theoretical model independent methodology to identify new physics. In Chapter 5, semi-supervised Neural Network (NN) models are shown to not only reduce model dependencies, but also classify LHC data with comparable success to standard fully supervised methods.

However, semi-supervised classifiers are not without their own challenges.

The implementation of fully supervised models uses pre-defined signal and background samples to train. Once trained, the predefined signal and background samples can be used to calculate the optimal response distribution cut to maximise the signal extracted. This is not the case in semi-supervised classifiers which do not have a predefined signal sample. Thus, in order to determine the optimal response distribution cut, a scan is used. This scan of the response distribution introduces an additional look-elsewhere effect that can manifest itself as "fake signals". These fake signals can significantly distort resonance patterns and influence the interpretations [1]. Hence, it's imperative to evaluate the probability of fake signals produced in the implementation of these classifiers, and expose it in the form of a trials factor.

When conducting any resonance search, it is crucial to take into account the look-elsewhere effect to arrive at a statistically meaningful result. This effect refers to the increased probability of finding apparent signals (that are actually statistical fluctuations) when searching over an extended range, usually mass. This effect can lead to unreliable results if not properly addressed. Therefore, the look-elsewhere effect introduced by semi-supervised classifiers has to be included on top of those measuring the probability of statistical fluctuations over an extended mass range, and thus occurs even if the mass of the new resonance is fixed (known).

This chapter presents the methodology and results of a frequentist study aimed at estimating the prevalence of additional look-elsewhere effects introduced by semi-supervised NN models, in the form of a trials factor. In a frequentist approach, the sum of trials, conducted as pseudo-experiments, estimates the probability of outcomes based on their relative frequency over numerous repetitions. Motivated by the multi-lepton anomalies and the excess around 150 GeV, the simulated $Z\gamma$ final state dataset, introduced in this Chapter, is used. An additional component to the analysis is an evaluation of ML-based data generators for scaling and sampling datasets in HEP analyses.

6.2 Final State $Z\gamma$ Searches

As introduced in Section 2.3, substantial evidence indicates that the scalar, S, with a mass, m_S, of approximately 152 GeV predominantly decays into the $\gamma\gamma$ final state, presenting a large excess in this channel. However, a more subtle excess is observed in the $S \to Z\gamma$ decay mode, a phenomenon that remains largely unexplained [38, 53]. As a result, we search for $S \to Z\gamma$, at the mass of 150 GeV. Such an inquiry is pivotal for understanding the nuances of the S decay mechanisms and potentially uncovering new physics beyond the Standard Model.

The dominant background source in the search for the $S \to Z\gamma$ decay mode of new resonances is due to the non-resonant production of a prompt photon, γ, and a Z boson. The Standard Model non-resonant $Z\gamma$ represents more than 92% of total background [124]. A secondary smaller contribution comes from the production of Z bosons in association with jets (Z+jets), with one jet misidentified as a photon. The Z+jets background is not considered in this study as its contribution is derived from real data at the LHC.

The production of the Higgs like heavy scalar decaying to $Z\gamma$ ($pp \to H \to Z\gamma$) events, where $Z \to \ell^+\ell^-$. This means that the Z boson decay into two leptons of opposite sign, either two muons, $\mu^+\mu^-$, or two electrons, e^+e^-. The production of $\ell^+\ell^-\gamma$ are exposed using Feynman diagrams, Figure 6.1.

6.2.1 Monte Carlo Event Generation

To thoroughly optimize and assess the strategy, a comprehensive simulation of 110,000 $Z\gamma$ events was carried out using `MadGraph5_aMC@NLO 2.6.7`, with Next-to-Leading Order precision in QCD [111]. The processes of parton showering and hadronisation were executed with `PYTHIA 8.2` [126], employing the ATLAS A14 tuning and the NNPDF2.3 Leading Order parton distribution functions [127].

Simulations were facilitated using `Delphes3` [128], offering a fast simulation approximating the ATLAS experiment's current setup. The specific generator-level criteria were applied to identify relevant events, necessitating photons to have a transverse momentum exceeding 25 GeV and the $Z\gamma$

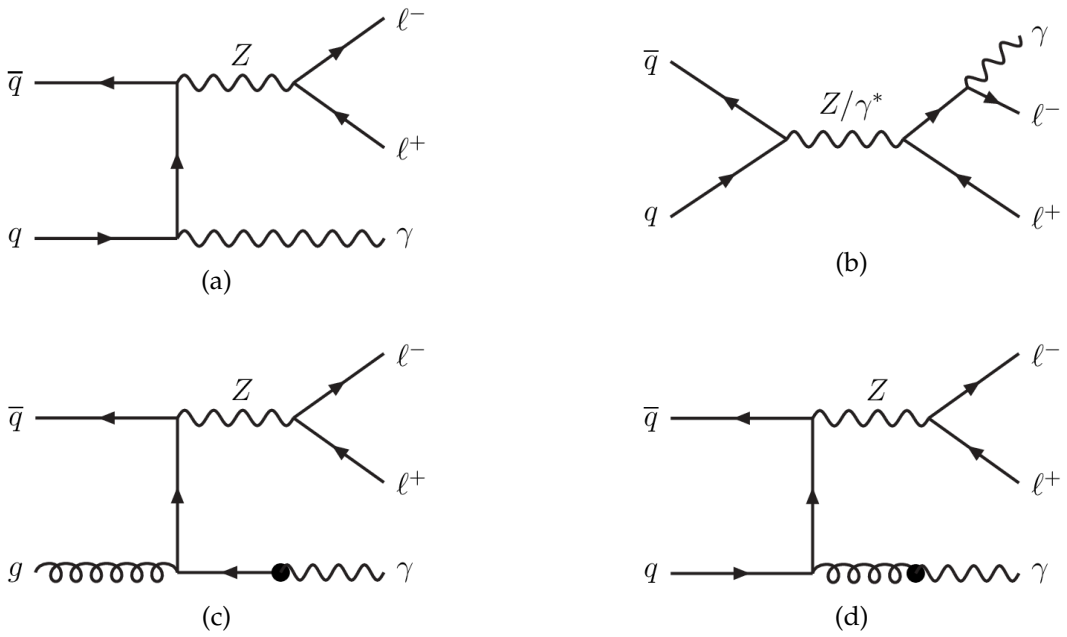


FIGURE 6.1: Feynman diagrams for $\ell^+\ell^-\gamma$ production: (a) photon radiation from a quark leg; (b) final-state photon radiation from a lepton; and (c, d) contributions from Z + q(g) processes in which a photon is produced from the fragmentation of a quark or gluon [125].

invariant mass to lie within the 132 to 168 GeV range. The reconstruction of hadronic jets was performed using the anti-k_t algorithm [129], with a radius parameter of $R = 0.4$, via the `FastJet` 3.2.2 [130]. Only jets with a transverse momentum greater than 30 GeV and a pseudo-rapidity of $|\eta| < 4.7$ were considered. Furthermore, jets derived from bottom quarks were classified as b-jets using specific b-tagging methods [131]. Reconstructed jets overlapping with photons, electrons, or muons within a cone of $R = 0.4$ were excluded. Electrons and muons were mandated to have a p_{T} greater than 25 GeV and a pseudo-rapidity within $|\eta| < 2.5$.

6.2.2 Invariant Mass Fixed Study

In the proposed theoretical model, we are interested in evaluating the $Z\gamma$ final state dataset as the primary background for the scalar, S, around the motivated mass of 150 GeV.

This study therefore confronts mass window and side-band regions of the

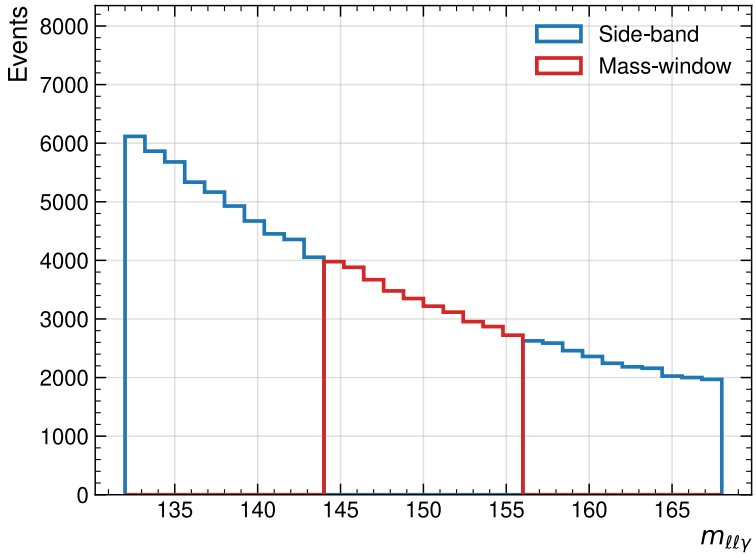


FIGURE 6.2: Invariant mass distribution highlighting the defined mass-window and side-band regions.

invariant mass, $m_{\ell\ell\gamma}$, spectrum. The mass window considers the mass range around 150 GeV, where the signal is expected to be found. The side-band region considers the mass regions surrounding the mass window where only background is expected.

The mass window is fixed to the range from 144 to 156 GeV. The side-band is fixed to the range from 132 to 144 GeV excluding the mass window. The invariant mass distribution exposing side-bands and mass window can be visualised in Figure 6.2.

In order to best describe the background distribution and potential signal resonances, functional fits are used.

Control region

The $Z\gamma$ invariant mass distribution, shown in Figure 6.2, is the control region (background contribution to the signal region of interest). In resonance searches a pure background distribution represents the null hypothesis, for the case where there is no signal resonances. A function must be selected to best describe the background distribution. On initial inspection the shape

	Exponential	**Quadratic-Exp**	**Polynomial**	**Poly-Log-Exp**
C0	$49.289 \cdot 10^{4}$	$19.330 \cdot 10^{3}$	$65.501 \cdot 10^{3}$	23.362
C1	$34.342 \cdot 10^{-3}$	6.257	-735.161	1.054
C2	-	$-75 \cdot 10^{-3}$	2.115	$648.860 \cdot 10^{-3}$
C3	-	$0.14 \cdot 10^{-3}$	-	-
χ^2	$3.932 \cdot 10^{-3}$	$2.502 \cdot 10^{-3}$	$2.858 \cdot 10^{-3}$	$2.454 \cdot 10^{-3}$

TABLE 6.1: Comparison of background functional fits on $Z\gamma$ invariant mass distribution.

of the distribution appears exponential. However, it is important to consider more complex functional forms to best describe the distribution. The functional forms considered are, an exponential function,

$$f(x) = c_0 \cdot e^{-c_1 \cdot x}, \tag{6.1}$$

a quadratic exponential function,

$$f(x) = c_0 \cdot e^{-c_1 \cdot x + c_2 \cdot x^2}, \tag{6.2}$$

a polynomial function,

$$f(x) = c_0 + c_1 \cdot x + C_2 \cdot x^2, \tag{6.3}$$

and a mixed poly-log-exponential function [132],

$$f(x) = (1 - x)^{c_0} \cdot x^{c_1 + c_2 \cdot \log(x)}. \tag{6.4}$$

Each of the background functional forms are evaluated by fitting them to the invariant mass distribution of the entire MC dataset. The fit is done using 36 bins, one per GeV, using the entire dataset, 106,474 events. The success of each fit is evaluated using the difference between the data and background fit as well as the χ^2 value. A tabulated summary and a visual comparison of the resulting background functional fits are exposed in Table 6.1 and Figure 6.3 respectively.

The comparison of background functional forms, on the specified invariant

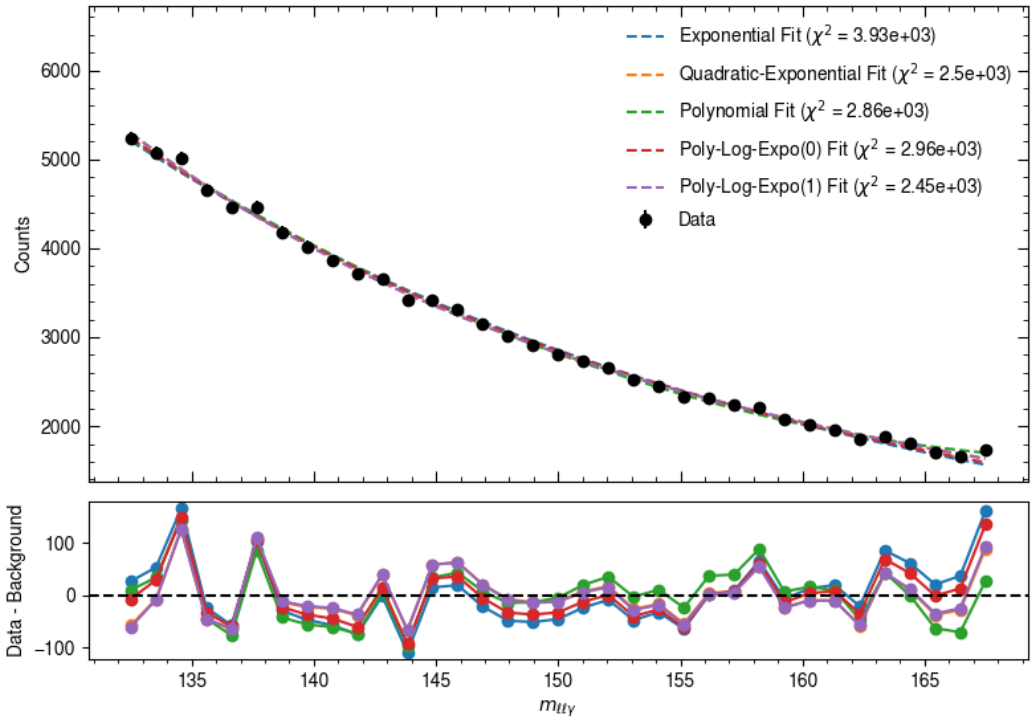


FIGURE 6.3: Background functional form fit comparison on MC $Z\gamma$ invariant mass distribution.

mass range, affirms that the mixed poly-log-exponential function, selected in Ref. [132], has the minimal χ^2, and is therefore selected as the optimal background function for this analysis.

Signal region

In order to measure the amount of signal events found in the signal region, a signal functional form is used. A Gaussian function is selected as the functional form to expose signal events in this mass window. The Gaussian functional form is shown in Equation 6.5, as:

$$g(m) = c_s \cdot e^{-\frac{(m-\mu_S)^2}{2\cdot\sigma_S^2}}, \tag{6.5}$$

where m is the mass, c_s is a scalar measuring the curve's peak, μ_s is the position of the centre of the peak, and σ_s is the standard deviation. In this

analysis, the μ_s can be fixed at 150 GeV and σ_s can be set to the search resolution of 2.4 [124].

6.2.3 Kinematic Features

The key kinematic features of the $Z\gamma$ final state dataset are categorised into three groups. First, we consider the individual characteristics of the leading lepton, ℓ_1, sub-leading lepton, ℓ_2, and the photon, γ. These include their transverse momentum, Pt, azimuthal angle, ϕ, pseudo-rapidity, η, and energy, E.

Next, we examine the kinematics of the $Z\gamma$ system as a whole, including its invariant mass, $m_{\ell\ell\gamma}$, transverse momentum, $Pt_{\ell\ell\gamma}$, missing transverse energy, E_T^{miss}, the azimuthal angle of the missing transverse energy, $\Phi_{E_T^{miss}}$, the number of jets, N_j, and the number of central jets, N_{cj}.

Finally, we assess the relational kinematic features. These include the distance $\Delta R_{\ell\ell}$, which measures how far apart the two leptons are in the $\eta - \phi$ space[1]. We also consider the difference in the azimuthal angles of the two leptons, $\Delta\Phi_{\ell\ell}$, and the difference in azimuthal angle between the missing transverse energy and the $Z\gamma$ system, $\Delta\Phi(E_T^{miss}, Z\gamma)$. The feature distributions of the $Z\gamma$ MC dataset are shown in Figure 6.4.

6.3 Pseudo-Experiment Design

In this study, we employ the frequentist approach to quantify the prevalence of additional look-elsewhere effect introduced in semi-supervised learning models. Central to this approach is the rigorous repetition of a carefully designed pseudo-experiment, conducted thousands of times to yield statistically robust and accurate probabilities of outcomes. This methodological rigour is essential for the reliable interpretation of results in the context of semi-supervised learning.

The core structure of the pseudo-experiment, specifically tailored for this study, comprises three primary components: data sampling or generation,

[1]The distance ΔR is defined as $\Delta R = \sqrt{(\Delta\eta_{\ell\ell})^2 + (\Delta\phi_{\ell\ell})^2}$.

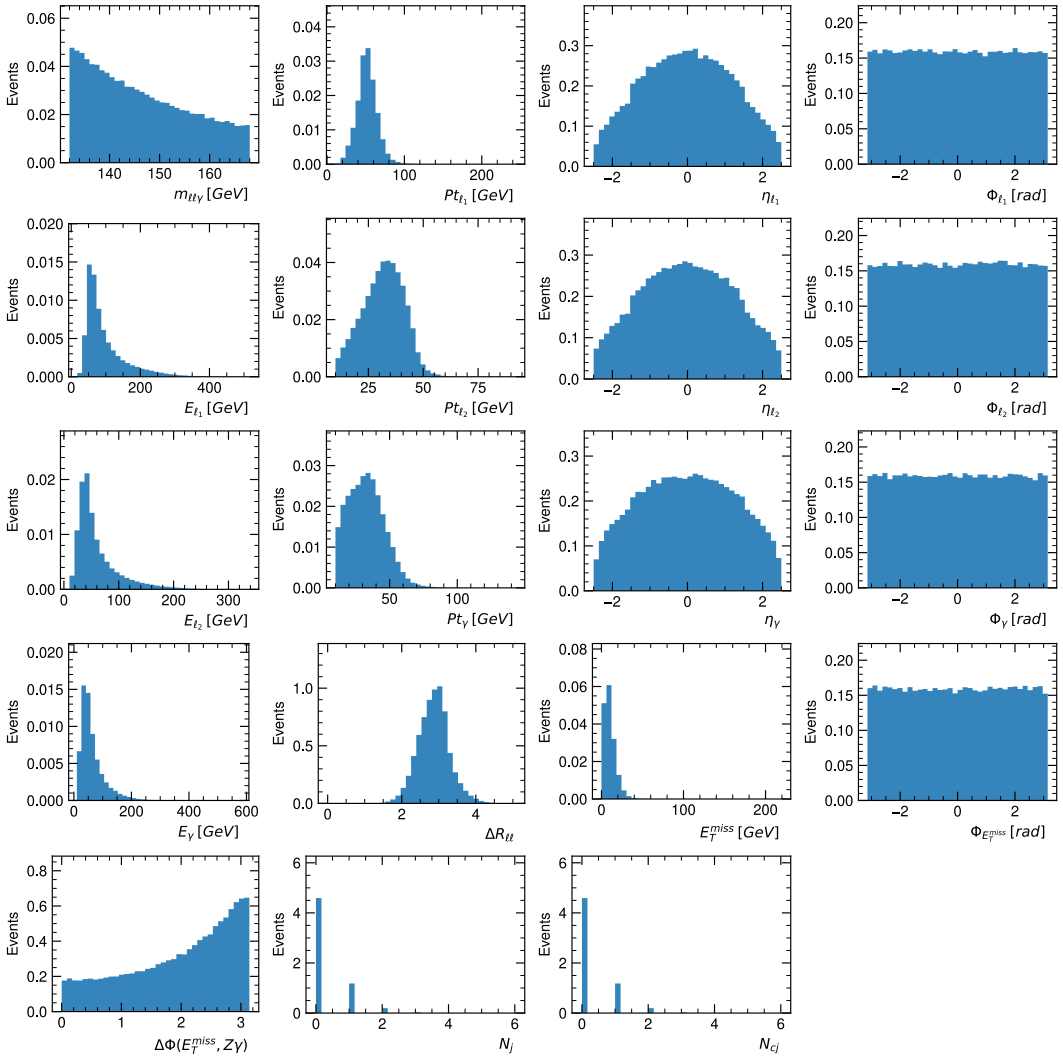


FIGURE 6.4: Monte Carlo $Z\gamma$ dataset kinematic feature distributions.

semi-supervised training, and the measurement of signal significance. A simplified visualization of the pseudo-experiment, contextualised within the frequentist framework, is provided in Figure 6.5.

To further delineate the extent of false signals and uncover the look elsewhere effect, the design of the pseudo-experiment is elaborated. The output response of the trained semi-supervised classifier is scanned, producing local samples of extracted events reflective of the process to optimise the extraction signal events. Each of the local samples are then mapped to their corresponding invariant mass, where signal and background fits reveal the

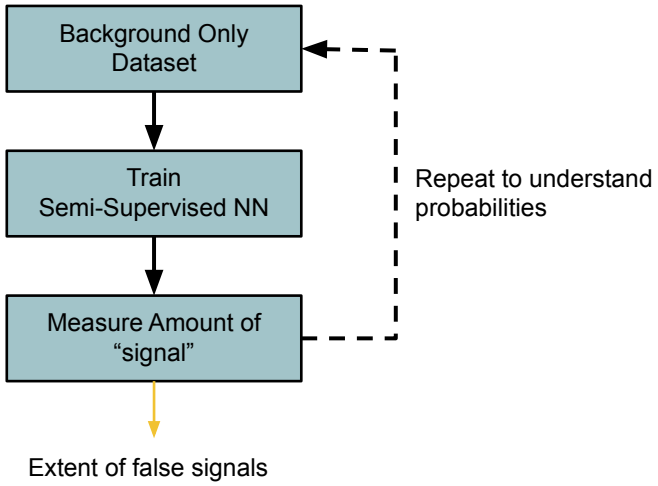


FIGURE 6.5: Simplified diagram of the frequentist study, pseudo-experiment.

significance of fake signals in each sample. The expanded overview and detailed flow diagram of the pseudo-experiment is illustrated in Figure 6.6.

In the following sections, each component of the pseudo-experiment will be expanded upon, providing a comprehensive exploration and detailed analysis of their individual roles and interplay in the context of this study.

6.3.1 Data sampling and generation

Frequentist studies require a sufficient number of repetitions of an experiment to have statistically meaningful results. For each pseudo-experiment iteration, a sample of 80,000 events is used to train and evaluate the NN classifier, in order for the results not to be limited by the MC statistics. However, repeatedly generating these samples of events with pure MC methods would require substantial computational resources and time [133]. Fortunately, machine learning-based data generators have recently emerged as a novel and powerful tool for generating and scaling data samples [134, 135]. These advanced algorithms leverage the potential of ML and deep learning to create realistic, high-quality simulated datasets. Such data generators utilise various architectures, most notably Generative Adversarial

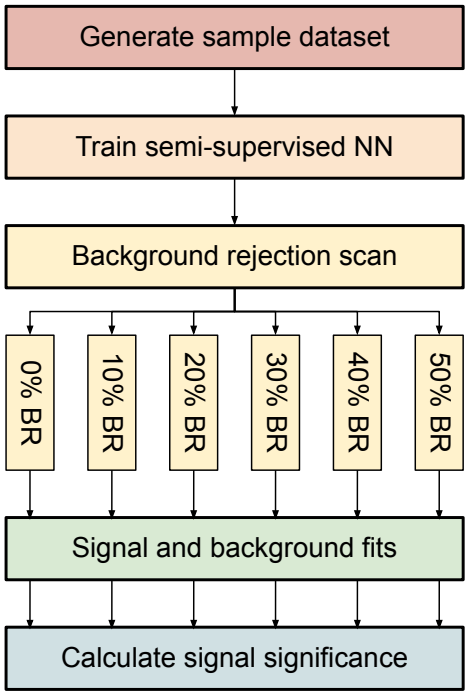


FIGURE 6.6: Diagrammatic overview of the designed pseudo-experiment.

Networks (GANs) and Variational Autoencoders (VAEs), each demonstrating unique capabilities and potential use-cases in particle physics.

An additional study, presented in Section 6.4, considers and compares the KDE, GAN and VAE data generation methods. The optimal model is then selected and used for the frequentist study to sample/generate statistically accurate and independent datasets for each pseudo-experiment.

6.3.2 Semi-supervised neural network

The core component of the analysis is the evaluation of the semi-supervised NN classifier. In this study, the NN model used is kept consistent with the model used in Ref. [90]. The model architecture consists of an input layer, three hidden layers and an output layer, shown in Figure 6.7. The input layer consists of seven nodes corresponding to the input features. The hidden layers have 26, 26 and 13 nodes, respectively. The output layer consists

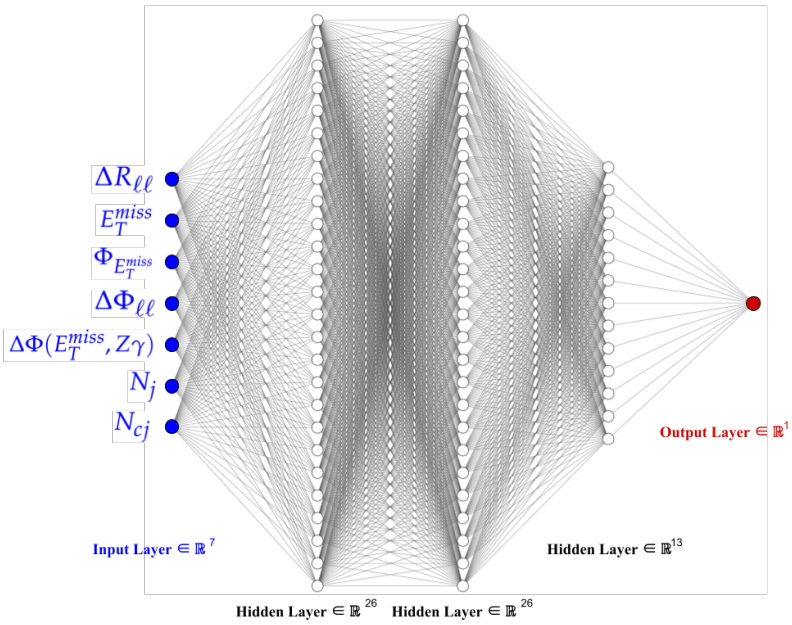


FIGURE 6.7: Neural network model architecture.

of a single node that outputs the NN response. The neurons in the hidden layers each use the *Tanh* activation function. Finally, the output neuron utilises the sigmoid activation function, forcing the output to a value between zero and one.

The training features are selected to reflect the $Z\gamma$ kinematics excluding features with direct correlation to the $m_{\ell\ell\gamma}$. To this end the features used to train the NN are $\Delta R_{\ell\ell}$, E_T^{miss}, $\phi_{E_T^{miss}}$, $\Delta\phi_{\ell\ell}$, $\Delta\phi(E_T^{miss}, Z\gamma)$, N_j, and N_{cj}.

The loss function implemented in the training of the semi-supervised NN is binary cross-entropy. The binary cross-entropy loss function is described using Equation 4.1. The model weights are optimised using the *Adam* optimiser, as described in Section 4.5.2.

The semi-supervised model is trained using two samples, as described in Section 4.3.3. The samples are each defined on the $m_{\ell\ell\gamma}$ distribution. The pure background sample, sample 0, consists of events in the side-band region. The mixed signal and background sample, sample 1, consists of events in the mass-window region. The mass-window and side-band regions are defined in Section 6.2.2.

6.3.3 Background rejection scan

Once trained, the NN output response distribution is analysed. Due to the Sigmoid activation on the NNs output node, response values vary between 0 and 1. The NN model attempts to classify side-band events, sample 0, as background with responses tending towards 0, and mass-window events, sample 1, as signal with responses tending towards 1. Due to the fact that in semi-supervision sample 1 contains signal and background events, only the signal events should generate a response close to 1, while the background events should not be differentiable from sample 0 events and classified as such. The NN response distribution in the study is therefore expected to show minimal separation between the samples, as there is no signal events present in sample 1.

In the semi-supervised context, the extraction of signal events using the NN response can rather be understood as the rejection of background events. By rejecting background events, events with response tending to 0, we are left with a local sample of signal events. The optimal background rejection cut on the response distribution must therefore be determined to remove as many background events as possible from the sample.

Fully supervised classifiers have predefined signal events, allowing for the optimal response distribution cut to be selected. However, semi-supervised classifiers do not have predefined signal samples and must use a scan to determine the optimal background rejection cut. This scan may introduce additional look elsewhere effects. In order to take this into account and measure the extent of the effect, a background rejection scan is used. The background rejection scan consists of extracting samples of events with an increasing number of rejected background events. The amount of background events rejected in each sample is 0, 10, 20, 30, 40 and 50 percent of the total events, respectively. The events from each background rejection cut forms a local sample, and can be mapped to their respective invariant mass for analysis. Figure 6.8 provides a visual example of the background rejection scan and extraction of local event samples.

Performing a scan of the NN model's response yields critical insights into

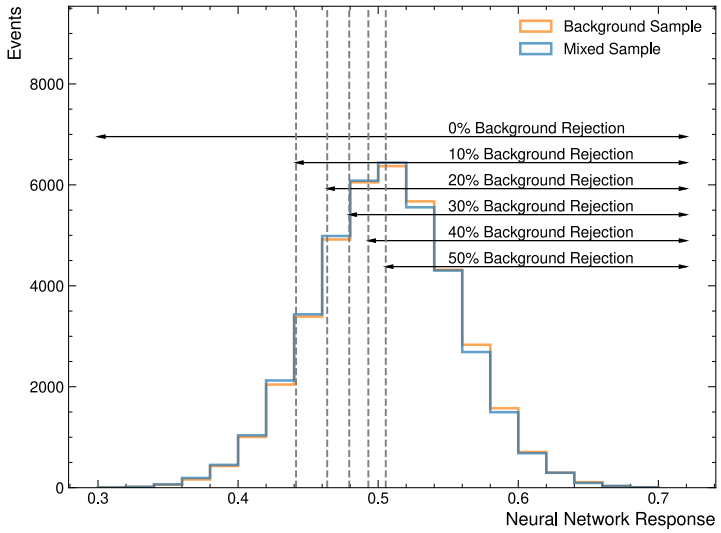


FIGURE 6.8: Diagram of background rejection scan.

the semi-supervised model. This process plays a pivotal role in optimizing signal efficiency and background rejection by identifying the most effective thresholds to enhance classification capabilities. It also provides a comprehensive view of the model's performance across various thresholds, highlighting potential areas of over- or under-optimization and suggesting necessary adjustments or refinements. Furthermore, scanning over the NN response adds a layer of robustness against overfitting, ensuring that the model remains generalizable and not overly tailored to the specific traits of the training data.

Another vital aspect of the response scan is the evaluation of systematic uncertainties. In fields like particle physics, where minor variations in input parameters can significantly impact results, assessing how the NN response is affected by these uncertainties is crucial. This sensitivity analysis helps in understanding the reliability and stability of the model under varying conditions.

Most importantly, particularly in the context of this study, the response scan facilitates hypothesis testing. By generating localised samples for analysis, it enables us to evaluate the significance of any signal present within these samples, offering a robust method to test and validate our hypotheses.

6.3.4 Invariant mass fits

In each background rejection sample, we observe a local analysis subset. For a precise measurement of false signals within the signal region, events from these samples are mapped to their corresponding invariant mass. The local distributions of $m_{\ell\ell\gamma}$ are then scrutinised using functional fits, a method essential for revealing any signal presence in the mass-window region. The functional forms employed for representing the background and signal are detailed in Sections 6.2.2 and 6.2.2, respectively.

To synthesise the signal and background representations, we combine them into a comprehensive model probability density function (PDF). This process is central to our analysis and involves testing the background-only, null hypothesis. By adopting this hypothesis testing approach, we assess the probability of observing signal-like events purely as statistical fluctuations within the background. This critical step aids in accurately identifying true signal events against the backdrop of background noise.

In the analysis, we fit a combined PDF to the invariant mass data. This is achieved by minimizing the negative log-likelihood. The fitting involves comparing the model PDF with the observed data. For this purpose, the ZFIT library is used [136]. ZFIT leverages the robust capabilities of the ROOT framework, offering a sophisticated environment for both fitting and statistical analysis. This combination of advanced tools enhances the accuracy of the fits and strengthens the hypothesis testing.

In Figure 6.9, we demonstrate this fitting process. It illustrates three key components: the background-only model (dashed blue line), the combined signal and background PDF model (dashed red line), and the signal component in isolation (solid blue line). This representation clearly distinguishes each element within the fitting procedure.

By combining meticulous fitting processes with comprehensive hypothesis testing, the study aims to clearly distinguish genuine signal events from statistical noise, thereby augmenting our understanding of false signals in semi-supervised models.

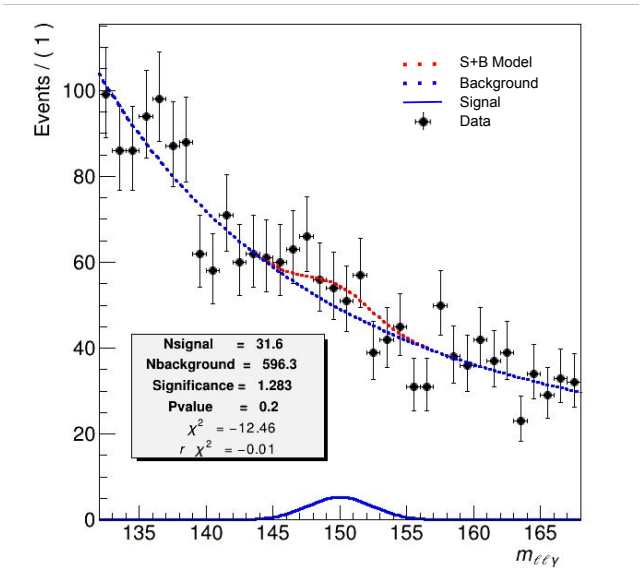


FIGURE 6.9: Example signal and background fits to the $Z\gamma$ invariant mass distribution.

6.3.5 Significance calculation

In the analysis of each local sample, the background and signal are modelled through functional fits, optimised using the negative log-likelihood method. This optimization facilitates the precise capture of any signal events within the Gaussian function. Subsequent integration over the signal and background fits yields the respective numbers of events, denoted as N_s and N_b, respectively. These event counts are pivotal for computing the significance of the signal within the mass window, a crucial step in evaluating the potential for discovery.

The concept of discovery significance, Z, is a statistical measure used to ascertain the likelihood that the observed signal is not a result of background fluctuations. In simpler terms, it quantifies the 'unusualness' of the signal, providing a benchmark to distinguish genuine discoveries from statistical anomalies. The calculation of Z is grounded in the relationship between the observed signal and the expected background.

A general formula for calculating discovery significance is:

$$Z = \frac{N_s}{\sqrt{N_b}}, \tag{6.6}$$

where N_s represents the number of signal events and N_b the number of background events. However, this formula assumes an ideal scenario with perfect background estimation and no additional systematic uncertainties.

In our study, we employ a more nuanced approach to accommodate the complexities of real-world data. The significance is computed using an expanded form of the above equation, which factors in the interplay between signal and background events more comprehensively:

$$Z = \sqrt{2 \cdot \left((N_s + N_b) \cdot \log\left(1 + \frac{N_s}{N_b}\right) - N_s \right)}. \tag{6.7}$$

In this equation, the term $\log\left(1 + \frac{N_s}{N_b}\right)$ accounts for the relative proportion of signal to background, providing a more accurate representation of the significance.

Each pseudo-experiment conducted in the study generates six local significances, corresponding to each of the background rejection batches. These significances are then analysed to understand the robustness of the detected signal, and to ascertain whether what is observed is indeed indicative of a genuine physical phenomenon or merely a statistical fluctuation.

6.4 Machine Learning Data Generation Evaluation

In order to effectively generate sufficient quantity and quality of data for the given frequentist analysis, a comprehensive evaluation of various machine learning data generation techniques was undertaken. This assessment focused on three advanced methodologies: the Kernel Density Estimation (KDE), the Generative Adversarial Networks (GAN), and the Variational Auto-encoder (VAE). Each of these approaches offers distinct capabilities in producing large volumes of high-fidelity data. This section delineates the comparative analysis conducted on these methods, elucidating their respective strengths and limitations in the context of the specific data generation needs.

6.4.1 Evaluation of different data generators

Each data generation method evaluated required a base training dataset. The $Z\gamma$ final state MC dataset introduced in Chapter **??**, is used to train and evaluate each model. The dataset is constrained to the mass range of the frequentist study and a wide selection of features are used to provide an in-depth understanding of the capabilities of each model to produce realistic particle events. The selected features for the data generation study, defined in Section 6.2.3, are $m_{\ell\ell\gamma}$, $p_{T_{\ell 1}}$, $p_{T_{\ell 2}}$, p_{T_γ}, $E_{\ell 1}$, $E_{\ell 2}$, E_γ, $\eta_{\ell 1}$, $\eta_{\ell 2}$, η_γ, $\phi_{\ell 1}$, $\phi_{\ell 2}$, ϕ_γ, $\Delta R_{\ell\ell}$, E_T^{miss}, $\phi_{E_T^{miss}}$, N_j, and N_{cj}.

The quality of data produced using each method is evaluated both visually and using selected metrics. The success of each model is firstly assessed in terms of the generated feature distributions. The success of the model to produce feature distributions reflecting those of the MC training dataset, is evaluated visually, using the bin-wise relative difference and the Kolmogoroc-Smirnov score [137]. In addition to feature distributions it is vital that the generated events reflect the feature-wise correlation of the MC events. To this end the feature correlation is also compared visually, using the Spearman correlation [138] (coefficient and p-value) and the absolute correlation difference metrics. The visual and metric comparisons serve a dual purpose in our analysis. Firstly, they establish a criterion for success, which is pivotal in the optimization process of the models. Secondly, these metrics create a benchmark, enabling a systematic evaluation and comparison of the results produced by each model.

A comprehensive introduction of the evaluation metrics, as well as model development and optimisation strategies are presented in Appendix A.

6.4.2 Data generation results

Each of the data generation models were carefully constructed, trained and optimised to generate $Z\gamma$ events. The model architectures, training mechanisms and considerations, as well as the resultant feature distribution and correlation plots, for each of three methods, are exposed in Appendix A.

The three most successful methods identified were the Wasserstein GAN (WGAN), the VAE with an additional discriminator (VAE+D), and the KDE. Each of these models was used to generated 80,000 events per iteration. To facilitate a comprehensive comparison, the average result of 500 generated samples, using each method was evaluated. These results are presented in Table 6.2.

Evaluation Metric	**WGAN**	**VAE+D**	**KDE**
Relative difference	2.67×10^{-3} $(\pm 3.1 \times 10^{-5})$	4.42×10^{-3} $(\pm 2.2 \times 10^{-5})$	3.21×10^{-4} $(\pm 1.6 \times 10^{-5})$
Kolmogorov-Smirnov score	2.67×10^{-2} $(\pm 2.89 \times 10^{-4})$	5.55×10^{-2} $(\pm 3.04 \times 10^{-4})$	4.50×10^{-2} $(\pm 2.18 \times 10^{-4})$
Spearman correlation coefficient	6.96×10^{-1} $(\pm 5.99 \times 10^{-3})$	6.23×10^{-1} $(\pm 7.07 \times 10^{-3})$	9.38×10^{-1} $(\pm 1.37 \times 10^{-2})$
Spearman p-value	3.43×10^{-23} $(\pm 5.19 \times 10^{-23})$	1.38×10^{-17} $(\pm 1.67 \times 10^{-17})$	8.95×10^{-54} $(\pm 1.84 \times 10^{-52})$
Correlation difference	2.14×10^{-2} $(\pm 2.84 \times 10^{-4})$	2.62×10^{-2} $(\pm 3.27 \times 10^{-4})$	2.47×10^{-3} $(\pm 2.11 \times 10^{-4})$

TABLE 6.2: Resultant comparison of the WGAN, VAE+D, and KDE data generation methods. The evaluation metrics value's mean (standard deviation) are calculated for 500 generated samples of 80,000 events.

This comparative analysis quantifies the ability of each data generation technique. The three best performing methods where found to be the WGAN, VAE+D, and KDE methods. Each method demonstrates a high proficiency in accurately simulating $Z\gamma$ events. Within the scope of this research, which concentrates on investigating semi-supervised classifiers, the KDE approach is identified as the most appropriate. The selection of the KDE method is primarily due to its exceptional performance in delivering high-quality results. Notably, compared to other methods, KDE introduces significantly less potential error. This aspect is crucial, particularly in influencing the outcomes of frequentist study results. The method's remarkable accuracy and reduced error propensity make it ideal for our research. These attributes render the KDE method particularly beneficial, aligning seamlessly with the specific objectives and methodological rigour of the study.

6.5 Frequentist Study Results

In order to measure the extent of look-elsewhere effects introduced by semi-supervised NN classifiers, the pseudo-experiment was repeated 125,000 times and the results were analysed. In this section, the resulting components of an example pseudo-experiment are first explored before the final frequentist study results are presented and discussed.

6.5.1 Pseudo-experiment results

In this study, 125,000 independent pseudo-experiments are conducted. Before the final study results can be evaluated, each component of the pseudo-experiment must be assessed.

Data sampling

The first component of the experiments is the sampling or generation of a statistically accurate and independent dataset. The evaluation and final selection of the model used to achieve this is presented in Section 6.4. The KDE model was selected as the method to generate samples for each experiment in this study. Therefore, for each given pseudo-experiment, the KDE is used to generate 80,000 independent $Z\gamma$ events. An example of the feature distributions and event-wise correlation plots of generated events are exposed and compared to the MC training dataset in Figures 6.10 and 6.11, respectively.

Semi-supervised NN training

The generated dataset is used to train the semi-supervised NN classifier. The NN is trained on all the generated features except the invariant mass which is used to define the signal and background regions as well as for the resonance search. As there is no signal introduced when training the model, the model is expected to find no significant discrimination between mass-window and sideband events. This is shown in Figure 6.12a, where the signal (mass-window) and background (side-band) response distributions show negligible separation. This is reinforced by the ROC curve presented

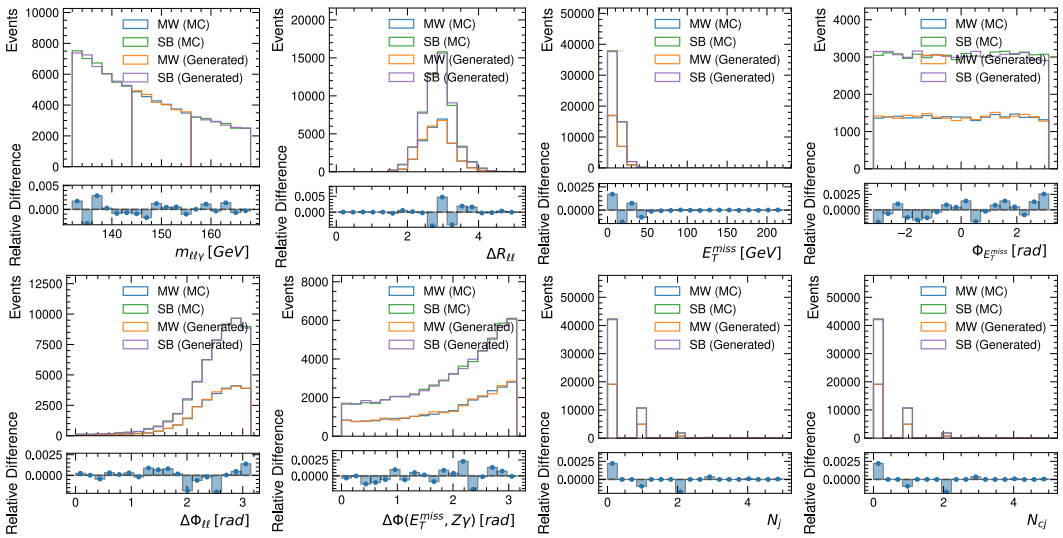


FIGURE 6.10: Example generated feature distributions for pseudo-experiment using the KDE method.

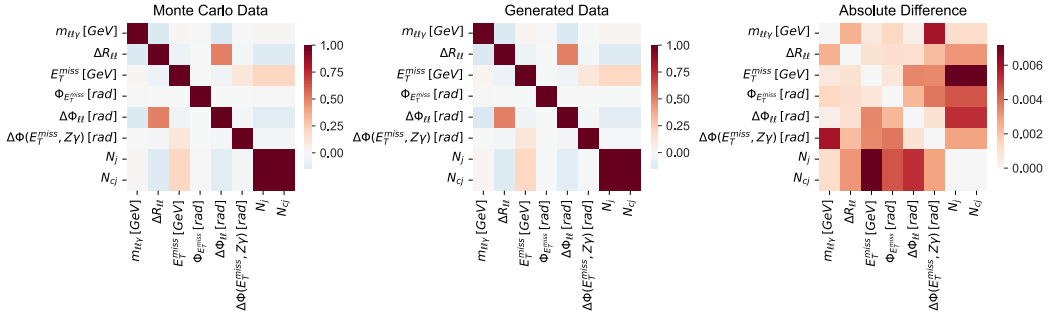
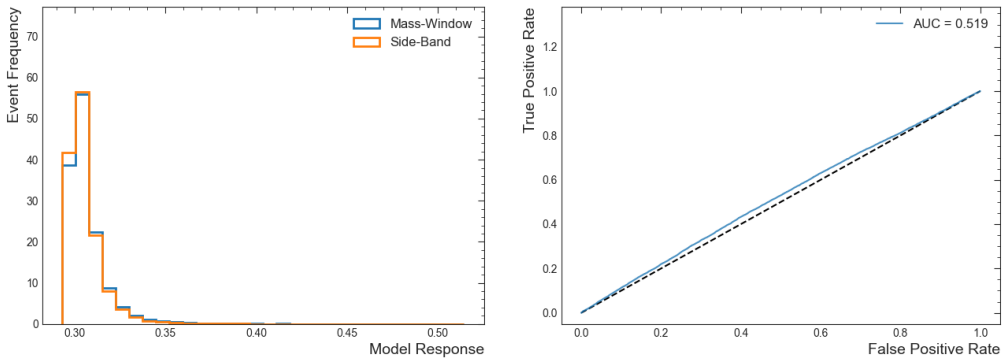


FIGURE 6.11: Example generated feature correlation comparison for pseudo-experiment using the KDE method.

in Figure 6.12b, that shows an AUC of 52%. This AUC exposes that the model is approximately as likely to label an event as signal or background.

Furthermore, the features selected for the training of the NN must be evaluated in terms of their influence on the NN model. Features with high model influence may be correlated to the invariant mass, which would enable the NN to distinguish between the signal and background regions. In order to assess the feature influence, the SHapley Additive exPlanations (SHAP) feature ranking is used [139]. Figure 6.13 exposes that the final selected kinematic features have negligible SHAP scores and therefore have no significant correlation with the invariant mass.



(a) (b)

FIGURE 6.12: Example pseudo-experiment response distribution (a) and ROC curve (b) for the trained semi-supervised neural network.

Invariant mass fits

For each local background rejection sample, the invariant mass distribution is fit with the background (null hypothesis) and signal+background PDF (alternative hypothesis) models. The minimising of these fits exposes any signal found in the mass window of each sample. Figure 6.14 shows an example of the local fits for a single pseudo-experiment iteration.

The example pseudo-experiment's local background rejection fits, expose expected distributions with minimal signal significance. The likelihood of attaining fluctuations in the fitting data around 150 GeV should occur at normal probabilities. The local signal significance values in all conducted pseudo-experiments ranges between -3.5σ and 3.5σ.

6.5.2 Final results and discussion

In order to measure the additional look-elsewhere effect originating from the implementation of the semi-supervised classifier, a frequentist approach is used. To this end, the pseudo-experiment is repeated 125,000 times to obtain necessary statistics. This means that for each pseudo-experiment the KDE model, pre-trained on the $Z\gamma$ MC-generated dataset of 110,000 events, is used to generate a new sample of 80,000 statistically independent events. These events are then used to train the semi-supervised NN. Once

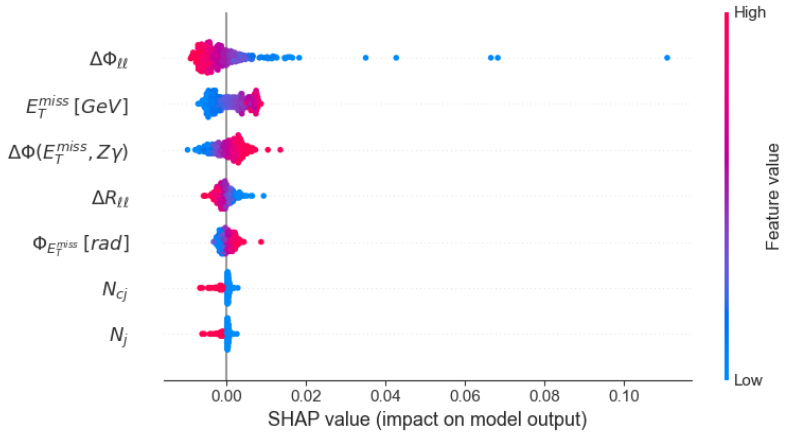
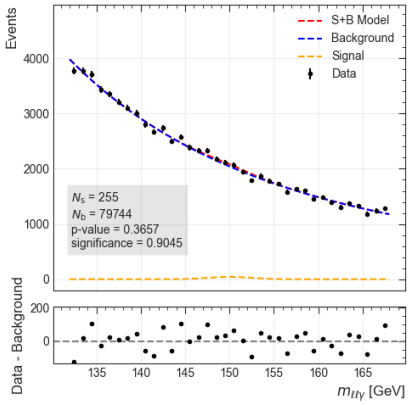


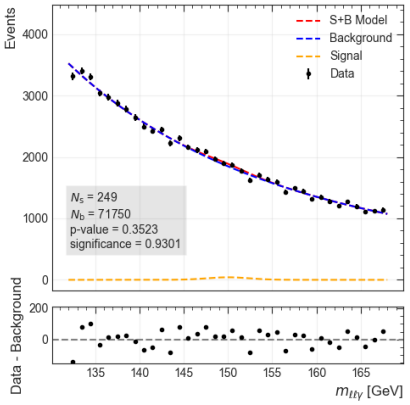
FIGURE 6.13: Example NN SHAP feature ranking for single pseudo-experiment iteration.

trained, the event samples for the background rejection cuts of 0%, 10%, 20%, 30%, 40% and 50% are extracted via the NN response distribution. The invariant mass distributions of these samples are then used to calculate the significance of a narrow resonance at 150 GeV. Therefore, each pseudo-experiment will produce six local significances, corresponding to each of the background rejection thresholds.

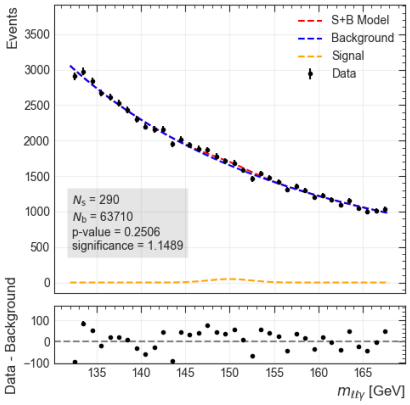
This methodology, similar to mass scans in high-energy physics, segments our analysis through the use of background rejection samples, where each threshold acts as a local category. This approach facilitates an examination of the impact of background rejection levels on our results, enabling a nuanced understanding of their influence on our findings. The resulting distributions of the local significances for each of the background rejection cuts (BR) are shown in terms of their probability density function (PDF), $f(Z)_{\mathrm{BR}}$, in Figure 6.15a. These distributions have an approximately Gaussian shape, however with a standard deviation between 0.65 and 0.71. This decrease in variation, with respect to the standard Gaussian, is due to the constraints on the resonance fit imposing a priori via the shape of the resonance and the side-bands. Note that the peak of the distributions shifts to the right with increasing background rejection. Such a shift emerges from the interplay between signal and background modelling, amplified by the diminishing event counts in samples subjected to higher background rejection cuts.



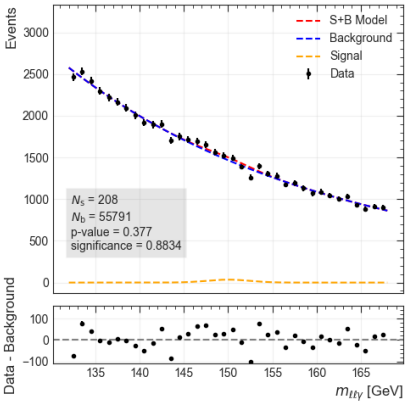
(a) 0% Background Rejection



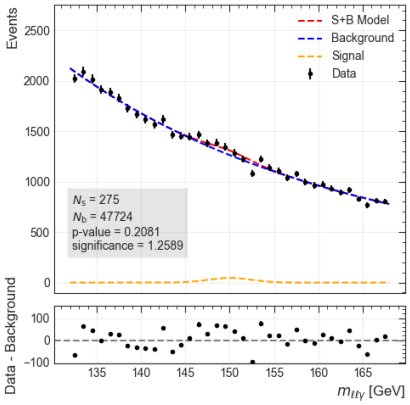
(b) 10% Background Rejection



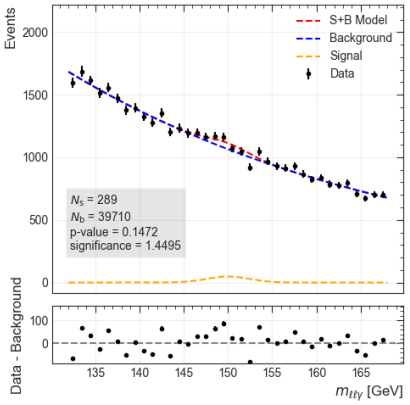
(c) 20% Background Rejection



(d) 30% Background Rejection

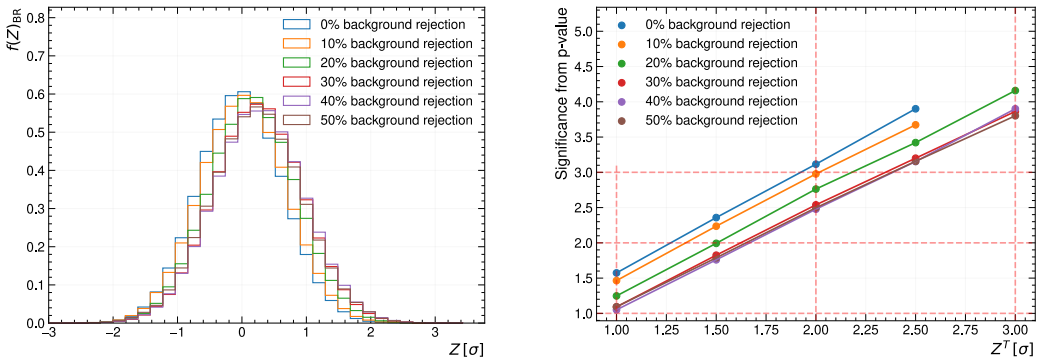


(e) 40% Background Rejection



(f) 50% Background Rejection

FIGURE 6.14: Example pseudo-experiment signal and background fits for background rejections of 0% (a), 10% (b), 20% (c), 30% (d), 40% (e) and 50% (f).



(a) (b)

FIGURE 6.15: Analysis of local frequentist study results. Distribution of local signal significance values for each local background rejection sample (a) and local p-values (converted to significance) for each local background rejection sample compared to a standard Gaussian (b).

To further interpret these results and quantify the deviations from a Gaussian shape, the probability of getting a local significance of up to 1σ, 2σ and 3σ can be calculated for each BR cut sample. For this, we consider only positive values of the significances for an excess at 150 GeV, as only they can reflect instances where a fake signal is generated. The corresponding normalised probability density function is:

$$f^+(Z)_{\text{BR}} = \frac{f(Z)_{\text{BR}}}{\int_0^\infty f(Z)\,dZ}, \qquad \text{for } Z > 0, \tag{6.8}$$

where BR is a label indicating the local background rejection cut. We then compute the probability that, for each background rejection cut, a significance of up-to a value $Z^T = 1\sigma$, 2σ and 3σ is obtained via the cumulative distribution function (CDF):

$$F(Z^T)_{\text{BR}} = \int_0^{Z^T} f^+(Z)_{\text{BR}}\,dZ. \tag{6.9}$$

Therefore, the corresponding p-values calculated as $1 - F(Z^T)_{\text{BR}}$, quantify the probability of detecting an excess of up-to Z^T. These p-values can be converted to significance, before being directly compared to Z^T, corresponding to if the distribution was normally distributed. This relationship is

shown in Figure 6.15b, where one can see that the local samples conform to a large degree with expected probabilities from a Gaussian distribution. Only at higher significance thresholds above $\approx 2.5\sigma$ slight deviations from the diagonal are seen. It is also shown that for background rejection cuts of 0% and 10%, no 3σ significance excesses were found during the 125′000 iterations.

To understand the extent of additional look-elsewhere effect introduced, the 0% BR sample can be used as a baseline. This inclusive sample contains all the events classified by the NN without any cuts applied, and therefore provides a sample unaffected by the influence of the classifier and its response scan. The additional look-elsewhere effect can thus be revealed through a comparison of the samples that include the influence of the NN, from response cuts, to the inclusive baseline.

The "global" sample which consists of the statistics from all background rejection categories is calculated as:

$$f(Z)_{\text{global}} = \sum_{i=1}^{6} f(Z)_i, \tag{6.10}$$

where i represents the different background rejection cuts. It is shown in Figure 6.16 from where one can see that it approximates a Gaussian distribution with a standard deviation of 0.69 centred around a mean of 0.15. The corresponding global p-value is calculated as

$$p_{\text{global}}(Z^T) = 1 - \int_0^{+Z^T} f^+(Z')_{\text{global}} dZ', \tag{6.11}$$

where f^+ is defined analogously to Equation 6.8.

The comparison of the p-values from the different BR cuts sample and the corresponding global PDF are compared to the 0% BR cut p-value to quantify additional look-elsewhere effects. This means a trials factor is defined as the ratio of these p-values, as shown in Figure 6.17. For lower significances, up to approximately 1.5σ, the trials factor is close to 1, implying minimal inflation in the observed significance. However, at higher significances, of greater than 2σ thresholds, the trials factor displays elevations, indicating

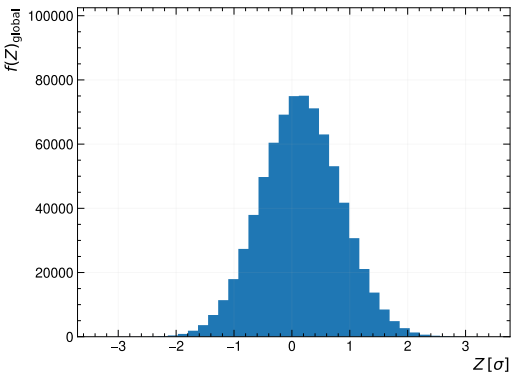


FIGURE 6.16: Distribution of the significance from the global sample, i.e. ensemble of outcomes across all background rejection cuts.

an inflation of significance due to the NN procedure.

Therefore, despite the inherent challenges for resonant searches using weak supervision, the depicted results demonstrate the efficacy of these semi-supervised techniques and their ability to maintain additional LEEs within acceptable bounds. This means that the extent of false signals introduced by the NN are under control.

With the trials factor established, we adjust the global significance levels for the LEE to accurately assess the rate of false discovery. In the statistical analysis of experimental data, particularly when multiple hypotheses are tested simultaneously, the need to control for the inflation of Type I error is paramount. The Type I error, or the false positive rate, is the probability of incorrectly rejecting a null hypothesis when it is in fact true. The inflation of this error rate is a concern when multiple comparisons are made because the likelihood of observing at least one statistically significant result due to chance alone increases with the number of comparisons. This necessitates the use of adjustments to maintain the overall error rate at a desired significance level. Two such adjustments are the Bonferroni correction and the Gross-Vitells adjustment, each with its own methodology and context of application.

The Bonferroni Correction is one of the simplest and most conservative methods to control the family-wise error rate, which is the probability of

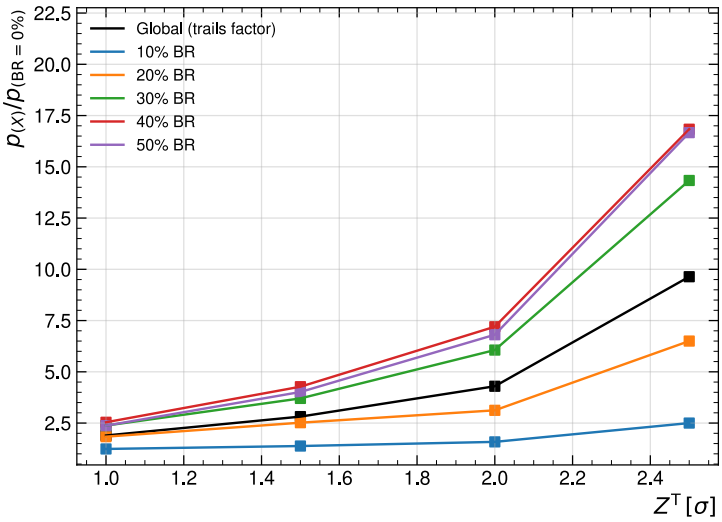


FIGURE 6.17: The ratio $p(\mathrm{X})/p(\mathrm{BR} = 0\%)$ for $X = 10\%, 20\%, 30\%, 40\%, 50\%$ BR cuts and $p_{\mathrm{global}}/p(\mathrm{BR} = 0\%)$, corresponding to the trails factor, as a function of the significance threshold, Z^T.

making one or more Type I errors among all the hypotheses being tested. The Bonferroni method involves dividing the desired overall alpha level (the probability of a Type I error) by the number of hypotheses tested. For instance, if an experiment involves testing 100 hypotheses and the desired alpha level is 0.05, the Bonferroni correction would adjust the significance threshold for each individual test to 0.0005. This method ensures that the cumulative probability of making at least one Type I error across all tests does not exceed the overall alpha level. However, the conservative nature of the Bonferroni correction can lead to a decrease in statistical power, which is the probability that the test correctly rejects a false null hypothesis.

The Gross-Vitells adjustment, on the other hand, is a more sophisticated technique that considers the spatial structure of the statistical tests. Originating from the field of random field theory, the Gross-Vitells method takes into account the correlation between test statistics across the search space, which is often the case in high-dimensional datasets like those analysed in particle physics. This method uses the expected Euler characteristic of the excursion sets above a certain threshold to estimate the effective number of independent tests, which is generally smaller than the actual number of

tests due to correlations. The adjustment made by the Gross-Vitells method is less stringent than the Bonferroni correction and can lead to higher power while still controlling the family-wise error rate. The Bonferroni and Gross-Vittels adjusted global significances are compared to the unadjusted significance in Figure 6.18. This exposes the rate of false discovery, while accounting for the LEE.

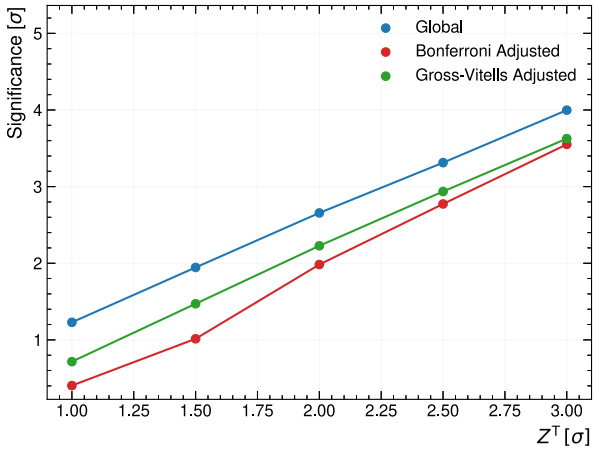


FIGURE 6.18: Comparison of unadjusted global significance to Bonferroni and Gross-Vitells adjusted global significance as a function of local significance.

In Figure 6.18, the blue line represents the unadjusted global significance, which rises linearly with local significance. When we apply the Bonferroni correction (red line), the global significance values are significantly reduced. The Gross-Vitells method (green line), offers a less conservative but more nuanced adjustment, suggesting that some significance can be retained without inflating the Type I error rate as much as the Bonferroni method does.

At low local significances, the Bonferroni adjustment displays a divergence from the unadjusted significance by up to 0.9σ, suggesting a highly conservative approach that may overestimate the probability of error. As the local significance increases, this deviation lessens, reaching a minimal disparity of 0.3σ at high local significances. This pattern indicates that although the Bonferroni method is robust in controlling the family-wise error rate, it

could diminish the test's sensitivity to detect true effects, especially when local significance is low.

Conversely, the Gross-Vitells adjusted global significance, illustrated by the green line in Figure 6.18, ascends more incrementally with local significance. This adjustment, by accounting for correlations inherent in the data, offers a nuanced moderation of the error rate. It diverges from the unadjusted global significance by no more than 0.4σ at low local significance levels, reducing to a marginal 0.3σ, as local significance grows. These values demonstrate the Gross-Vitells method's effectiveness in accounting for the LEE without unduly compromising the error rate. The approach underscores the efficacy of semi-supervised models in maintaining reasonable control over LEE and error rates, essential for robust statistical inference in high-dimensional data analyses.

In conclusion, this frequentist study meticulously evaluates the LEE and the prevalence of false signals during the semi-supervised NN classifiers' training for $Z\gamma$ final state signal isolation. The local and global significance distributions maintain alignment with Gaussian distributions, highlighting the minimal introduction of false signals throughout the training process. Our analysis demonstrates that the LEE is negligible at lower levels of significance, affirming the classifier's reliability in these regions. However, as local significance intensifies, the LEE becomes significant, underscoring the necessity for proper adjustments. The incorporation of the trials factor into our analysis has been instrumental in providing critical insights into the scale of adjustment required to accommodate the LEE. It reveals a persistent adjusted difference of $\pm 0.35\sigma$ and a slight increase in LEE at low local significance levels, emphasizing the impact of high-significance local fits on the trials factor results, as depicted in Figure 6.17.

Furthermore, the convergence of the empirical data with theoretical adjustments via the Gross-Vitells method highlights the robustness of the analytical approach. It controls the potential for false signals attributed to the LEE, thereby ensuring the results' significance is not overstated. The Gross-Vitells adjustments not only provide a refined assessment of the LEE but

also uncover the intricate relationship between LEE and the observed significances. This dual analysis of the trials factor and the adjusted significance levels constructs a comprehensive framework that guards against the over interpretation of the observed phenomena, thus bolstering confidence in the veracity of the potential discovery of true signals. The study, through its nuanced consideration of LEE adjustments, reinforces the semi-supervised model's efficacy in classifying $Z\gamma$ events and underscores the importance of such statistical considerations in high-energy physics research.

Chapter 7

Conclusions

The quest for understanding the fundamental constituents of the universe and their interactions has been a driving force for theoretical and experimental physics. Central to this pursuit is the LHC at CERN, where the ATLAS detector serves as one of the primary instruments probing the frontiers of our knowledge. This dissertation has presented a detailed study within this context, focusing on searches for new bosons, using semi-supervised classifiers. This research is motivated by the multi-lepton anomalies found at the LHC, and focuses on the $\mathrm{S}(150) \to Z\gamma$ final state.

While model-dependent searches are crucial, they inherently rely on specific theoretical assumptions that may not encompass the full spectrum of possible new physics scenarios. This recognition has spurred the exploration of model-independent searches, which aim to be less reliant on theoretical preconceptions and more inclusive of unexpected phenomena. An innovative approach within this paradigm is the application of semi-supervised machine learning techniques. These techniques leverage both labelled and unlabelled data, making fewer assumptions about the underlying model, thereby providing a more general strategy for new physics searches. This dissertation has not only implemented these semi-supervised techniques within the context of the $2HDM + S$ model but has also critically assessed their efficacy and robustness. It is evident that these methods offer a reduced model dependency, opening a wider aperture for potential discoveries.

This concluding chapter will reflect on the journey taken through the ATLAS data within the $2HDM + S$ landscape, highlighting the novel application of semi-supervised learning. It will discuss the key findings, their implications for the field, and the potential pathways that future research could take to discover new physics BSM.

The base investigation into semi-supervised learning, as reported in Chapter 5, stands as a cornerstone of this thesis. Here, a rigorous comparison was drawn between semi-supervised and fully-supervised machine learning classifiers, with a focus on the classification of di-lepton signals. The analysis employed three prominent machine learning architectures: BDT, MLP and DNN, each renowned for their prowess in pattern recognition tasks. Key findings reveal that both BDT and DNN models excel in their respective techniques, showcasing AUC scores of 0.855 and 0.845 in semi-supervised settings, against 0.859 in fully-supervised ones. The proximity in performance between the semi- and fully-supervised approaches underscores the potential of semi-supervised methods to closely match the predictive strength of their fully-supervised counterparts while utilizing less model dependency.

Although both BDT and DNN models demonstrate strong performance, the DNN was selected for ongoing work due to its marginally better resistance to overfitting and its inherent flexibility in model architecture — advantages that, while modest, are nonetheless valuable in the high-stakes pursuit of potential new physical phenomena. The consolidation of this research into semi-supervised methods, particularly within the realm of DNNs, not only demonstrates their viability but also sets the stage for their more widespread adoption in particle physics research.

For the intricate tasks associated with HEP analysis, particularly those concerning the evaluation of systematic uncertainties and the look-elsewhere effect, the generation of sufficient and realistic synthetic physics data is of paramount importance. This necessity was addressed by evaluating several advanced data generation techniques, and comparing their abilities to generate events. The results of this comparison, presented in Section 6.4, compare the KDE, WGAN-gp, and VAE+D. Each of the methods are shown

to generate high quality synthetic datasets that mimic the statistical properties of genuine physics events.

However, the focus on the precision required for frequentist study methodologies led to the conclusion that KDE, with its minimal error introduction, was the most appropriate choice for this thesis. Despite the more sophisticated generative capabilities of WGAN, the KDE provides the balance of accuracy and computational efficiency needed for the rigorous evaluation of false signals and LEE in semi-supervised classifiers. The specific attributes of KDE that contributed to its selection include its simplicity, interpretability, and lower propensity to introduce systematic errors into the analyses, which are critical when conducting precise statistical studies.

The frequentist methodology developed and utilised within this research critically evaluated the efficacy of semi-supervised NN classifiers within the realm of narrow resonance searches in high-energy physics, gauging their tendency to produce additional look-elsewhere effects in the form of fake signals. This additional look-elsewhere effect, originating from the scan of the NN response distribution, can be considered a consequence of model independent searches. In this study, the semi-supervised NN classifier, was developed and analysed to identify false signals and measure the LEE in high-energy physics, particularly for $\mathrm{S}(150) \rightarrow Z\gamma$ final state signal isolation. The methodology was tested using $125,000$ pseudo-experiment iterations, yielding significant insights into the behaviour of local and global significance distributions and the impact of LEE on statistical analyses.

The frequentest study results expose local significance distributions generally follow a Gaussian pattern, but shift right with increased rejection of background events. At higher significance levels, particularly around 2.5 and 3σ, the results indicated potential deviations due to insufficient statistics, emphasizing the rarity of such high significance levels.

The global significance, computed as a sum of local distributions, was found to be normally distributed around zero. However, when considering the LEE through a comparative analysis of local and global p-values, it became evident that global p-values show reduced significance, indicating an inflation in observed significance due to multiple testing, especially at higher

levels.

To address this, we employed a trials factor, as the ratio of global to inclusive p-values. While the methodology demonstrates that this effect is very small at lower significance values, it is enhanced with decreasing p-values. Despite these complexities, the difference between the baseline with 0% BR cut and p-values for higher BR cuts remain limited, showing the method's proficiency in classifying events. Critically, the results corroborate that semi-supervised NN classifiers introduce controlled fake signals, cementing their viability in HEP analyses.

The analysis was further refined by applying the Bonferroni and Gross-Vitells adjustments to control for Type I error inflation. The Bonferroni method, while conservative, significantly reduced global significance values, potentially affecting the test's sensitivity. Conversely, the Gross-Vitells method provided a more balanced approach, showing less deviation from unadjusted significances and better accounting for data correlations.

The analysis demonstrates the effectiveness of semi-supervised models in controlling LEE and error rates in high-dimensional data analyses. The integration of the trials factor and the application of statistical adjustments like the Gross-Vitells method have been crucial in understanding the impact of the LEE. These approaches ensure that the significance of results is not overstated, thus maintaining the reliability and accuracy of the statistical inferences in the search for true signals in high-energy physics.

In conclusion, the study underscores the importance of considering LEE and statistical adjustments in high-energy physics research. It highlights the robustness of semi-supervised NN classifiers in handling complex datasets, and their potential in aiding the discovery of true signals, while minimising model dependency thereby contributing significantly to the field. The results presented in this thesis therefore establishes a comprehensive foundation for future studies to more critically evaluate the use of machine learning classifiers and thus implement these semi-supervised techniques in extracting subtle signals from data produced at the ATLAS experiment.

Appendix A

Evaluation of the WGAN data generation model

A.1 Introduction

This appendix extends the Monte Carlo and data generation methods introduced in Section 4.6. Section 6.4 is dedicated to a comprehensive evaluative comparative study, assessing the efficacy of various machine learning data generators in synthesizing LHC data. The introduction of this chapter introduces the foundational concepts and methodologies underpinning the Kernel Density Estimation (KDE), Variational Autoencoders (VAE), and Generative Adversarial Network (GAN) models, alongside the metrics used for their evaluation.

The second segment of the chapter is devoted to an in-depth exploration of the Wasserstein GAN with gradient penalty (WGAN). This section aims to offer a detailed analysis of the development, optimization, and results of the WGAN model. This comprehensive examination is pivotal in understanding the model's applicability and efficiency in generating high-energy physics data, thereby contributing to the field of machine learning in scientific data generation.

A.1.1 Kernel Density Estimator

The Kernel Density Estimator (KDE) introduced in Section 4.6.2 is selected as a promising unsupervised fitting method. As the KDE model is fit to

the training dataset, the only requirement when implementing the method is to select the optimum bandwidth and kernel function. When comparing different kernel functions on $Z\gamma$ feature distributions, it was determined that the Gaussian function was optimal. The bandwidth parameter was selected by scanning over a large range of bandwidths to determine the best value.

A.1.2 Variational Auto-Encoder

The development and optimisation of the VAE and VAE+D was conducted in parallel. The architecture, hyperparameters and training strategies are summarised in Ref. [140]. The VAE+D provides a GAN like VAE model that utilises a discriminator network, which not only enhances the model's ability to generate high-fidelity data but also improves its performance in learning complex data distributions. This integration leads to a more efficient and effective generative model, capable of tackling a broader range of applications compared to its traditional counterparts.

A.1.3 Generative Adversarial Network

The Generative Adversarial Network with focus on its variation that includes the Wasserstein loss and gradient penalty, was selected as the machine learning based data generator with the most potential for this application. The GAN and WGAN-gp are introduced in detail in Sections 4.6.3 and 4.6.4, respectively. This Chapter provides an in-depth analysis of the implementation and optimisation of the training and testing of the WGAN-gp model.

A.1.4 Evaluation metrics

Evaluation metrics are selected to measure and compare the success of each data generation method on the $Z\gamma$ dataset. The evaluation metrics consider success as generated events that accurately reflect the real particle events feature distributions and correlations. To evaluate the success of each of the generated feature distributions to reflect real physics the relative difference and Kolmogorov-Smirnov test are employed. The analysis of the

generated invariant mass, $m_{\ell\ell\gamma}$, is of specific interest as it will be used as the base of signal and background fits. The quality and shape of the generated $m_{\ell\ell\gamma}$ distribution must therefore be smooth and consistent. Furthermore, as the study will confront sidebands with mass window, the distributions of each feature can use a more focused comparison between MC and generated events in the mass-window and sideband respectively.

The Relative Difference serves as a metric for relativistic comparison between the binned feature distributions of MC data and generated data. This measure quantitatively assesses the divergence between the two distributions. The relative difference is calculated using the formula:

$$\frac{p_{MC} - p_G}{\max(p_{MC})}, \tag{A.1}$$

where p_{MC} represents the binned feature distribution of the MC data, and p_G corresponds to that of the generated data. This formulation normalises the difference by the maximum value in the MC distribution, providing a scaled perspective of the discrepancy between the distributions.

The Kolmogorov-Smirnov Test (KS), in its two-sample variant, is used to determine if there is a significant difference between two empirical distributions, which are denoted as p_{MC} and p_G. This test is especially valued for its non-parametric nature, meaning it does not rely on assumptions about the form of the underlying distributions. The KS test compares the cumulative distribution functions (CDFs) of the two distributions, identified as $F_{p_{MC}}(x)$ for p_{MC}, and $F_{p_G}(x)$ for p_G:

$$D = \sup_x |F_{p_{MC}}(x) - F_{p_G}(x)|. \tag{A.2}$$

Here, the KS statistic D is defined as the supremum of the absolute differences between the CDFs of p_{MC} and p_G over all values of x. The supremum function, denoted as $\sup_x$, effectively captures the largest difference between $F_{p_{MC}}(x)$ and $F_{p_G}(x)$. This approach quantifies the extent of divergence between the two empirical distributions, with the KS test focusing on the maximum discrepancy across the entire range of the combined data

set. The KS test is implemented using the **SciPy** library [141] which follows Hodges' treatment [137].

Although these two tests provide statistics on the success of the generated feature distributions, generated events must reflect the accuracy of the feature wise correlation. Therefore, in order to evaluate if the feature correlation of events is reflective or real physics the absolute correlation difference and Spearman correlation test are employed.

The Absolute Correlation Difference is measured as the absolute difference between the feature correlation of the MC data minus the generated data. This provides a general measurement of the extent to which the generated data feature correlation presents real physics. It also exposes the features with the highest and lowest correlation difference, providing an in-depth feature wise evaluation of the generated event correlation.

The Spearman Correlation Test is a rank-order correlation coefficient used to evaluate feature correlations between the generated data events and the MC events. This non-parametric measure of rank correlation assesses the monotonic relationship between features, offering insights into the predictive consistency of generated data against MC events. Its suitability for handling non-linear relationships and outliers is advantageous for the complex HEP data. The test is implemented using the **SciPy** library [141], and is critical for understanding feature correlations.

A.2 Wasserstein Generative Adversarial Network

The Wasserstein Generative Adversarial Network (WGAN), introduced in Section 4.6.4, is an extended version of the GAN to generate superior quality data while reducing the effects of vanishing gradients. The model uses two competing neural networks to converge on one being able to generate realistic data and the other to classify if an event reflects real physics or not. The complexity of training the WGAN originates in the fact that the optimisation of each of the two networks influences the ability of the other. Therefore, the two models must be optimised simultaneously, increasing the complexity of optimisation multifold.

The implementation of the WGAN-gp model is executed using the PyTorch library, a powerful and flexible deep learning framework in Python [142]. This model is based on the architecture of the WGAN described in the Ref. [118].

To ascertain the viability of WGAN in simulating LHC events, a meticulous optimisation of the model is imperative. This includes an in-depth analysis of the model's results and the influence of parameters on that result. This section will firstly present the development process of the model, outlining the strategies employed for its optimisation, and discusses the major factors influencing its performance. Following this, we present a detailed exposition of the final model that was chosen after rigorous evaluation.

A.2.1 Model development and optimisation

Despite the interdependent training of the Critic and Generator models, it is imperative to enhance each model's structures and architectures independently for optimal performance. This section delves into the multifaceted journey of model development for the WGAN, involving many training iterations evaluating various architectures, hyperparameters, training mechanisms, and data processing techniques. A summary highlighting the most impactful modifications, will be presented.

The base components of the implemented model are the dataloader, critic model, generator model and the main WGAN-gp or trainer model. The dataloader facilitates storing and preprocessing the training dataset as well as sampling batches of data during training. The critic and generator models are neural networks configured to classify and generate synthetic events respectively. Finally, the trainer model is used to initialise and implement the WGAN-gp training using the dataloader and neural network models.

To monitor and evaluate the training process of various model configurations, key metrics such as loss are meticulously tracked using TensorBoard [143]. This tool enables the detailed visualization of training dynamics, facilitating insightful comparisons across different model runs.

Additionally, the model saves progress distribution and correlation plots at regular intervals, providing a visual representation of its learning trajectory. To further refine and optimise the WGAN-gp model, for generating $Z\gamma$ data, a systematic approach is employed. This involves controlled scans of various parameters and configurations, allowing for precise adjustments and enhancements to the model. Through this methodical process, the aim is to incrementally improve the model's performance and adaptability in generating high-fidelity $Z\gamma$ data, ensuring it meets the specific requirements and standards of the application.

Building upon the foundational setup for monitoring and optimizing the WGAN-gp model, it is crucial to understand the training cycle of the base model, particularly focusing on the concept of an epoch. An epoch in the context of training a GAN, like the WGAN-gp, represents one complete cycle through the entire training dataset. During each epoch, the WGAN-gp trainer model undergoes a single training iteration.

For a given epoch the WGAN trainer model uses the dataloader to sample the entire training dataset in batches. The size of these batches is determined by the hyperparameter, batch-size. Each batch of training events is used for a critic training iteration and a generator training iteration.

During a critic training iteration, the gradient penalty and critic loss are calculated for the current critic and generator outputs using Equations 4.31 and 4.30. The critic model weights are updated using the selected optimiser on the calculated loss.

During a generator training iteration, the generator loss is calculated for the current critic and generator output using Equation 4.29. The generator model weights are updated using the selected optimiser on the calculated loss. In initial training epochs the generator model is only trained once for every five critic training iterations. This allows the critic model to discriminate with sufficient quality to best inform the generator model updates.

Using this base model, various optimisation scans of architectures, activations, hyperparameters and algorithms can be assessed. The key results to these experiments are presented in Section A.2.2. One such experiment,

to evaluate the influence of increasingly complex architectures will be exposed.

Example WGAN architecture optimisation

In the example experiment, four different configurations of hidden layers and nodes of increasing complexity where designed for the generator and critic models independently. These are summarised in Table A.1.

Configuration	Number of nodes per hidden layer	
	Critic	Generator
1	$[D, D, D]$	$[D, D, D]$
2	$[D, 2 \cdot D, D]$	$[D, 2 \cdot D, D]$
3	$[D, 2 \cdot D, 2 \cdot D, D]$	$[D, 2 \cdot D, 2 \cdot D, 2 \cdot D, D]$
4	$[D, 2 \cdot D, 4 \cdot D, 2 \cdot D, D]$	$[D, 2 \cdot D, 4 \cdot D, \cdot DD, D]$

TABLE A.1: Table showing example WGAN architecture optimisation for increasing complexity of hidden layers. Where D is the input dimension selected as 256 for given example.

The critic and generator model configurations influence the training of one another and therefore to gauge the influence of the different configurations, each combination of generator and critic combinations must be tested. Each model combination is trained using the same fixed hyperparameters for 500 epochs. Each combination is denoted as $G_j C_k$, where j and k are the generator and critic configurations respectively. The trained models are then evaluated and compared on their; critic loss, generator loss, gradient penalty, generated feature distributions, and generated feature correlations.

The **critic loss** per epoch serves as an indicator of the critic model's convergence. Typically, critic loss values are negative and approach zero as they converge. Figure A.1 illustrates the critic loss per epoch for each model. In the initial 50 epochs, each model's critic loss reaches its maximum negative value. Models with more complex critic networks exhibit larger dips, whereas those with simpler networks show minimal troughs. Post reaching their peak negative values, the critic loss in all models tends towards zero. Notably, models with higher complexity demonstrate a more gradual convergence, while simpler models converge more rapidly.

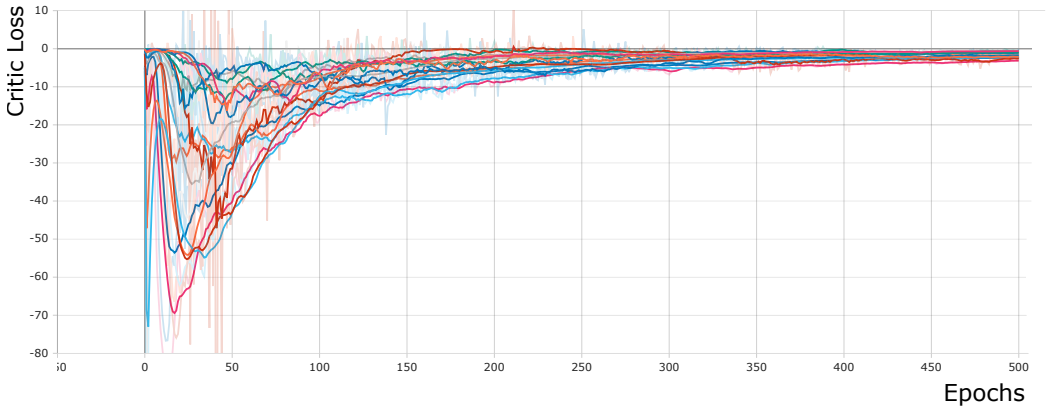


FIGURE A.1: Plot comparing example model critic loss per epoch.

A detailed comparison, as depicted in Figure A.1, reveals significant insights about the critic model configurations. Configurations 1 and 2, lacking in complexity, converge swiftly but only to a mediocre level of quality. Configuration 4, despite its large initial dip indicating substantial learning, has difficulty converging successfully, suggesting a potential limit to beneficial complexity. Finally, configuration 3 emerges as the most effective, striking an optimal balance with its gradual convergence and achieving the highest quality in the resultant models. This underscores that an intermediate level of complexity in the critic network, as seen in configuration 3, provides the most favourable outcomes for accurately classifying events in this context.

The **generator loss** per epoch similarly serves as an indicator of the generator model's convergence. The generator loss values are positive and form an initial peak, for early epochs, before tending towards zero and convergence. Figure A.2 compares the example model configurations in terms of the generator model loss per epoch.

In the comparative analysis of generator loss as depicted in Figure A.2, several patterns emerge, highlighting the performance of different model configurations. Notably, the configurations G_3C_1 and G_4C_1, depicted by green and blue lines respectively, display loss values consistently below zero. This suggests an insufficiency in the critic model, potentially limiting the effectiveness of training.

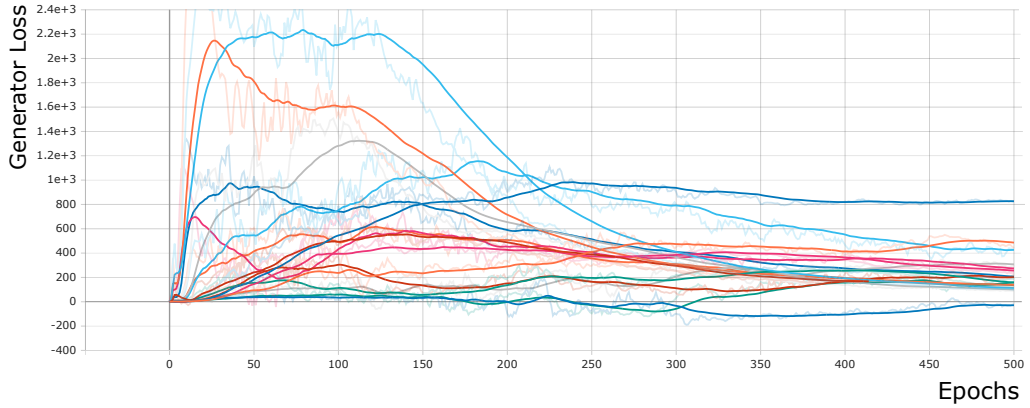


FIGURE A.2: Plot comparing example model generator loss per epoch.

Additionally, the middle dark blue line, representing the G_1V_2 configuration, demonstrates a notable inability to converge. Its loss trajectory diverges from the expected stabilization path, indicating potential issues in its parameter settings or training approach.

In contrast, the light blue (G_1C_4) and orange (G_2C_4) configurations exhibit the largest initial peaks in generator loss, due to a combination of low complexity generator model and high complexity critic model. Despite this, they show tendencies towards decent convergence.

Among the models analysed, the grey (G_2C_3) and light blue (G_1C_4) configurations stand out for their efficient performance. These models maintain loss values within an optimal range, steadily approaching convergence. This consistent performance suggests a well-calibrated balance in the generative adversarial network, highlighting their effectiveness in the model configuration.

The implementation of **gradient penalty** in training WGANs serves as a critical mechanism for enforcing the Lipschitz constraint on the critic. By penalizing the gradient's norm deviating from a target value, typically set to 1, the gradient penalty ensures a smoother and more stable training process. The analysis of gradient penalty per epoch offers insights into the critic's performance throughout the training. Higher penalty values suggest more

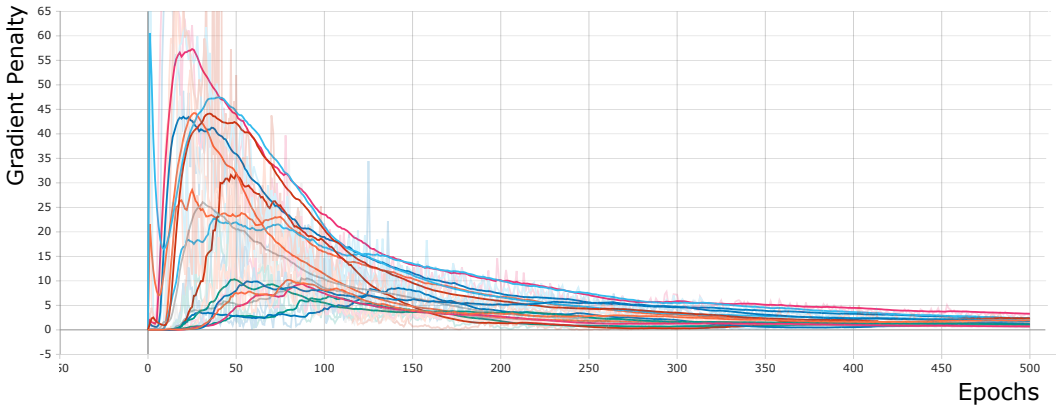


FIGURE A.3: Plot comparing example model gradient penalty per epoch.

significant deviations from the Lipschitz constraint, indicating potential issues in model training or architecture.

In the model comparison of gradient penalties as illustrated in Figure A.3, several key observations can be made. Firstly, the gradient penalty trends largely mimic those of the critic loss. This parallel indicates a close correlation between the critic's effectiveness in enforcing the Lipschitz constraint and the overall model performance.

Among the various configurations, the pink line, corresponding to the G_3C_4 configurations, exhibits the largest initial peaking in gradient penalties. This suggests an initial struggle to adhere to the Lipschitz constraint, potentially due to aggressive critic training or suboptimal parameter settings. Following closely are the light blue (G_1C_4), red (G_1C_3), and orange (G_2C_4) configurations, each showing significant but slightly lesser initial peaks compared to G_3C_4.

A common trend observed across most models is the convergence of gradient penalties around the value of 1, with a range typically between 0 and 3. This convergence signifies the models' increasing adherence to the Lipschitz constraint as training progresses, reflecting an improvement in the critic's performance.

Notably, as models improve and stabilise, there is a discernible decrease in the amount of gradient penalty. This decrease is indicative of the models'

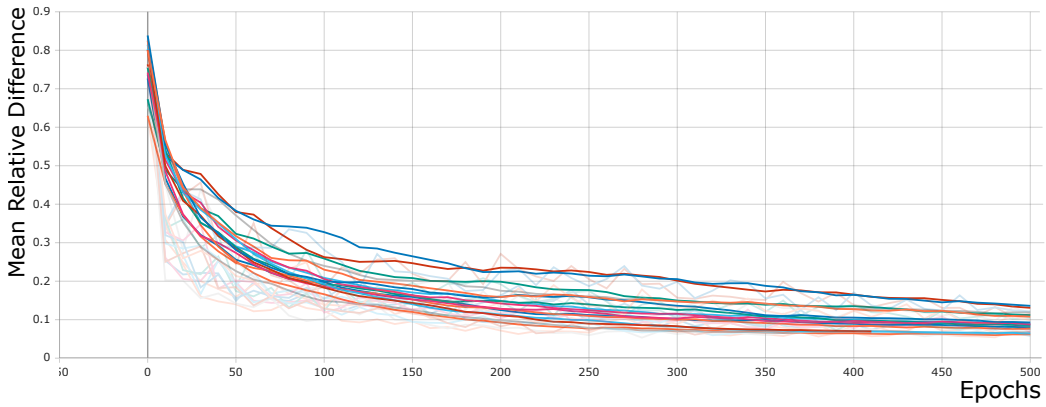


FIGURE A.4: Plot comparing example model mean relative difference per epoch.

enhanced ability to maintain the Lipschitz constraint naturally, reducing the need for penalization. This trend underscores the effectiveness of the gradient penalty mechanism in guiding the models towards more stable and reliable training outcomes in the context of WGANs.

Using the evaluation metrics detailed in Section A.1.4, the **mean relative difference** and **Kolmogorov-Smirnov score** are used to assess the quality of the generated feature distribution during the training of WGANs.

An analysis of the mean relative difference per epoch, as shown in Figure A.4, clearly shows that the quality of generated data improves as the number of epochs increase. This improvement suggests that the models are progressively learning to better replicate the real data distribution. Among the various configurations, the orange (G_2C_4), grey (G_2C_3), and light blue (G_1C_4) models demonstrate the best performance. These models show a consistent reduction in relative difference over epochs, and the best convergence ability, underscoring their effectiveness in closely mimicking the real data distribution.

Similarly, the Kolmogorov-Smirnov score, as depicted in Figure A.5, affirms that the quality of generated data consistently improves across epochs. The KS score decreases over time, indicating an improving match between the generated and real distributions. The orange (G_2C_4), grey (G_2C_3), and light

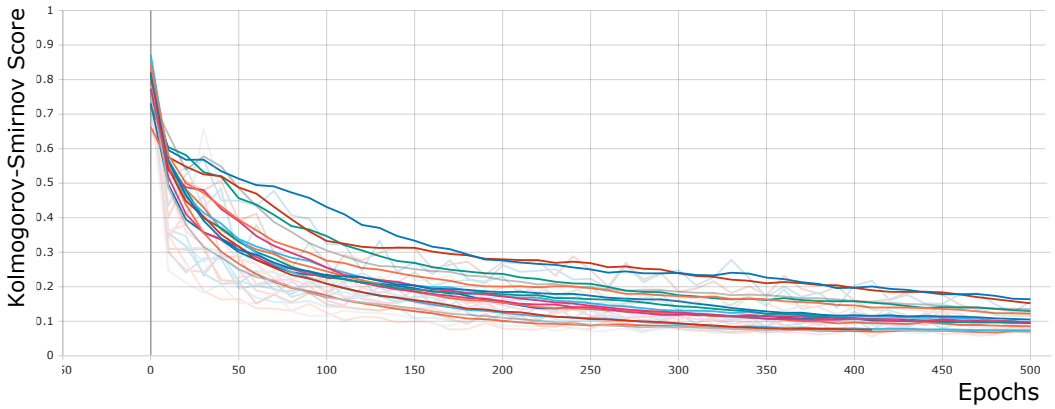


FIGURE A.5: Plot comparing example model mean Kolmogorov-Smirnov score per epoch.

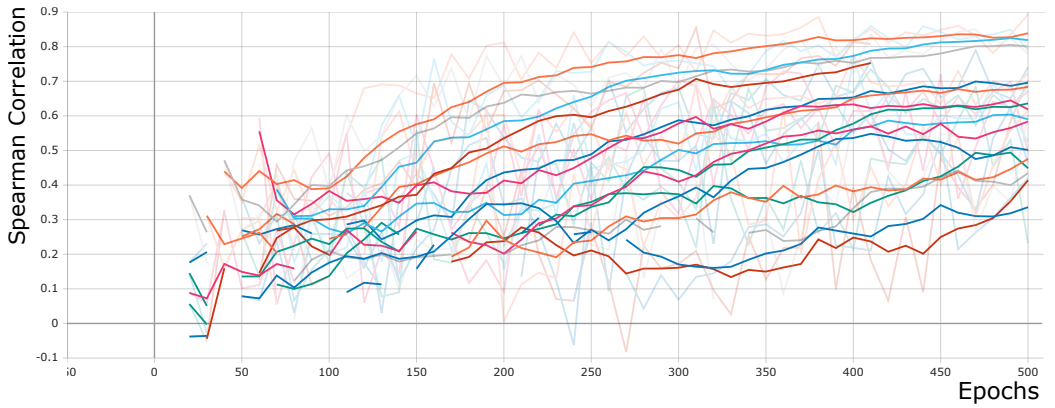


FIGURE A.6: Plot comparing example model mean Spearman Correlation score per epoch.

blue (G_1C_4) configurations again emerge as the best performers. The declining trend in their KS scores across epochs suggest that these models are increasingly successful in generating data that closely aligns with the real distribution, validating the effectiveness of their generative capabilities.

Finally, the **Spearman Correlation Test**, is used to evaluate the generated feature correlations. The Spearman score and p-value provide insights into how well the generated data replicates the feature-wise correlations present in the real data.

The mean Spearman Correlation score per epoch, as presented in Figure A.6, demonstrates a clear trend of improvement in feature-wise correlation of

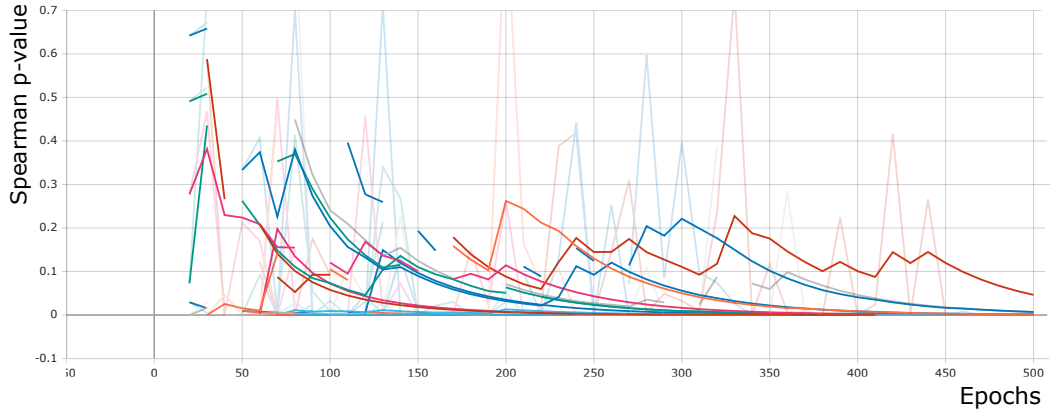


FIGURE A.7: Plot comparing example model mean Spearman Correlation p-value per epoch.

the generated data over the training epochs. This enhancement reflects the increasing ability of the models to capture and replicate the intricate relationships between features in the real data. Among the models, the orange (G_2C_4), grey (G_2C_3), and light blue (G_1C_4) configurations stand out for their superior performance. These models exhibit a consistent improvement in their Spearman scores, indicating their effective replication of real data correlations.

The mean Spearman Correlation p-values, illustrated in Figure A.7, typically trend towards zero as models approach convergence. This is an expected behaviour, confirming the statistical significance of the correlation scores. However, notable exceptions are observed in the dark red (G_3C_2) and dark blue (G_3C_1) models. These models exhibit higher p-values, suggesting difficulties in achieving convergence due to possibly insufficient complexity in their critic models. The persistently higher p-values in these configurations signal a failure to adequately capture the real data's feature-wise correlations, reflecting a critical area for model improvement or redesign.

Early Stopping

In the study of optimizing WGAN data generators, the above example relied on a fixed number of epochs for model training. However, as the complexity of the models increased, it became evident that a larger number of

epochs was necessary to achieve convergence. This variability in epoch requirements across different model complexities posed a challenge in conducting fair comparisons. To address this issue, the strategy of early stopping was implemented. Early stopping serves as a mechanism to terminate the training process of a model, not arbitrarily after a set number of epochs, but once it meets a predefined convergence criterion.

The general criterion for convergence in this context was defined based on the performance of the generator and critic. Specifically, training would cease when there was no further improvement in the loss of both the generator and critic networks over a designated number of epochs. This method allowed for a more adaptive approach to training, ensuring that each model was allowed sufficient time to converge without the risk of overfitting due to excessive training.

Building upon this concept, a more nuanced early stopping mechanism was developed, particularly for complex physics-based WGAN models. This advanced form of early stopping went beyond simply monitoring the loss values. It incorporated the introduced evaluation metrics, offering a more comprehensive and dynamic approach to determining the optimal point for halting training. This innovation in the early stopping methodology not only enhanced the efficiency of the training process but also ensured a higher degree of accuracy and reliability in the model's output, tailored to the intricate needs of complex physics simulations.

A.2.2 Summary of important model optimisations

The model development and optimisation exposed how models can be evaluated and compared. In this section, a summary of the model optimisation considerations and experimental resulting insight is provided. These can be broken down into the following categories: data distributions and transformations, activation functions, and hyperparameters. This section will explore each of these categories.

Data distributions and transforms

The first consideration when training a machine learning based data generator, is the feature distributions of the training data. Data generators including the WGAN have difficulty generating uniformed distributions, such as the ϕ distributions. They similarly have difficulty with uniformed exponential, of the important $m_{\ell\ell\gamma}$ distribution, and the rounded η distributions. In order to provide more optimal training distributions, transforms are used. In this study, the Scikit framework was used to perform data transforms [144]. This study explored various data transformation techniques to optimise the training distributions for the data generator. The transforms considered include MinMaxScaler, QuantileTransform, PowerTransform, and LogTransform. These transformations were evaluated not only on individual features but also on the entire dataset, aiming to identify the most effective method for enhancing the generator's performance.

When applying each data transform to the training data, it was crucial to store the transform scalars. This ensured that the inverse transformation could be consistently applied to the output generated data, maintaining uniformity and comparability between the training and generated datasets.

Through this evaluation process, each transform demonstrated varying degrees of potential in improving the data distributions. However, it was observed that the LogTransform and PowerTransform consistently led to inferior results compared to the other methods. The QuantileTransform, while slightly enhancing the outcomes, did not significantly impact the quality of the generated data. On the other hand, the MinMax Scaler transform emerged as the most effective, consistently yielding better results. Consequently, the MinMaxScaler was selected as the preferred data transformation method for the entire training dataset. This choice was based on its demonstrated ability to consistently improve the quality of the data distributions, thereby facilitating more effective training of the WGAN model.

Additionally, to improve the quality of the output distributions, with focus on the invariant mass distribution, it was found that by increasing the range provides a buffer region on the low and high mass values. This extended

region is removed post generation, providing smoother and more accurate distributions at the minimum and maximum masses.

Activation functions

In the process of optimizing the WGAN data generator, a detailed examination of activation functions for both the critic and generator networks was conducted. These networks were treated independently, allowing for a more focused and effective optimization strategy. Various activation functions were considered, both individually and in combination. This analysis included ReLU, Sigmoid, Tanh, LeakyReLU, ELU, and Softmax. For the critic network, the most effective configuration was found to be the exclusive use of ReLU between layers. Although combinations of Sigmoid and Tanh with ReLU showed promise, they did not match the performance of ReLU alone.

For the generator network, a similar trend was observed, with ReLU emerging as the most effective activation function. Notably, the implementation of batch normalization (BatchNorm) layers significantly improved the quality of generated data. This improvement was also accompanied by a reduction in the number of epochs required for convergence, illustrating the synergistic effect of BatchNorm with the network's batch size hyperparameter. Additionally, the use of the Sigmoid function as the output node of the generator network, in combination with min-max normalization, was considered. This approach aimed to constrain the output data within a 0 to 1 range, adhering to the min-max normalization principles.

The analysis revealed a predominant advantage in models using ReLU for both the generator and critic networks. The combination of BatchNorm with ReLU particularly stood out in enhancing the generator's performance. However, the inclusion of Sigmoid functions, even at the output node of both networks in conjunction with MinMax normalization, did not yield any substantial improvement in results. This comprehensive study underscores the critical role of tailored activation function selection in optimizing neural network architectures, particularly in the context of generative adversarial networks.

Hyperparameters

Following the detailed exploration of activation functions in the WGAN data generator optimization, attention was shifted towards the fine-tuning of various hyperparameters. Key hyperparameters under scrutiny included the latent dimension, batch size, learning rate, λ (gradient penalty constant), β_1, and β_2. These parameters play a crucial role in the performance and efficacy of the generative models, making their optimization vital for achieving the best possible results.

The selection of optimal hyperparameters was carried out for each optimization iteration and experiment. One of the significant findings was regarding the latent dimension, where a value around 8 was identified as optimal. This dimension size proved to be particularly effective in balancing the model's complexity and its ability to generate high-quality data. Additionally, the batch size, especially when used in conjunction with batch normalization in the generator network, showed optimal performance in the range of 258 to 768 events. This range was found to strike the right balance between computational efficiency and the model's capacity to learn from the data.

The learning rate, another pivotal hyperparameter, was dynamically adjusted over the epochs. Initially set around $1 \cdot 10^{-3}$ to allow for larger weight updates in the early stages of training, it was gradually decreased to as low as $1 \cdot 10^{-6}$ as the model neared convergence. This strategy facilitated fine-tuning of the model, enabling more precise updates and thereby enhancing the quality of convergence. For the parameters λ, β_1, and β_2, fixed values were found to be most effective, as $1 \cdot 10^{-3}$, 0, and 0.9 respectively. These fixed values provided a stable framework within which the model could reliably learn and adapt to the data, further contributing to the overall optimization of the WGAN data generator.

A.2.3 Final optimised WGAN model and results

The culmination of the optimization process for the WGAN data generator is presented in this final section, where we showcase the best performing model after extensive experimentation and refinement. This final model

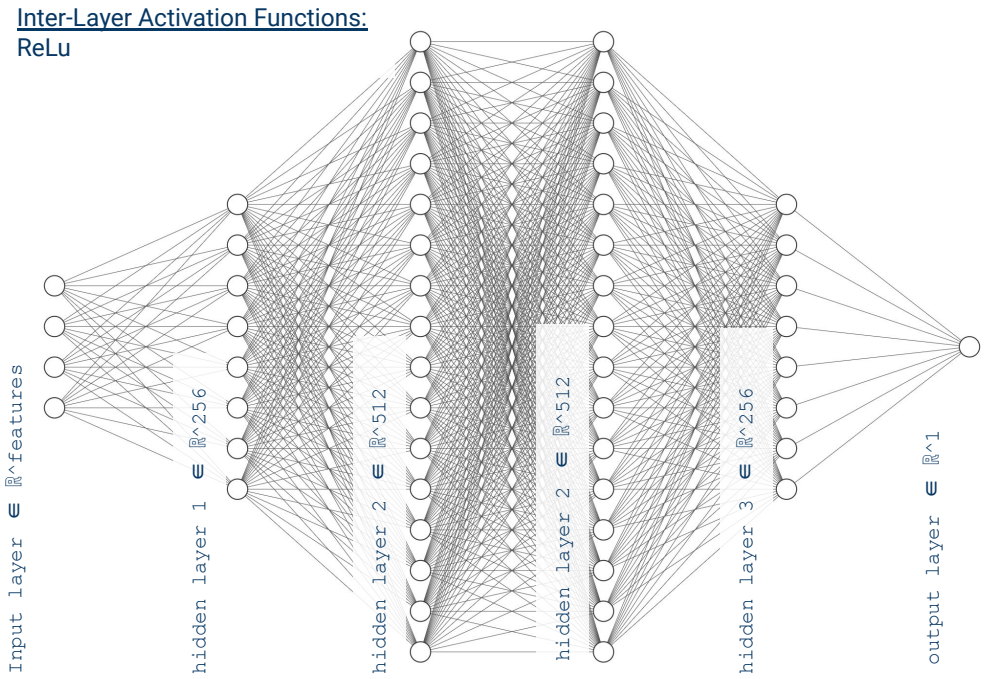


FIGURE A.8: Diagram of final optimised critic network architecture.

incorporates a series of preprocessing steps on the training data, including the application of the MinMax transform to feature distributions and an extension of the invariant mass range.

For the critic model, the architecture of the final optimised version is detailed in Figure A.8, providing a comprehensive visual representation.

The generator models architecture, which has been meticulously refined through the optimization process, is presented in Figure A.9.

The final selection of hyperparameters for this optimised model includes a latent dimension of 8, a batch size of 512, and a learning rate that dynamically adjusts from $1e^{-3}$ to $1e^{-6}$ in accordance with the early stopping criterion. The gradient penalty weight (λ) is fixed at $1 \cdot 10^{-3}$, and the values for β_1 and β_2 are fixed at 0 and 0.9, respectively. It's noteworthy that the critic model undergoes training five times for each generator training iteration, ensuring robustness and efficacy in the learning process.

The model's training can be exposed through the evaluation metrics per

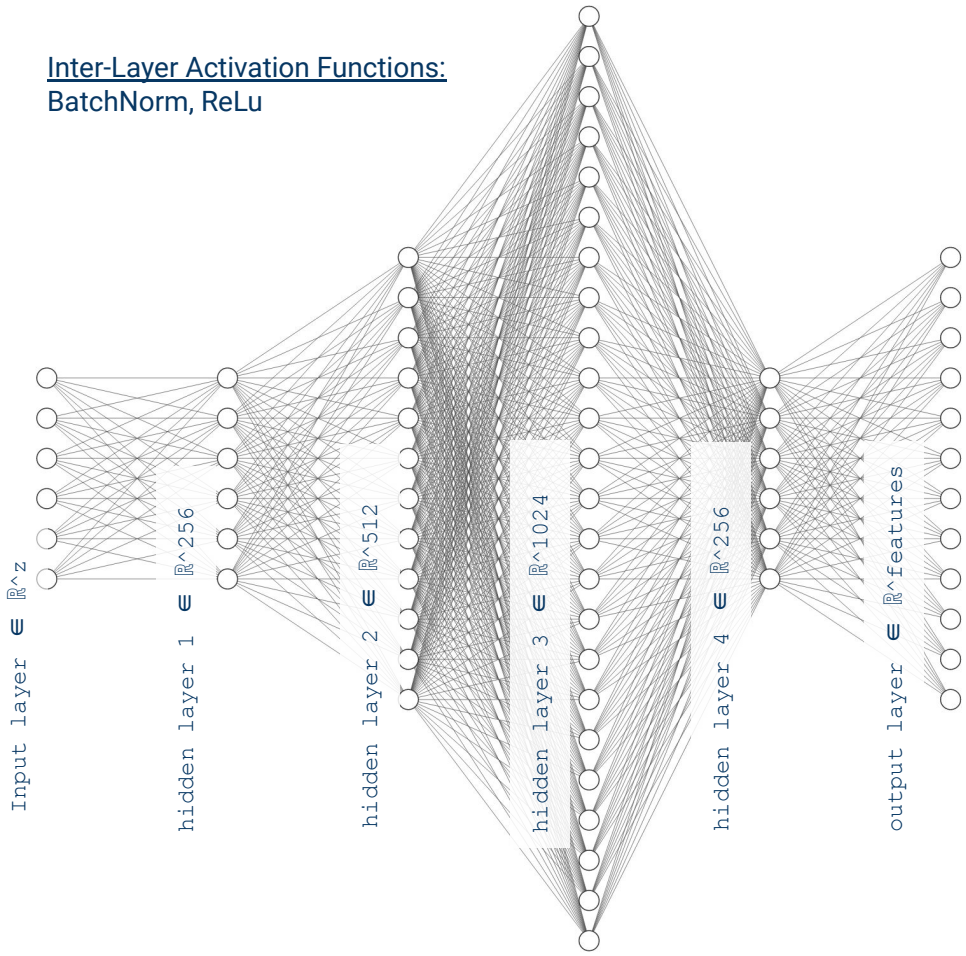


FIGURE A.9: Diagram of final optimised generator network architecture.

epoch. Figure A.10, exposes the model improvement per epoch for the feature distributions, in Figures A.10a and A.10b, and the feature correlations, in Figures A.10c and A.10d.

The results of this optimised model are visually represented through feature distribution plots and feature correlation plots, comparing the generated data with the training data. These results are encapsulated in Figures A.11 and A.12.

The feature distributions, shown in Figure A.11, exposes how generated data distributions mimic those of the MC data. However, the limitations

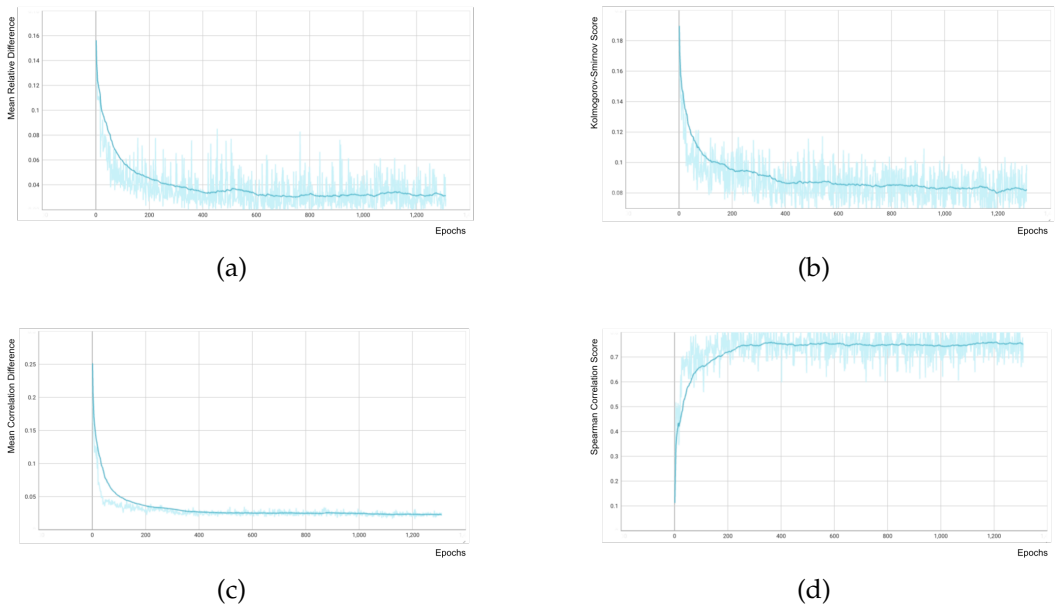


FIGURE A.10: Evaluation metrics per epoch for best performing WGAN model with (A) Relative difference; (B) Kolmogorov-Smirnov score; (C) Correlation difference; and (D) Spearman correlation score.

of the generated data is shown in the form of additional peaks and asymmetries in ϕ distributions, as well as minor fluctuations in the η distributions. It is important to note that the $m_{\ell\ell\gamma}$ generated data follows the MC distribution extremely well.

The feature correlations in Figure A.12, firstly shows that the maximum difference between the MC and generated feature correlation is approximately 0.06. This demonstrates the model's success in providing realistic correlation in event features. The features correlations with the most substantial differences are the correlations between the $\Delta R_{\ell\ell}$ and η_{ℓ_2}; the E_T^{miss} and η_γ; and the categorical N_j and N_{cj}.

These results underscore the WGAN's capability to realistically produce $Z\gamma$ physics events, demonstrating the success of the optimization process in enhancing the model's accuracy and reliability. This achievement marks a significant step forward in the field of generative modelling for complex physical phenomena.

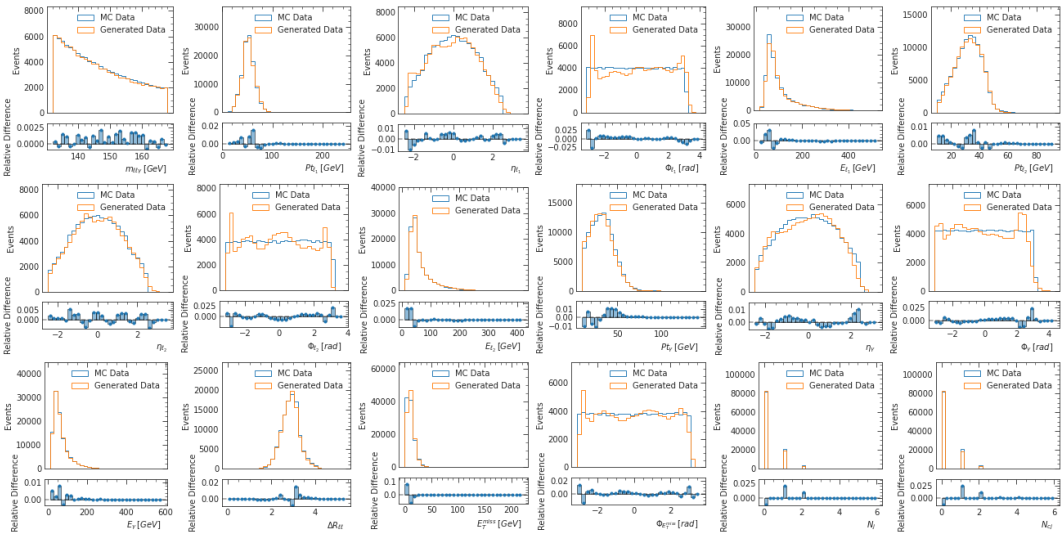


FIGURE A.11: Final optimised WGAN generated data feature distributions.

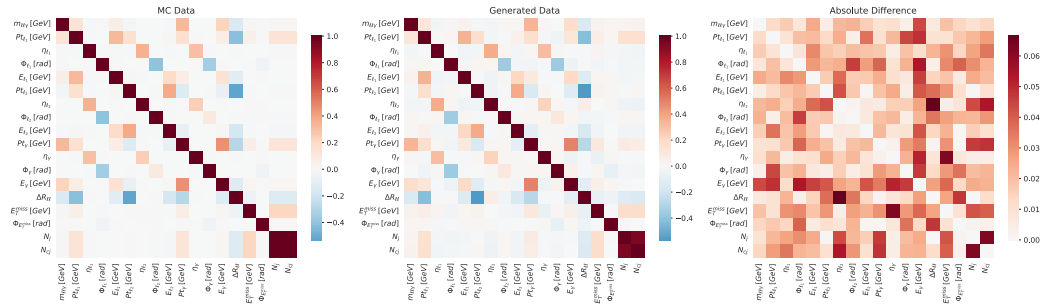


FIGURE A.12: Final optimised WGAN generated data feature correlations.

Appendix B

Contributions to Modelling the COVID-19 Pandemic

B.1 Introduction

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, marked a significant turning point in global health, societal norms, and economic structures. The virus, initially identified in Wuhan, China, during December 2019 [145], quickly spread globally, leading the World Health Organisation (WHO) to declare it a Public Health Emergency of International Concern in January 2020, and subsequently a pandemic on March 11, 2020 [146].

The SARS-CoV-2 virus, a novel coronavirus, belongs to the same family as the SARS and MERS viruses, which were responsible for severe respiratory syndrome outbreaks in 2002-2003 and 2012 respectively [147].

COVID-19, an abbreviation for "coronavirus disease 2019," predominantly presents with respiratory symptoms, ranging from mild cold-like symptoms to severe pneumonia. The virus can also affect multiple organ systems and has a diverse range of clinical manifestations, making it a complex disease to manage [148]. Additionally, asymptomatic and presymptomatic transmission played a significant role in the spread of the disease, adding to the challenge of containment.

COVID-19 spread rapidly, reaching nearly all corners of the world. By early 2020, many countries started reporting their first cases, and by the end of the year, the pandemic had led to millions of infections and hundreds of

thousands of deaths globally. In addition to the direct health impact, the pandemic has led to extensive societal and economic disruption. The speed of its spread led to unprecedented pressure on healthcare systems worldwide, necessitating extraordinary measures like city-wide lockdowns and widespread travel restrictions.

In response to the pandemic, a global initiative was undertaken to model its progression, aiming to enhance the decision-making process for public health policies [149]. Numerous research teams and national response units delved into the investigation of intervention strategies tailored to specific countries, exploring their influence on the virus's transmission rate. Additionally, these studies have assessed how the unique characteristics of each country prior to the pandemic affect the spread of the virus [150, 151]

Simultaneously, a global race to develop effective vaccines began. Through an unprecedented collaboration and investment, the first vaccines received emergency use authorisation by the end of 2020. Vaccine distribution, however, has raised further challenges, with issues such as equitable access and vaccine hesitancy becoming prominent global concerns.

As of mid-2023, the COVID-19 pandemic continues to evolve. Variants of the virus, differing in their transmissibility and disease severity, are causing renewed waves of infection in many parts of the world. Global cooperation, scientific research, public health measures, and vaccination remain crucial in managing the ongoing crisis.

B.1.1 The South African COVID-19 pandemic

On the 5^{th} of March 2020, South Africa registered its first COVID-19 case, followed three weeks later by a comprehensive government-mandated lockdown, on the 27^{th} of March [152]. This lockdown was a critical part of a broader, multi-level strategy designed to correspond with varying risk levels, as detailed in Ref. [153]. A detailed timeline of the nation's adjustments in lockdown levels is presented in Table B.1. The initial wave of the South African pandemic persisted until October 2020, after which the number of cases became more manageable. However, by late November 2020,

Alert Level	Wave	Start Date	Total Cases	Recoveries	Fatalities
5	1	27 March 2020	927	12	0
4	1	1 May 2020	5951	2382	116
3	1	1 June 2020	34357	17291	705
2	1	18 August 2020	592144	485468	12264
1	2	21 September 2020	661936	591208	15992
3	2	29 December 2020	1021451	858456	27568
1	2	1 March 2021	1513959	1431336	50077
2	3	31 May 2021	1665617	1559337	56506
3	3	16 June 2021	1774312	1620317	58223

TABLE B.1: South Africa's COVID-19 pandemic dynamics, including alert level progression for the first, second and third wave.

the country experienced a resurgence in cases, marking the onset of the pandemic's second wave. The implemented risk-adjusted strategy facilitated a dynamic approach to reopening and closing the economy. This strategy was influenced by several factors, including the rate of virus transmission, the number of active cases, the healthcare system's capacity, the effectiveness of public health measures, and the broader economic and societal impacts of ongoing restrictions.

The University of Witwatersrand, in collaboration with iThemba LABS, has played an important role in the Gauteng Premier's COVID-19 Advisory Committee, contributing detailed analyses of the pandemic's progression in the province [154]. The involvement in this committee primarily centres around modelling efforts that equip government officials with crucial data. This data serves as a statistical foundation for decision-making processes, guiding the adjustment of alert levels and the allocation of resources. Our models offer a robust statistical basis, facilitating informed and effective responses to the evolving pandemic situation.

In this section an overview of the data driven methodologies that were used to understand and inform policy decisions in Gauteng. Firstly the epidemiological modelling of COVID-19 will be introduced and the most important extensions and implementations presented. Secondly the methodology and

implementation of a geo-spacial COVID-19 hot-spot model is presented.

B.2 Epidemiological Modelling

Epidemiological modelling is a critical tool for informing public health, employing mathematical and statistical techniques to understand, predict, and control the spread of infectious diseases. These models simulate disease dynamics, taking into account various factors like transmission rates, recovery rates, and population susceptibility. By offering insights into potential epidemic trajectories and the effectiveness of intervention strategies, epidemiological models are indispensable for informed decision-making in public health policy and response planning.

B.2.1 Epidemiological data

In epidemiological modelling, as with any data-driven research, the accuracy and detail of the outcomes heavily depend on the quality and granularity of the available data. The foundational data for epidemiological models typically include counts of susceptible, infected, recovered individuals, and fatalities related to the virus, as depicted in Figure B.1. This dataset is essential for constructing SIRD (Susceptible, Infected, Recovered, Deceased) models. To enrich these models and provide deeper insights, additional data sources are often integrated. These include detailed hospital data (covering aspects like occupancy rates, admissions, ward types, fatalities, and comorbidities), statistics on excess fatalities (which may not be directly attributed to the virus), and geolocation data. Such comprehensive data integration enables a more nuanced understanding and modelling of the pandemic's spread and impact.

B.2.2 Compartmental di-SIRD COVID-19 model

In an effort to understand and mitigate the impact of the COVID-19 pandemic, and to provide valuable insights for policymakers, a compartmental di-SIRD (Dual Infectious-Susceptible-Infected-Recovered-Deceased) model was developed. A detailed introduction and explanation of this model is

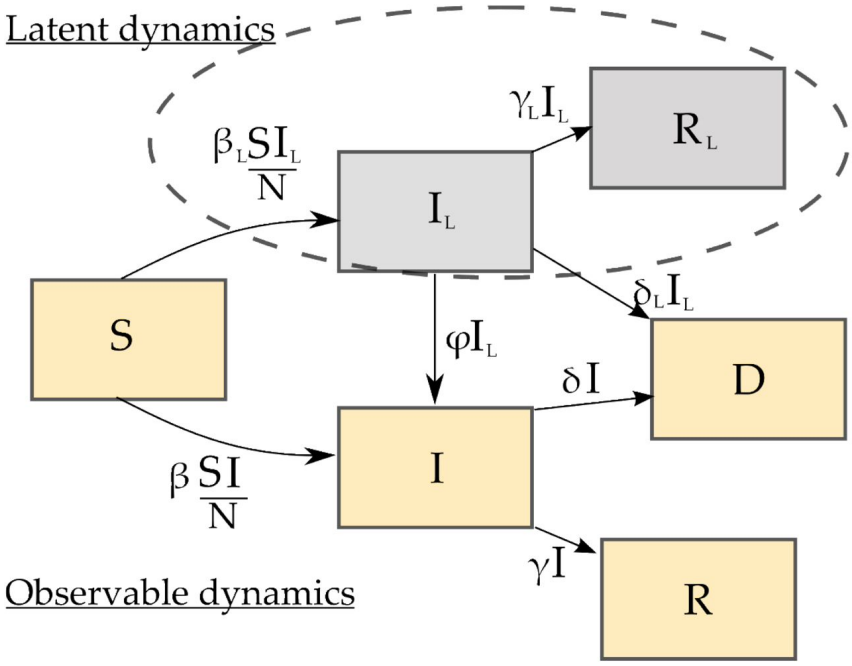


FIGURE B.1: Diagram of di-SIRD model with unobserved latent dynamics [155].

provided in Ref. [155]. This model is specifically designed to address the unique aspects of the COVID-19 pandemic, particularly the phenomenon of asymptomatic yet infectious individuals [156]. The di-SIRD model incorporates both observable and latent variables to accurately represent the dynamics of the current outbreak, thereby facilitating effective control measures and informed decision-making. The latent and observable contributions and interactions in the di-SIRD model is shown in Figure B.1

To effectively implement an epidemiological model, it is crucial to understand its underlying mathematical equations. A fundamental equation representing the dynamics of the entire pandemic posits that the total population (N) is equal to the sum of the numbers of susceptible (S), infected (I), latent infected (I_L), latent recovered (R_L), recovered (R), and deceased (D) individuals:

$$S + I + I_L + R_L + R + D = N. \tag{B.1}$$

The temporal evolution of these compartments is captured through first-order derivatives, as expressed in Equation B.1, yielding the following relationship:

$$\frac{\partial S}{\partial t} + \frac{\partial I}{\partial t} + \frac{\partial R}{\partial t} + \frac{\partial D}{\partial t} = 0. \tag{B.2}$$

To predict future trends using this model, we employ derivative equations, taking into account the interrelations depicted in Figure B.1 and the overarching equation in B.1. The derivative equations are formulated as follows:

$$\frac{\partial S}{\partial t} = -\beta_f \cdot \frac{S \cdot I}{N}, \tag{B.3}$$

$$\frac{\partial I}{\partial t} = \beta_f \cdot \frac{S \cdot I}{N} - \gamma \cdot I, \tag{B.4}$$

$$\frac{\partial I}{\partial t} = \phi \cdot I - \delta \cdot I - \gamma \cdot I, \tag{B.5}$$

$$\frac{\partial R}{\partial t} = \gamma \cdot I, \tag{B.6}$$

$$\frac{\partial D}{\partial t} = \delta \cdot I, \tag{B.7}$$

where γ is the recovery rate, ϕ is the infection rate, δ is the fatality rate and β_f is the kernel function incorporating intervention strategies. These equations collectively form the mathematical backbone of our epidemiological model, enabling predictions and analyses of pandemic trends.

The kernel function, β is conditioned on the control measure p. This control measure takes into account that the implementation of nationwide control measures against COVID-19 is influenced by random variations in both the timing of execution and the level of public compliance. These factors depend on the specific execution plan and the general attitude of the population. To model this, the kernel function can be written as:

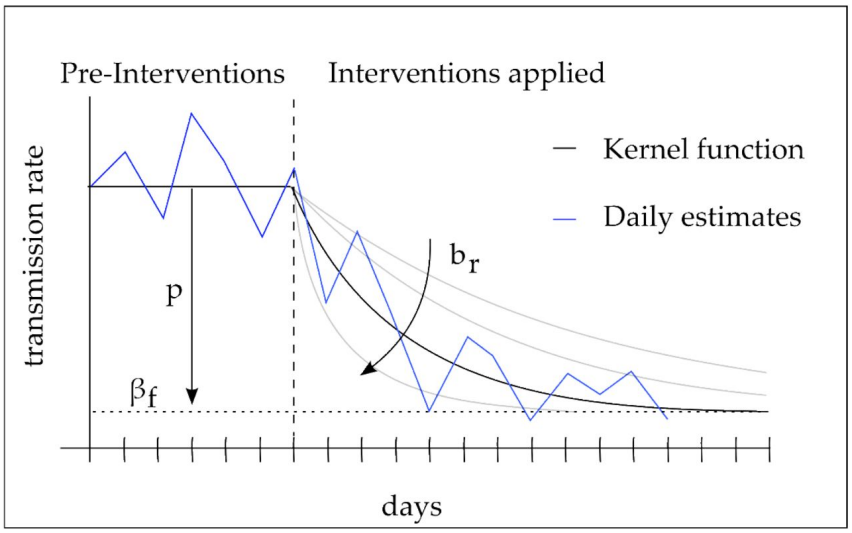


FIGURE B.2: The transmission rate kernel function depicting the typical adjustment rate, β_r and final transmission rate β_f. The control index p affects the total reduction in transmission.

$$\beta_f(t) = \beta_0 + (1 - \exp^{-b_r \cdot t})\beta_0 \cdot \alpha_s \cdot p(t), \tag{B.8}$$

where β_0 models the initial transmission rate before the application of the control measure and b_r is the typical rate of adjustment.

Here, p represents the Stringency Index of the non-pharmaceutical interventions (NPIs), adapted from the Oxford COVID-19 Government Response Tracker (OxCGRT), specifically for South Africa's five-level alert system. This index quantifies the extent of social distancing enforced by NPIs. The parameter α_s denotes the efficiency of the NPIs, reflecting the degree of adherence to social distancing measures. A higher index value indicates greater adherence, with this parameter being derived from empirical data.

B.2.3 Alert levels and inclusion in model

The stringency index, $p(t)$, for each alert level was calculated by scoring the indicators appropriately based on the description of the Alert Levels provided by the government. The development and integration of the stringency index is exposed in detail in Ref. [157]. The alert levels therefore show the extent of government implemented NPI. The stringency index is directly

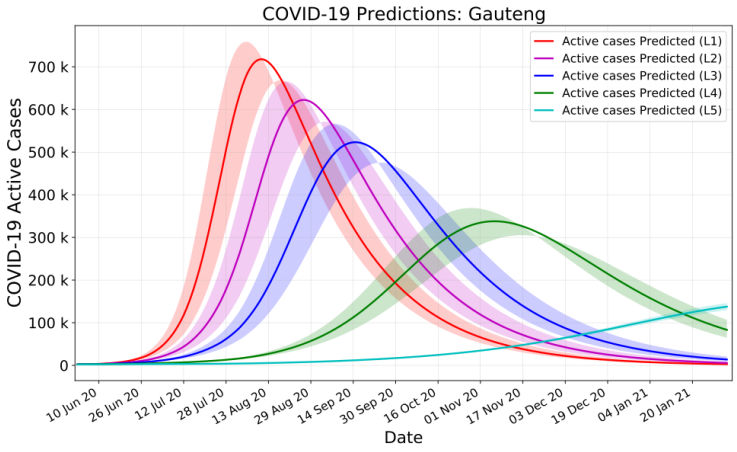


FIGURE B.3: Influence of South Africa's alert levels on SIRD model predictions.

related to the alert levels, and therefore influences the model's prediction, at a given time, via the β parameter, as shown in Figure B.3. By decreasing the stringency of government implemented intervention (low alert levels) leads to an initial high rate of infection and active cases that decreases slowly. However, as the stringency increases (higher alert levels) the rate of infection slows, leading to a decreasing number of active cases at the wave peak.

The implemented alert level must therefore be carefully selected by the stakeholders to allow for sufficient hospital resources to cope with the pandemic wave, while limiting the negative effects on the socio-economic dynamics caused by the interventions.

The stringency index is dependent on time and must be adjusted accordingly to implement interventions as affectively as possible. Figure B.4 exposes the implemented stringency index for Gauteng Province during the 2021 period.

B.2.4 Calibration of predictive model

Before implementing the epidemiological model, the model must be calibrated on the specific virus dynamics. At a basic level this includes the

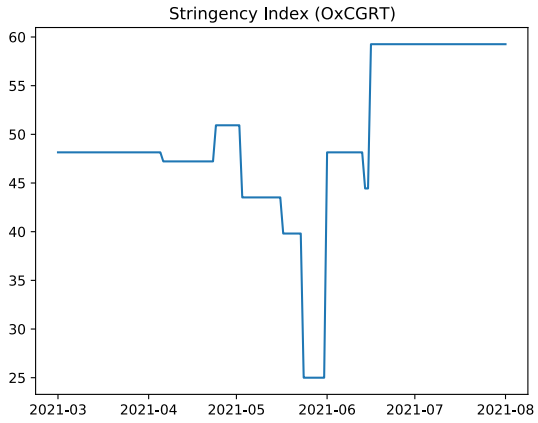


FIGURE B.4: Gauteng Province stringency index for March to August 2021.

calibration through fitting of the fatality rate, λ, the recovery rate, γ, and the infection rate, β, with rate of social adherence, α.

The parameters are calibrated on the most recent and informative data available to provide parameters that accurately represent the local pandemic. The calibration of the infection rate includes the stringency index and simultaneously calibrates the level of adherence using Equation B.8. The $\alpha_s \cdot p(t)$ term is expanded to $\alpha_i \cdot p_i(t) - \alpha_s \cdot p_s(t)$ where subscript i represents the previous stringency and adherence, and subscript s represents the calibration parameters of interest. Figure B.5, exposes the calibration of the infection rate for the month preceding the 21^{st} of May 2021.

B.2.5 Predicting the spread of COVID-19

Once the model parameters are calibrated to the best available data, the model can be used to predict the dynamics of the current/next wave of the pandemic. The prediction exposes the number of cases, recoveries, fatalities and susceptible population over time. This exposes data driven insight into both the scale and timeframe of a wave. This includes the growth and estimated peak number of cases as well as the date that the wave reaches its peak. The example prediction on Gauteng's third pandemic wave active cases is shown in Figure B.6. The plot exposes the predicted active cases

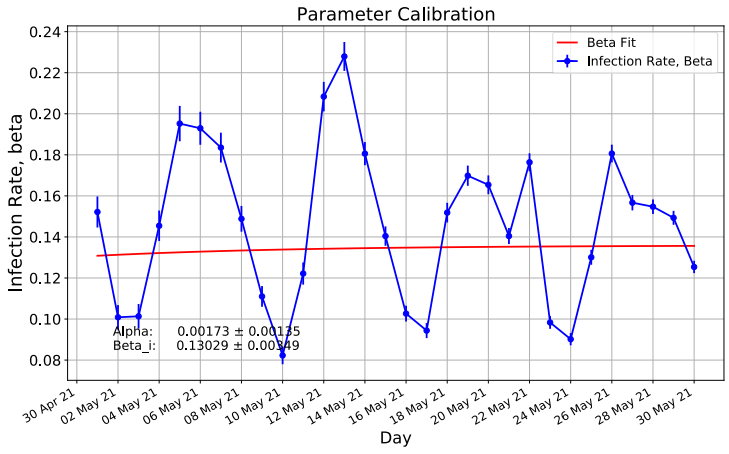


FIGURE B.5: Example infection rate calibration for May 2021 in Gauteng province.

over time together with the actual number of active cases, allowing validation of the prediction as the data follows the prediction.

Furthermore, Figure B.7 exposes further predicted dynamics, including the total number of cases, the total number of recoveries, the number of active cases, and the total number of fatalities. The figure demonstrates the predictive model's ability to model the entire virus dynamic, through the extent that the data is following the prediction.

B.2.6 Extending models for hospital occupancy

One of the most important extensions to the model is to expose the predicted number of hospital occupancies. This is achieved through combining an in-depth analysis of the hospital data. Hospital data includes hospitalisation dates, hospital ward information, patient comorbidities, patient age, patient recovery/fatality, and private/public hospitalisations. Therefore, through fitting and projecting specific hospital dynamics, as shown in example Figure B.8, they can be combined with the SIRD projection to predict specific progressions.

These predictions include the predicted number of occupancies in the general ward, high care, and intensive care unit respectively. It can be further

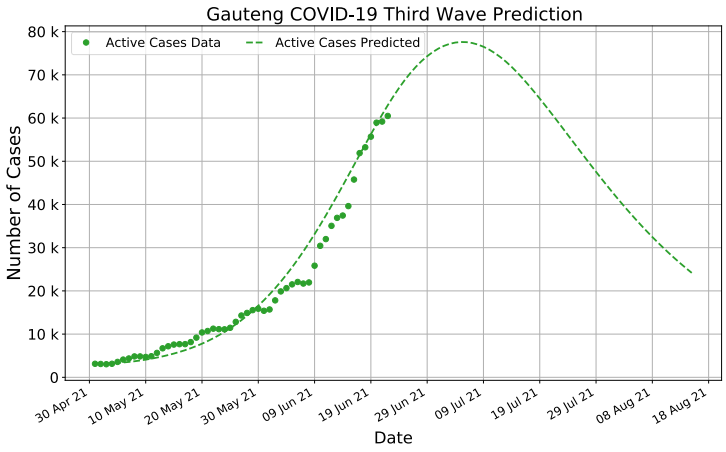


FIGURE B.6: Active cases model prediction vs data for the third wave in Gauteng.

used to expose required alert level adjustments, to be made by policymakers, to limit the number of occupancies at the peak of a pandemic wave to be within the limited number of beds available. Additional hospital analysis are used to expose which comorbidities in combination with COVID-19 cause increased need for hospitalisation. An example of the predicted number of hospital occupancy for the ages of 50 to 59 during the period of the fourth pandemic wave, is presented in Figure B.9.

B.2.7 Extending models for vaccinations

As vaccinations became accessible to the public, the predictive model must absorb the effect of people who are vaccinated. To this end the vaccination data is fit to predict the changing number of vaccinations into the future, as shown in Figure B.10, in order for it to be absorbed into the predictive model.

B.3 Geo-spacial Hot-Spot Modelling

In order to expose the location specific COVID-19 dynamics within a given area, a geolocation based model was developed. The unsupervised model

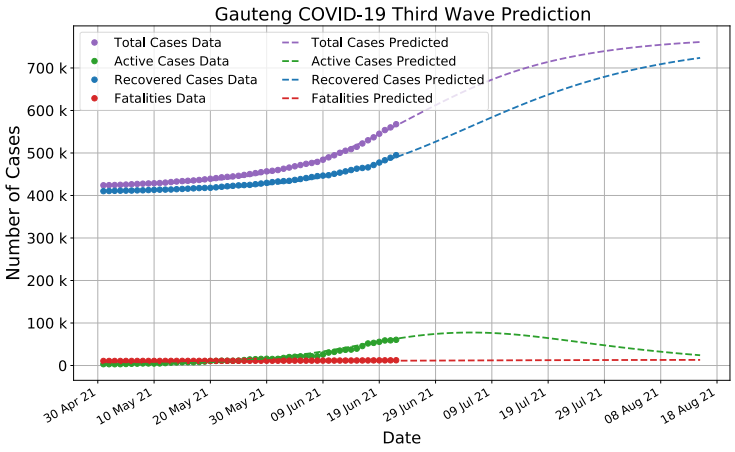


FIGURE B.7: Full model prediction vs data for the third wave in Gauteng.

is implemented in conjunction with an intricate analysis to produce a granular insight into the location specific virus dynamics with emphasis on its severity. The methodology and results are presented in this section and are published in Ref. [158].

B.3.1 Geo-spacial pandemic data

The granularity of data required for precise hot-spot geo-localisation necessitated detailed datasets, such as the anonymised data supplied by the Gauteng Department of Health. This dataset was comprehensive, encompassing variables like Case ID, test date, recorded address, and precise geo-localisation coordinates (latitude and longitude).

The National Institute of Communicable Diseases (NICD) played a pivotal role by aggregating and disseminating daily updates on COVID-19 cases tested across both public and private sectors throughout South Africa. Their reports, providing a detailed provincial breakdown, were integral to this study.

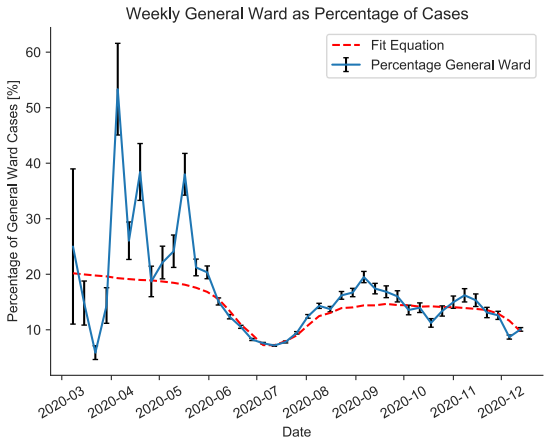


FIGURE B.8: Fit to number of general ward cases as a percentage of total cases.

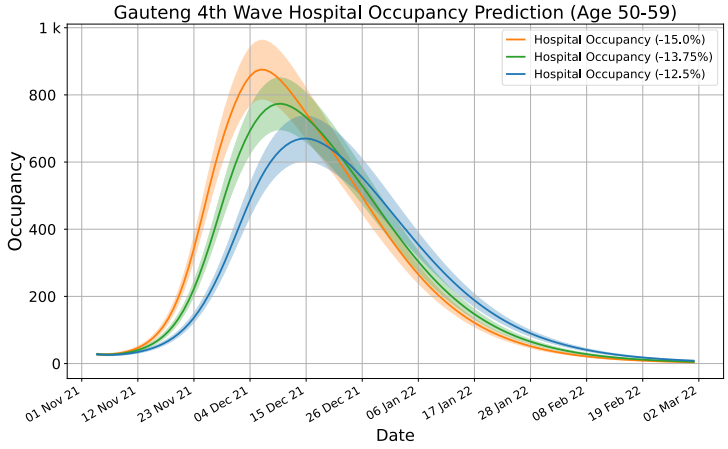


FIGURE B.9: Hospital occupancy prediction for Gauteng's fourth wave.

This data, upon acquisition by the Gauteng Department of Health, underwent a geocoding process to assign accurate geolocations to each case. Ensuring confidentiality, the data was anonymised prior to distribution to external analysts. Prior to the analytical phase, a thorough data cleansing was conducted to eliminate any records with incorrect or unprocessable address data.

Throughout the first wave of the pandemic, from March to October 2020, a

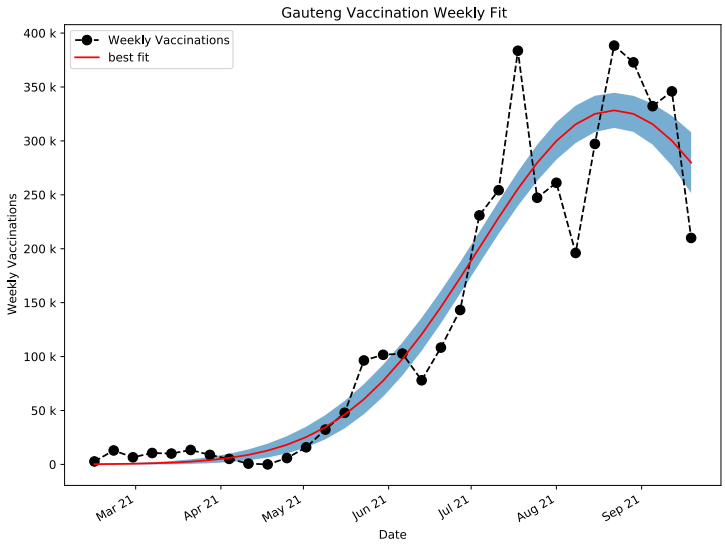


FIGURE B.10: Fit to weekly number of vaccinations.

total of 218,720 geocoded cases were scrutinised. The second wave, extending from November 2020 to February 2021, saw the analysis of an additional 191,750 case samples. It is important to note that post-May 2021, access to such geocoded case data was discontinued.

B.3.2 Clustering cases by geolocation

To effectively dissect the spatial distribution of COVID-19 cases, artificial intelligence techniques are employed with notable efficiency. Our research leverages the capabilities of the unsupervised machine learning method of Gaussian Mixture Models (GMM), which are instrumental in identifying and grouping cases based on geographic coordinates. This methodology facilitates the detailed analysis of case clusters, providing insights into the spatial dynamics of the viral spread within predefined locales.

The practical execution of this clustering involved generating GMM distributions and the subsequent visualisation of these clusters on detailed web-based maps. This process was carried out using Python, in conjunction with the Scikit-learn API package [159], which is renowned for its comprehensive

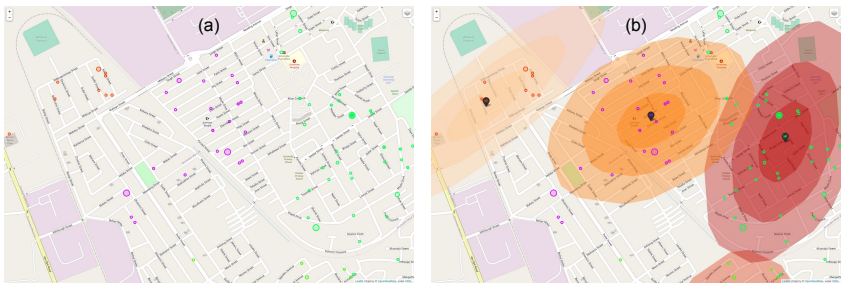


FIGURE B.11: Map visualisation of Gaussian Mixture Model clustering of COVID-19 cases in Gauteng: (a) Input case data. (b) Clustering results.

suite of machine learning tools that simplify data mining and data analysis initiatives.

Gaussian mixture model clustering

In our analysis of COVID-19's geographic impact within Gauteng, we applied Gaussian Mixture Models (GMM) to cluster cases based on residential coordinates. These clusters are subsequently scrutinised to classify areas as either hot-spots or otherwise.

GMMs were preferred over other unsupervised machine learning methods, such as k-means, for their probabilistic foundation. They excel in situations where clusters have the potential to intersect, a realistic aspect of the spatial dynamics of a pandemic. Each cluster is assigned a weight, ϕ, reflecting its relative prominence and aiding in distinguishing genuine hot-spots from spurious ones. The application of these models is graphically represented in Figure B.11, depicting both the input case data and the clustering results.

The GMM algorithm, when operational, estimates a set of k 2-dimensional Gaussian distributions — each defined by its mean, μ_j, covariance, Σ_j, and weighting, ϕ_j. The probability of a new case, $p(x)$, occurring at a given point x is a linear combination of probabilities from all the generated clusters:

$$p(x) = \sum_{j=1}^{k} \phi_j N(x|\mu_j, \Sigma_j), \tag{B.9}$$

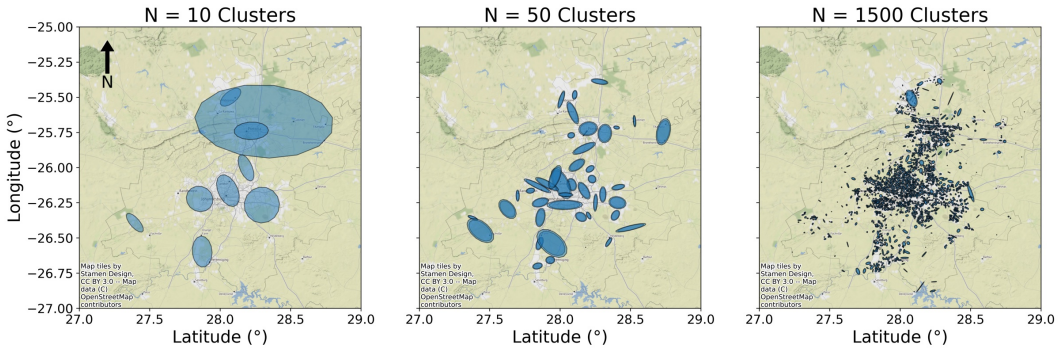


FIGURE B.12: Visual comparison of GMM clustering with different cluster counts: (a) 10 clusters. (b) 50 clusters. (c) 1500 clusters.

where N symbolises the normal distribution. To optimally fit the known case coordinates, the Expectation-Maximisation algorithm is used. This iterative process involves computing the likelihood of each data point belonging to a given cluster (expectation step) and updating the parameters of the clusters accordingly (maximisation step), as shown in Equations B.10 and B.11:

$$\gamma_{ij} = \frac{\phi_j N(x_i|\mu_j, \Sigma_j)}{\sum_{q=1}^{k} \phi_j N(x_i|\mu_q, \Sigma_q)}, \tag{B.10}$$

$$\phi_j = \frac{1}{N}\sum^{N} i\gamma ij, \quad \mu_j = \frac{\sum^{N} i\gamma ijx_i}{\sum^{N} i\gamma ij}, \quad \Sigma^2 j = \frac{\sum^{N} i\gamma_{ij}(x_i - \mu_j)(x_i - \mu_j)^T}{\sum^{N} i\gamma ij}. \tag{B.11}$$

For our specific study of Gauteng, determining the optimal cluster count was paramount. A fine balance was struck with 1500 clusters, enabling precise spatial resolution (1.9km^2 per cluster) and an average of 146 cases per cluster. This configuration, depicted in Figure B.12, was most effective for representing the virus's transmission patterns.

Susceptible-infectious curve

With the latent parameters (weights, means, standard deviations) of the Gaussian probability distributions derived from COVID-19 case data in

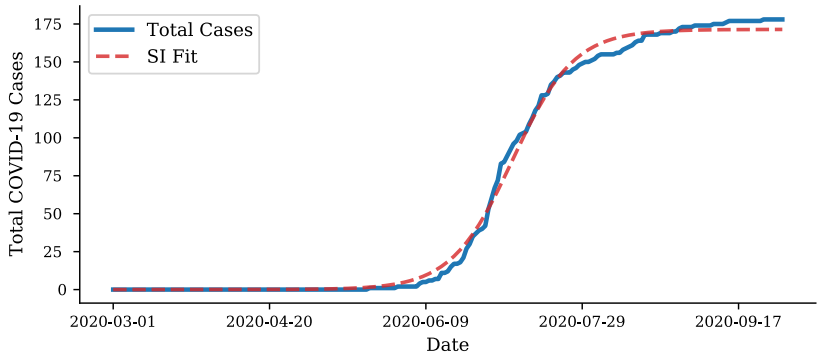


FIGURE B.13: Fit of the SI curve to cumulative COVID-19 case data within a cluster.

Gauteng, the next step involves distinguishing the highly infectious zones, or hot-spots, within the clusters identified. This differentiation is based on an examination of the time series data of COVID-19 cases, wherein the cumulative number of cases is plotted against the onset date of each case within the cluster.

The primary analytical model used here is the Susceptible-Infectious (SI) Curve, which predicts the growth of infections within a susceptible population over time, t. This model is encapsulated by the logistic function:

$$SI(t) = \frac{SI_0}{1 + e^{SI_1(t - SI_2)}}. \tag{B.12}$$

Here, SI_0 denotes the estimated total cases in a cluster at the infection's saturation point, SI_1 is the infection rate, and SI_2 represents the time until the outbreak's peak. The SI model is a solution to the logistic growth differential equation and is well-suited for epidemics; it forecasts a slow initial growth in infections, a subsequent rapid increase, and eventually a plateau as the population of susceptible individuals declines. This behaviour is visualised in Figure B.13, where the SI Curve is fitted to the cumulative case count of a representative cluster.

The fitting of an SI Curve to each cluster's epidemic curve provides localised parameters indicative of the virus's spread dynamics within that cluster. A

poor fit suggests that the cluster may not exhibit the characteristic progression of a hot-spot.

Upon completing the clustering and curve fitting, each cluster's profile includes its Total Cumulative Cases (N_{TC}), areas within one and two standard deviations ($A_{1\sigma}$ and $A_{2\sigma}$), SI model parameters (SI_0, SI_1, SI_2), and the associated ward and municipality. This comprehensive analysis allows for an in-depth understanding of the spatial distribution and progression of COVID-19 in Gauteng, thereby facilitating targeted public health interventions.

B.3.3 First wave clustering and hot-spot definition

To elucidate the spread and intensity of COVID-19 at the cluster level within Gauteng, we have employed the Gaussian Mixture Model (GMM) clustering technique as detailed in the previous section. The data corresponding to the first wave has been meticulously analysed using this model to form distinct clusters. For each cluster, we computed the Susceptible-Infectious (SI) parameters based on Equation B.12, representing the evolution of case numbers over time.

Leveraging these parameters, alongside the cluster attributes derived from the calibration data set, we established specific criteria to discern and classify clusters as hot-spots. This criterion, crafted using the entire data set from the first wave, serves as a baseline for recognising and interpreting subsequent waves. The first wave data comprises 218,720 case entries, which equates to an average of 146 cases for each of the 1,500 clusters identified.

This foundational work sets the stage for the following detailed analyses, which will refine our understanding of the pandemic's impact and guide future public health strategies.

Hot-spot classification on density

The distribution of case densities within the first wave's clusters in Gauteng is illustrated in Figure B.14. This distribution reveals a Gaussian-like pattern

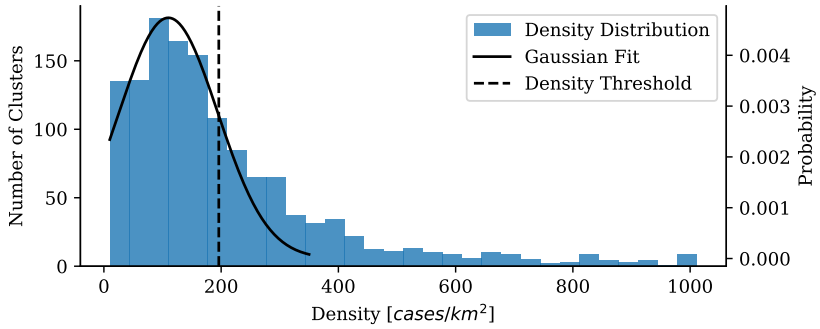


FIGURE B.14: Distribution of cluster densities during Gauteng's first COVID-19 wave.

in the lower density regions ranging from 0 to 350 cases/km^2 and exhibits a long, sporadic tail stretching to densities greater than 30,000 cases/km^2. A significant observation is that clusters with low densities tend to follow the predicted pattern of growth. By setting a cut-off threshold at the one standard deviation, σ, mark, we establish a density threshold, ρ_{th}, at 196.05 cases/km^2. Clusters that have densities surpassing this threshold demonstrate rapid and non-stochastic proliferation of cases, distinguishing them as hot-spots. Consequently, a hot-spot cluster is defined as any cluster whose density is in excess of ρ_{th}.

Hot-spot cluster activity definition

To assess the current state of a hot-spot cluster, it is essential to determine whether it has surpassed it's peak or reached a plateau in its case trajectory. The level of activity within a cluster at any given time can be quantified by the ratio of its total cases at that time, $N_{TC}(t)$, to the cluster's total predicted cases, SI_0. This ratio is defined in Equation B.13:

$$\frac{N_{TC}(t)}{SI_0} < L_{th}. \tag{B.13}$$

In this context, L_{th} signifies the activity threshold, which delineates the upper limit for the growth phase of clusters. The assumption is that post the first wave, only 1% of clusters continue to exhibit growth, with the remainder reverting to a baseline dynamic. The activity threshold for Gauteng,

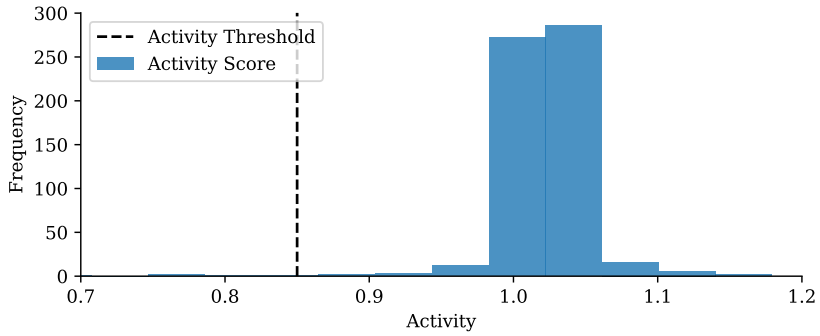


FIGURE B.15: Cluster activity distribution during the first wave in Gauteng.

based on empirical data from the first wave, is established at 0.85, as illustrated in Figure B.15.

Risk index definition

The risk index (RI) is a critical parameter for assessing the potential deviation of cluster behaviour from the expected single-wave pattern, providing a metric for the risk of non-stochastic progression in subsequent waves. The RI is mathematically articulated in the following criteria, outlined in Equations B.14 and B.15:

$$A(t) = 100 \cdot \left(\frac{N_{TC}(t) - SI_0(t)}{SI_0(t)} \right), \quad B(t) = 10 \cdot \left(1 + \frac{SI_0(t)}{N_{TC}(t)} \right)^{-1}, \tag{B.14}$$

$$RI = \begin{cases} A(t) + B(t), & \text{if } B(t) > 8, \\ A(t), & \text{if } B(t) \leq 8. \end{cases} \tag{B.15}$$

Here, $N_{TC}(t)$ denotes the total number of cases and $SI_0(t)$ represents the predicted total cases in the cluster for time, t. The application of Criterion B.15 to hot-spot and stochastic clusters yields the distribution seen in Figure B.16. The RI threshold is established on the premise that merely 1% of clusters pose a high risk in the aftermath of the first wave, accounting

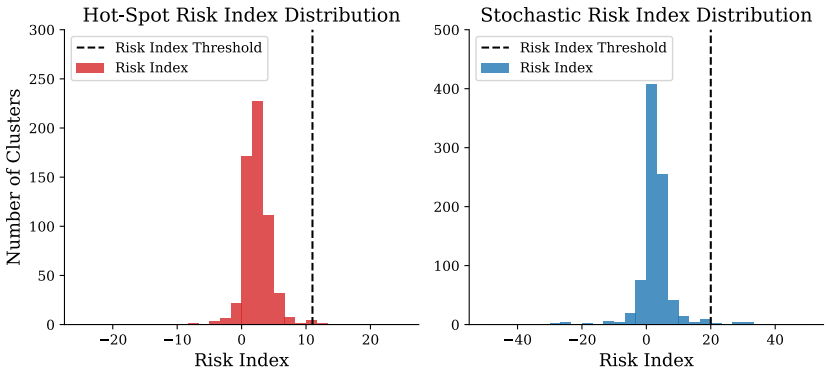


FIGURE B.16: Risk index distributions for the first wave in Gauteng. The distributions are categorised into hot-spot and stochastic clusters for the first wave period.

for the associated proportional error. Figure B.16 displays the RI value that qualifies a cluster as a high-risk entity within the Gauteng Province.

Thus, for subsequent waves in Gauteng, a hot-spot cluster with an RI exceeding 11 is deemed a high-risk hot-spot. Similarly, a cluster not previously identified as a hot-spot that registers an RI above 20 is regarded as an emergent high-risk cluster.

B.3.4 Subsequent wave clustering and hot-spot definition

The identification of clusters as hot-spots, and the subsequent definitions of activity and risk index, hinge on the calibration of historical data. Within the scope of this research, the calibration is based on the data from the first wave. With the onset of subsequent waves, clustering of cases is conducted in real-time with the incoming data, utilising the pre-calibrated criteria. As fresh data continues to be reported, clustering and the associated analyses are iteratively applied to the new dataset. It is crucial to note that for each new iteration, the data pertaining to the current analysis period are clustered anew, without being influenced by clusters previously identified.

This methodology affords a continuous and current assessment of the evolving landscape, pinpointing the locations, tracking the temporal dynamics, and gauging the intensity of active hot-spots. Furthermore, it facilitates the

early recognition of clusters that exhibit a high probability of escalating into hot-spots. The key points to this approach are:

- Clustering of new cases is performed iteratively with each new batch of data.
- Previous clusters are not used as a reference; each new data set is analysed independently.
- The approach enables real-time monitoring and analysis of disease progression during subsequent waves.
- It provides the ability to dynamically identify and evaluate the severity of active and potential hot-spots.

B.3.5 Results and discussion

Analysis of Cluster Definition Calibration on Gauteng Province's First Wave

The calibration of the density criterion was anchored to the first wave of COVID-19 case data in Gauteng Province. We denote $\rho_{cluster}(t)$ as the case density for a specific cluster at time, t and ρ_{th} as the established density threshold that characterises a cluster as a hot-spot. Out of $1,500$ clusters examined, 607 were classified as hot-spots post the application of the density threshold criterion, while the remaining 893 were considered normal clusters.

The validity of this classification was scrutinised by juxtaposing the susceptible-infection parameters of the designated hot-spot clusters against those classified as stochastic or non-hot-spot. As illustrated in Figure B.17, the average number of total cases in hot-spot clusters was observed to be higher by ± 180, in contrast to ± 90 in stochastic clusters. Furthermore, the hot-spot clusters exhibited a marginally higher exponential growth period of ± 10 days, as opposed to ± 11 days for stochastic clusters.

To further substantiate this definition, we analysed the distribution of total cases within stochastic and hot-spot clusters throughout the first wave. As

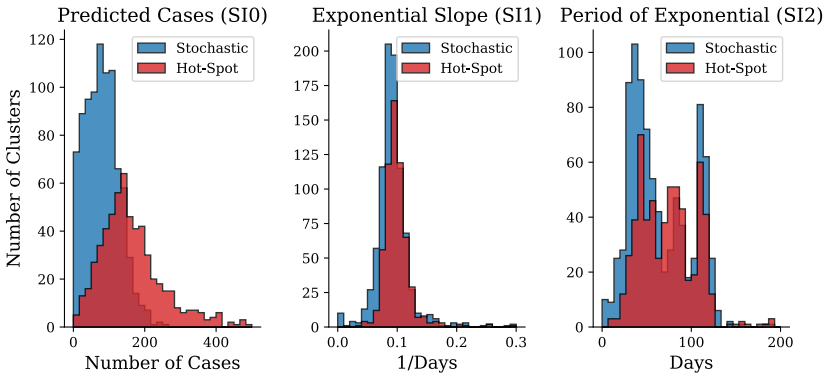


FIGURE B.17: Gauteng first wave cluster parameter distribution comparison. Distributions of susceptible-infectious parameters for clusters are depicted.

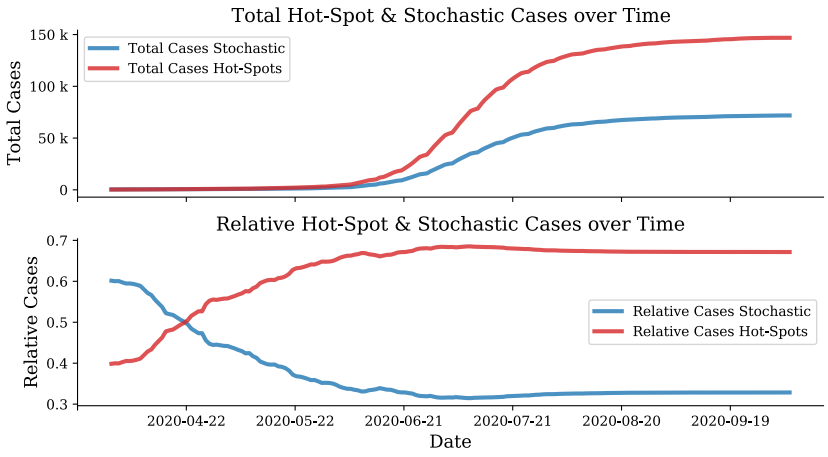


FIGURE B.18: Number of hot-spot cases over time during the first wave.

portrayed in Figure B.18, approximately two-thirds of the cases reported in Gauteng were localised within hot-spot clusters.

The distribution of cases was highly consistent with first wave projections made using a di-SIRD linear control model [155]. Figure B.19 delineates this correlation, exemplifying that the unexpected emergence of hot-spots in June 2020 deviated from the anticipated stochastic progression of the virus.

Conclusively, the density threshold of 196 cases/km^2, employed to demarcate hot-spot clusters, proved efficacious in segregating clusters with rapid and irregular growth from those exhibiting a more stochastic, randomised

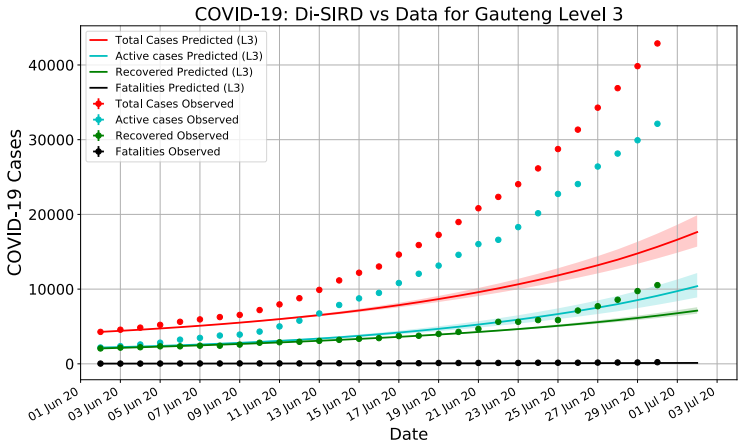


FIGURE B.19: Comparative analysis of the Di-SIRD model's stochastic predictions against actual data for June 2020 in Gauteng.

pattern of expansion.

Hot-spot activity analysis

The temporal evolution of newly identified hot-spots and the reversion of existing hot-spots to stochastic behaviour during the first wave can be investigated utilising Criterion B.13. These dynamics are depicted in Figures B.20 and B.21.

To elucidate the expansion of hot-spot clusters, an SI curve is fitted to the cumulative data of hot-spot clusters as shown in Figure B.20. The apex of daily hot-spot cluster increments is pinpointed in mid-July, corroborated by the SI_2 parameter which denotes the shift from exponential growth on July 10th, 101 days after the 1st of April. The plateau in the accumulation of hot-spot clusters synchronises with South Africa's transition from lockdown level 3 to level 2, during which time 594 out of the total 1,500 clusters had evolved into hot-spots. The SI model fitted to the cumulative hot-spot data prescribes the exponential growth phase to span roughly 12 days, $\frac{1}{SI_1}$.

Moreover, Figure B.21 elucidates not only the emergence but also the regression of hot-spot clusters as they revert to stochastic dynamics, as defined by Criterion B.13. Post mid-July, a predominant number of hot-spot clusters

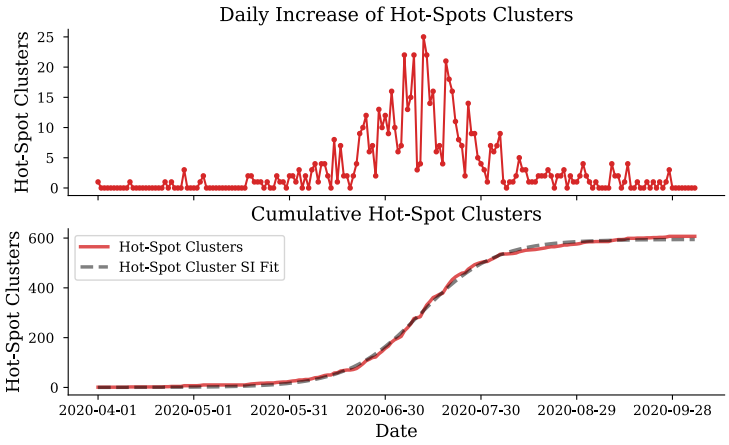


FIGURE B.20: First wave cumulative and emerging COVID-19 hot-spot clusters in Gauteng. Top: The emergence of new clusters transitioning into hot-spots. Bottom: The total number of clusters identified as hot-spots throughout the first wave in Gauteng.

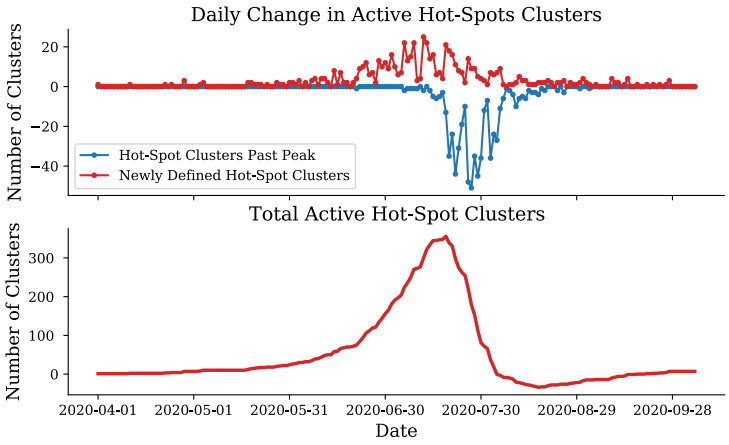


FIGURE B.21: Daily number of active hot-spot clusters. Top: The flux of clusters developing into active hot-spots and those reverting from hot-spot status. Bottom: The cumulative count of clusters evolving into active hot-spots.

attain their peak and consequently regress to stochastic profiles. By the conclusion of August, a pinnacle of 39 hot-spots is observed, and by the end of September, all but 21 clusters have regressed to their baseline dynamics.

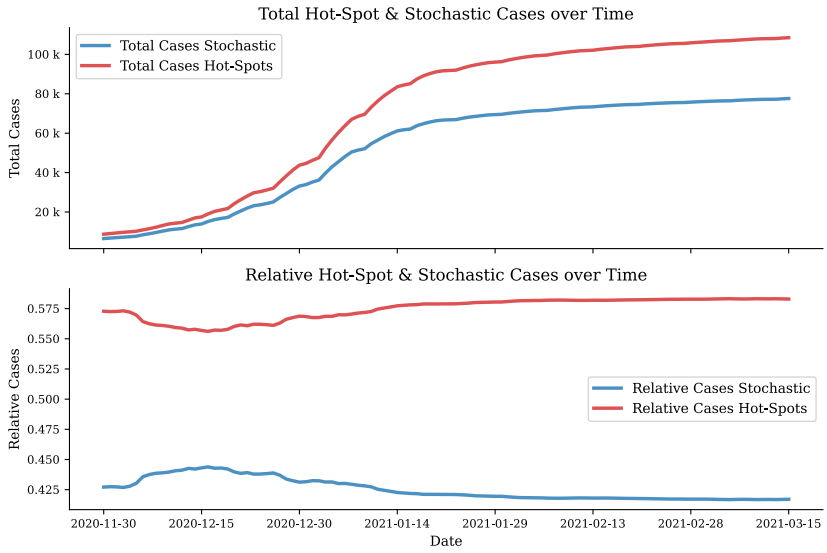


FIGURE B.22: Number of hot-spot cases over time during the second wave. The figure delineates a comparative analysis of hot-spot and stochastic growth via the number of cases per day.

Implementation of hot-spot definition on Gauteng's second wave

The refined cluster definitions derived from historical data were employed in the examination of subsequent waves. Given that hot-spots predominantly emerge during pandemic waves, it is crucial to apply the analysis to data from a specific wave. This section utilises first wave definitions to scrutinise the second wave data. Throughout the second wave, clustering and analysis were conducted on a weekly basis upon the receipt of new data, assessing a total of 191,750 cases. Figure B.22 illustrates the distribution of cases between hot-spots and normal clusters, indicating that nearly 60% of cases were linked to hot-spots during the second wave.

Out of the scrutinised clusters, 461 were designated as hot-spot clusters, while 1039 exhibited normal growth patterns, as depicted in Figure B.23.

As the second wave initiated on 11 November 2020, 48 clusters were active hot-spots. By the second wave's conclusion on 15 March 2021, 7 hot-spots persisted, with 454 reverting to normalcy.

Examining the clusters through the lens of their risk index revealed that

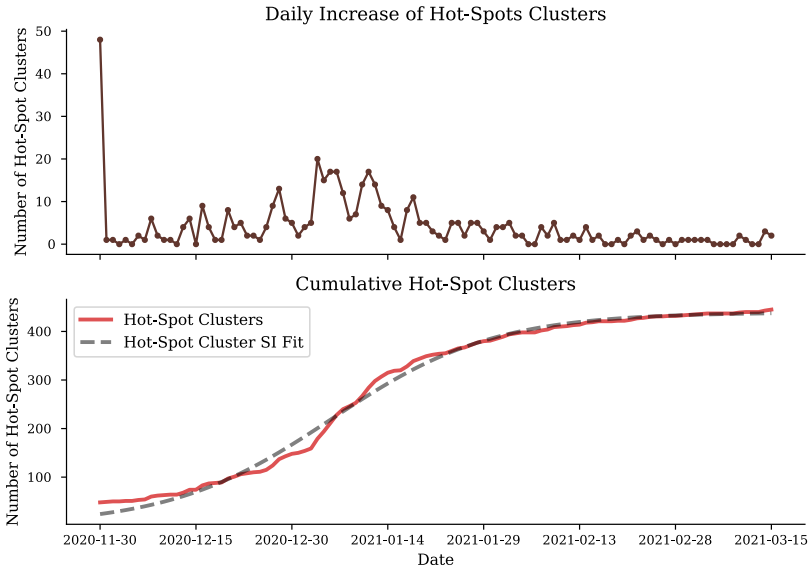


FIGURE B.23: Second wave cumulative and emerging COVID-19 hot-spot clusters in Gauteng. top figure: the emerging new clusters transitioning into hot-spots. Bottom figure totals the hot-spot clusters identified throughout the second wave in Gauteng.

313 hot-spot clusters were earmarked as high-risk, and 263 stochastic clusters were emerging clusters. Over 90% of the clusters initially identified as emerging evolved into hot-spots during the second wave. This risk index served to apprise local stakeholders of wards and municipalities corresponding with severe hot-spot clusters. The most acute wards during the second wave included ward 74804016 in Merafong City, wards 79700088 and 79700005 in Ekurhuleni, wards 74201033, 74201007, 74201008, 74201036, 74201010, and 74201024 in Emfuleni, wards 79900105, 79900082, and 79900061 in the City of Tshwane, and wards 79800053 and 79800061 in the City of Johannesburg municipalities.

With real-time analysis of cluster activity, severity, and location during the second wave, both provincial and municipal stakeholders were equipped to visualise and identify clusters of interest, thereby unveiling location-specific virus dynamics.

TABLE B.2: Summary of hot-spot model specifications for classified clusters.

Hot-Spot Classification	**Cluster Activity**	**Risk Index**
If cluster can be defined as a hot-spot or not	The time-dependent progression of the cluster	The severity of infection rate and scale of clusters

Exposure and applications of hot-spots:

The delineation and parameterisation of clustered cases afford several applications that aid stakeholders in their decision-making concerning COVID-19 interventions and preventive actions. This section explores two principal applications. The first, and arguably most critical, is the identification of locations experiencing extreme viral dynamics to guide intervention strategies, promote social awareness, and encourage the adoption of appropriate behaviours. The second application involves incorporating hot-spot dynamics into epidemiological models.

The fundamental impetus for classifying COVID-19 Hot-Spots is to pinpoint clusters or areas exhibiting anomalous, exponential growth. Utilising the hot-spot cluster definition established in this study, clusters can be effectively categorised, and their progression and dynamics can be tracked. Table B.2 presents a summary of the descriptive parameters for a classified cluster.

These parameters provide stakeholders with comprehensive insights into areas deemed high-risk for COVID-19 proliferation, including the duration of cluster activity and the intensity of the cluster's spread. These can be visualised on an interactive map, as depicted in Figure B.24, with the colour coding indicating severity based on the RI.

Spatio-temporal hot-spot analysis is vitally important for public health policymakers and decision-makers. It provides unique insights that cannot be achieved with other methods, allowing for the identification of specific clustering patterns in districts and areas that might otherwise be considered low

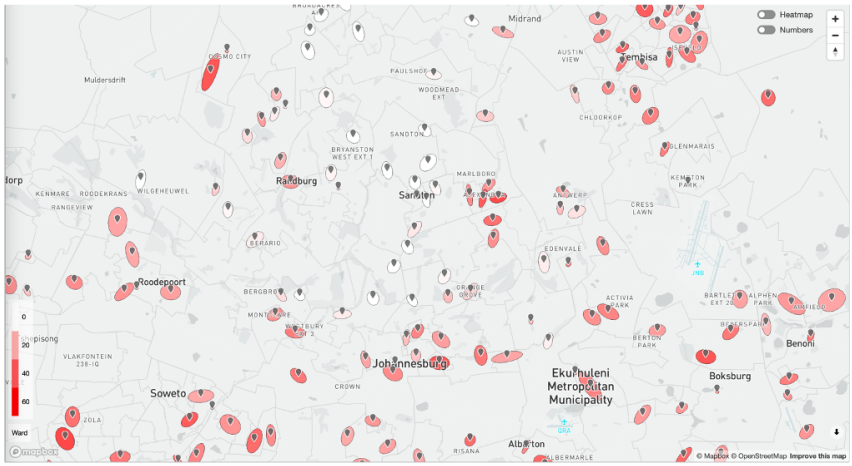


FIGURE B.24: Hot-Spot visualisation on gpcoronavirus.co.za. Courtesy of IBM South Africa.

risk for COVID-19 spread. Hot-spot analysis complements traditional epidemiological and surveillance strategies by illuminating the spatio- temporal progression of COVID-19 and potential future trends, while simultaneously facilitating the visualisation of data for stakeholder accessibility and informed decision-making.

One challenge in modelling the COVID-19 pandemic is the stochastic nature of SIRD models, which assume random spread (β) through a susceptible population. Yet pockets of cases often in densely populated regions, experience rapid, independent infection rates that deviate from broader models. These micro-system clusters, or hot-spots, exhibit distinct non-stochastic dynamics. To classify a group of cases as a hot-spot, the cases must first be aggregated and their attributes modelled, using the resulting characteristics to define a hot-spot cluster.

Hence, to generate actionable forecasts for policymakers and decision- makers, such as estimates of hospital bed occupancy, ICU ward usage, and peak timing — hot-spot cluster cases must be isolated from the data on which the stochastic SIRD model is calibrated. This adjustment allows the model to accurately track the progression of COVID-19 without the skew introduced by the non-standard hot-spot cases.

This isolation is achieved by separating the daily proportion of stochastic

cases from the total cases in all clusters and adjusting the reported data before its utilisation in the model:

$$I_{\text{stoch}} = \frac{I_s}{I_s + I_{\text{hs}}} I_d, \tag{B.16}$$

where I_{stoch} represents the stochastic active cases, I_s is the active cases in stochastic clusters, I_{hs} denotes the active cases in hot-spot clusters, and I_d is the recorded active cases.

B.3.6 Conclusions

Hot-spot analysis embodies a sophisticated statistical methodology that proves invaluable for the analytics and visual representation of outbreaks. It endows public health policymakers and decision-makers with the tools for a contemporaneous, precise appraisal of pandemic patterns and prospective trajectories. Additionally, this analysis augments traditional epidemiological investigations by revealing patterns otherwise overlooked and deemed low-risk. In summation, hot-spot analysis has shown immense utility in rapidly pinpointing high-risk clusters, thereby enabling the timely implementation or modification of public health strategies. Given the dynamic and ever-evolving nature of epidemics, it is reasonable to forecast that hot-spot analysis will continue to empower stakeholders to formulate informed, substantiated, and data-oriented decisions amidst epidemic waves and during critical initiatives such as vaccine distributions.

Bibliography

[1] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nature Communications*, vol. 5, no. 1, 2014. DOI: 10.1038/ncomms5308.

[2] *The standard model*. [Online]. Available: https://www.home.cern/science/physics/standard-model.

[3] G. Aad, T. Abajyan, B. Abbott, *et al.*, "Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc," *Physics Letters B*, vol. 716, no. 1, pp. 1–29, 2012. DOI: 10.1016/j.physletb.2012.08.020.

[4] S. Chatrchyan, V. Khachatryan, A. Sirunyan, A. Tumasyan, *et al.*, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Physics Letters B*, vol. 716, no. 1, pp. 30–61, 2012. DOI: 10.1016/j.physletb.2012.08.021.

[5] M. Tanabashi, K. Hagiwara, K. Hikasa, *et al.*, "Review of particle physics," *Phys. Rev. D*, vol. 98, p. 030 001, 3 2018. DOI: 10.1103/PhysRevD.98.030001. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[6] ATLAS Calibration, "A detailed map of higgs boson interactions by the atlas experiment ten years after the discovery," *Nature*, vol. 607, no. 7917, pp. 52–59, 2022. DOI: 10.1038/s41586-022-04893-w.

[7] S. L. Glashow, "Partial-symmetries of weak interactions," *Nuclear Physics*, vol. 22, no. 4, pp. 579–588, 1961, ISSN: 0029-5582. DOI: https://doi.org/10.1016/0029-5582(61)90469-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0029558261904692.

[8] S. Weinberg, "A model of leptons," *Phys. Rev. Lett.*, vol. 19, pp. 1264–1266, 21 1967. DOI: 10.1103/PhysRevLett.19.1264. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.19.1264.

[9] A. Salam, "Weak and Electromagnetic Interactions," *Conf. Proc. C*, vol. 680519, pp. 367–377, 1968. DOI: 10.1142/9789812795915_0034.

[10] P. W. Higgs, "Broken symmetries and the masses of gauge bosons," *Phys. Rev. Lett.*, vol. 13, pp. 508–509, 16 1964. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.13.508.

[11] P. W. Higgs, "Spontaneous symmetry breakdown without massless bosons," *Phys. Rev.*, vol. 145, pp. 1156–1163, 4 1966. DOI: 10.1103/PhysRev.145.1156. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.145.1156.

[12] C. T. Potter *et al.*, *Handbook of lhc higgs cross sections: 3. higgs properties: Report of the lhc higgs cross section working group*, en, 2013. DOI: 10.5170/CERN-2013-004. [Online]. Available: http://cds.cern.ch/record/1559921.

[13] G. Aad *et al.*, "Evidence for the spin-0 nature of the higgs boson using ATLAS data," *Physics Letters B*, vol. 726, no. 1-3, pp. 120–144, 2013. DOI: 10.1016/j.physletb.2013.08.026. [Online]. Available: https://doi.org/10.1016%2Fj.physletb.2013.08.026.

[14] S. Chatrchyan *et al.*, "Study of the mass and spin-parity of the higgs boson candidate via its decays Z to boson pairs," *Physical Review Letters*, vol. 110, no. 8, 2013. DOI: 10.1103/physrevlett.110.081803. [Online]. Available: https://doi.org/10.1103%2Fphysrevlett.110.081803.

[15] S. Haywood, *Symmetries and conservation laws in particle physics: an introduction to group theory for particle physicists*. World scientific, 2011.

[16] C. N. Yang and R. L. Mills, "Conservation of isotopic spin and isotopic gauge invariance," *Phys. Rev.*, vol. 96, pp. 191–195, 1 1954. DOI: 10.1103/PhysRev.96.191. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.96.191.

[17] J. M. Campbell, J. W. Huston, and W. J. Stirling, "Hard interactions of quarks and gluons: A primer for LHC physics," *Reports on Progress in Physics*, vol. 70, no. 1, pp. 89–193, 2006. DOI: 10.1088/0034-4885/

70/1/r02. [Online]. Available: https://doi.org/10.1088%2F0034-4885%2F70%2F1%2Fr02.

[18] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson, "Experimental test of parity conservation in beta decay," *Phys. Rev.*, vol. 105, pp. 1413–1415, 4 1957. DOI: 10.1103/PhysRev.105.1413. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.105.1413.

[19] M. Gell-Mann, "The interpretation of the new particles as displaced charge multiplets," *Il Nuovo Cimento*, vol. 4, no. S2, 848–866, 1956. DOI: 10.1007/bf02748000.

[20] T. Nakano and K. Nishijima, "Charge Independence for V-particles*," *Progress of Theoretical Physics*, vol. 10, no. 5, pp. 581–582, Nov. 1953, ISSN: 0033-068X. DOI: 10.1143/PTP.10.581. eprint: https://academic.oup.com/ptp/article-pdf/10/5/581/5364926/10-5-581.pdf. [Online]. Available: https://doi.org/10.1143/PTP.10.581.

[21] F. Englert and R. Brout, "Broken symmetry and the mass of gauge vector mesons," *Phys. Rev. Lett.*, vol. 13, pp. 321–323, 9 1964. DOI: 10.1103/PhysRevLett.13.321. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.13.321.

[22] T. W. B. Kibble, "Symmetry breaking in non-abelian gauge theories," *Phys. Rev.*, vol. 155, pp. 1554–1561, 5 1967. DOI: 10.1103/PhysRev.155.1554. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRev.155.1554.

[23] V. Sahni, "5 dark matter and dark energy," in *Lecture Notes in Physics*. Springer Berlin Heidelberg, Dec. 2004, 141–179, ISBN: 9783540315353. DOI: 10.1007/978-3-540-31535-3_5. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-31535-3_5.

[24] P. Di Bari, "On the origin of matter in the universe," *Progress in Particle and Nuclear Physics*, vol. 122, p. 103 913, 2022, ISSN: 0146-6410. DOI: https://doi.org/10.1016/j.ppnp.2021.103913. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0146641021000740.

[25] Y. Farzan and M. Tórtola, "Neutrino oscillations and non-standard interactions," *Frontiers in Physics*, vol. 6, p. 10, 2018.

[26] E. Gross and O. Vitells, "Trial factors for the look elsewhere effect in high energy physics," *The European Physical Journal C*, vol. 70, no. 1–2, 525–530, 2010. DOI: 10.1140/epjc/s10052-010-1470-8.

[27] J. F. Gunion, H. Haber, and S Dawson, *The Higgs hunter's guide*. ABP, PERSEUS PUBLISHING, Cambridge, 1990.

[28] "Theory and phenomenology of two-higgs-doublet models," *Physics Reports*, vol. 516, no. 1, pp. 1–102, 2012, Theory and phenomenology of two-Higgs-doublet models, ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2012.02.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0370157312000695.

[29] S. von Buddenbrock *et al.*, "Phenomenological signatures of additional scalar bosons at the lhc," *The European Physical Journal C*, vol. 76, no. 10, 2016, ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-016-4435-8. [Online]. Available: http://dx.doi.org/10.1140/epjc/s10052-016-4435-8.

[30] S. von Buddenbrock *et al.*, "The compatibility of LHC Run 1 data with a heavy scalar of mass around 270\,GeV," Jun. 2015. arXiv: 1506.00612 [hep-ph].

[31] S. von Buddenbrock, A. S. Cornell, M. Kumar, and B. Mellado, "The Madala hypothesis with Run 1 and 2 data at the LHC," *J. Phys. Conf. Ser.*, vol. 889, no. 1, p. 012 020, 2017. DOI: 10.1088/1742-6596/889/1/012020. arXiv: 1709.09419 [hep-ph].

[32] S. von Buddenbrock, A. S. Cornell, A. Fadol, M. Kumar, B. Mellado, and X. Ruan, "Multi-lepton signatures of additional scalar bosons beyond the Standard Model at the LHC," *J. Phys. G*, vol. 45, no. 11, p. 115 003, DOI: 10.1088/1361-6471/aae3d6. arXiv: 1711.07874 [hep-ph].

[33] S. von Buddenbrock *et al.*, "Constraints on a 2HDM with a singlet scalar and implications in the search for heavy bosons at the LHC," *J. Phys. G*, vol. 46, no. 11, p. 115 001, 2019. DOI: 10.1088/1361-6471/ab3cf6. arXiv: 1809.06344 [hep-ph].

[34] S. Buddenbrock *et al.*, "The emergence of multi-lepton anomalies at the LHC and their compatibility with new physics at the EW scale,"

JHEP, vol. 10, p. 157, 2019. DOI: 10.1007/JHEP10(2019)157. arXiv: 1901.05300 [hep-ph].

[35] D. Sabatta, A. S. Cornell, A. Goyal, M. Kumar, B. Mellado, and X. Ruan, "Connecting muon anomalous magnetic moment and multi-lepton anomalies at LHC," *Chin. Phys. C*, vol. 44, no. 6, p. 063 103, 2020. DOI: 10.1088/1674-1137/44/6/063103. arXiv: 1909.03969 [hep-ph].

[36] Y. Hernandez *et al.*, "The anomalous production of multi-lepton and its impact on the measurement of *Wh* production at the LHC," *Eur. Phys. J. C*, vol. 81, no. 4, p. 365, 2021. DOI: 10.1140/epjc/s10052-021-09137-1. arXiv: 1912.00699 [hep-ph].

[37] S. von Buddenbrock, R. Ruiz, and B. Mellado, "Anatomy of inclusive $t\bar{t}W$ production at hadron colliders," *Phys. Lett. B*, vol. 811, p. 135 964, 2020. DOI: 10.1016/j.physletb.2020.135964. arXiv: 2009.00032 [hep-ph].

[38] A. Crivellin *et al.*, "Accumulating evidence for the associated production of a new Higgs boson at the LHC," *Phys. Rev. D*, vol. 108, no. 11, p. 115 031, 2023. DOI: 10.1103/PhysRevD.108.115031. arXiv: 2109.02650 [hep-ph].

[39] G. Beck, R. Temo, E. Malwa, M. Kumar, and B. Mellado, "Connecting multi-lepton anomalies at the LHC and in astrophysics with MeerKAT / SKA," *Astroparticle Physics*, vol. 148, p. 102 821, 2023. DOI: 10.1016/j.astropartphys.2023.102821. [Online]. Available: https://doi.org/10.1016%2Fj.astropartphys.2023.102821.

[40] B. Abi *et al.*, "Measurement of the Positive Muon Anomalous Magnetic Moment to 0.46 ppm," *Phys. Rev. Lett.*, vol. 126, no. 14, p. 141 801, 2021. DOI: 10.1103/PhysRevLett.126.141801. arXiv: 2104.03281 [hep-ex].

[41] T. Aoyama *et al.*, "The anomalous magnetic moment of the muon in the Standard Model," *Phys. Rept.*, vol. 887, pp. 1–166, 2020. DOI: 10.1016/j.physrep.2020.07.006. arXiv: 2006.04822 [hep-ph].

[42] O. Fischer *et al.*, "Unveiling hidden physics at the LHC," *Eur. Phys. J. C*, vol. 82, no. 8, p. 665, 2022. DOI: 10.1140/epjc/s10052-022-10541-4. arXiv: 2109.06065 [hep-ph].

[43] A. Crivellin and B. Mellado, "Anomalies in Particle Physics," Sep. 2023. arXiv: 2309.03870 [hep-ph].

[44] S. Banik, G. Coloretti, A. Crivellin, and B. Mellado, "Uncovering New Higgses in the LHC Analyses of Differential $t\bar{t}$ Cross Sections," Aug. 2023. arXiv: 2308.07953 [hep-ph].

[45] G. Aad *et al.*, "Inclusive and differential cross-sections for dilepton $t\bar{t}$ production measured in $\sqrt{s}$ = 13 tev pp collisions with the atlas detector," *Journal of High Energy Physics*, vol. 2023, no. 7, Jul. 2023, ISSN: 1029-8479. DOI: 10.1007/jhep07(2023)141. [Online]. Available: http://dx.doi.org/10.1007/JHEP07(2023)141.

[46] M. Aaboud *et al.*, "Measurement of fiducial and differential W^+W^- production cross-sections at $\sqrt{s} = 13$ tev with the atlas detector," *The European Physical Journal C*, vol. 79, no. 10, Oct. 2019, ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-019-7371-6. [Online]. Available: http://dx.doi.org/10.1140/epjc/s10052-019-7371-6.

[47] ATLAS Collaboration, "Observation of four-top-quark production in the multilepton final state with the atlas detector," *Eur. Phys. J. C*, vol. 83, no. 6, p. 496, 2023. DOI: 10.1140/epjc/s10052-023-11573-0. [Online]. Available: https://doi.org/10.1140/epjc/s10052-023-11573-0.

[48] A. Hayrapetyan, A. Tumasyan, W. Adam, and CMS Collaboration, "Observation of four top quark production in proton-proton collisions at s=13tev," *Physics Letters B*, vol. 847, p. 138 290, 2023, ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2023.138290. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S037026932300624X.

[49] ATLAS Collaboration, "Observation of www production in pp collisions at s=13tev with the atlas detector," *Physical Review Letters*, vol. 129, no. 6, Aug. 2022, ISSN: 1079-7114. DOI: 10.1103/physrevlett.129.061803. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.129.061803.

[50] A. Tumasyan *et al.*, "Measurements of the Higgs boson production cross section and couplings in the W boson pair decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV," *Eur. Phys. J. C*, vol. 83,

no. 7, p. 667, 2023. DOI: 10.1140/epjc/s10052-023-11632-6. arXiv: 2206.09466 [hep-ex].

[51] A. Crivellin and B. Mellado, *Anomalies in particle physics*, 2023. arXiv: 2309.03870 [hep-ph].

[52] G. Coloretti, A. Crivellin, S. Bhattacharya, and B. Mellado, "Searching for low-mass resonances decaying into W bosons," *Phys. Rev. D*, vol. 108, no. 3, p. 035 026, 2023. DOI: 10.1103/PhysRevD.108.035026. arXiv: 2302.07276 [hep-ph].

[53] S. Bhattacharya *et al.*, "Growing Excesses of New Scalars at the Electroweak Scale," Jun. 2023. arXiv: 2306.17209 [hep-ph].

[54] G Aad, E Abat, J Abdallah, *et al.*, "The atlas experiment at the cern large hadron collider," *Journal of Instrumentation*, vol. 3, no. 08, S08003, 2008. DOI: 10.1088/1748-0221/3/08/S08003. [Online]. Available: https://dx.doi.org/10.1088/1748-0221/3/08/S08003.

[55] O. S. Brüning *et al.*, *LHC Design Report* (CERN Yellow Reports: Monographs). Geneva: CERN, 2004. DOI: 10.5170/CERN-2004-003-V-1. [Online]. Available: https://cds.cern.ch/record/782076.

[56] *Cern accelerating science*. [Online]. Available: https://home.cern/science/accelerators/accelerator-complex.

[57] E. Lopienska, *The cern accelerator complex, layout in 2022. complexe des accélérateurs du cern en janvier 2022*, General Photo, 2022. [Online]. Available: https://cds.cern.ch/record/2800984.

[58] A. Airapetian, V. Grabsky, H. Hakopian, *et al.*, *ATLAS detector and physics performance: Technical Design Report, 1* (Technical design report. ATLAS). Geneva: CERN, 1999. [Online]. Available: https://cds.cern.ch/record/391176.

[59] J. Pequenao, *Computer generated image of the whole ATLAS detector*, 2008. [Online]. Available: https://cds.cern.ch/record/1095924.

[60] J. Pequenao, *Computer generated image of the ATLAS inner detector*, 2008. [Online]. Available: https://cds.cern.ch/record/1095926.

[61] J. Pequenao, *Computer Generated image of the ATLAS calorimeter*, 2008. [Online]. Available: https://cds.cern.ch/record/1095927.

[62] J. Pequenao, *Computer generated image of the ATLAS Muons subsystem*, 2008. [Online]. Available: https://cds.cern.ch/record/1095929.

[63] A. M. Rodriguez Vera and J. Antunes Pequenao, *ATLAS Detector Magnet System*, General Photo, 2021. [Online]. Available: https://cds.cern.ch/record/2770604.

[64] W. Panduro Vazquez and A. Collaboration, "The ATLAS Data Acquisition system in LHC Run 2," CERN, Geneva, Tech. Rep. 3, 2017. DOI: 10.1088/1742-6596/898/3/032017. [Online]. Available: https://cds.cern.ch/record/2244345.

[65] A. collaboration, "Operation of the atlas trigger system in run 2," *Journal of Instrumentation*, vol. 15, no. 10, 2020. DOI: 10.1088/1748-0221/15/10/p10004.

[66] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "Gpt-4 passes the bar exam," *SSRN Electronic Journal*, 2023. DOI: 10.2139/ssrn.4389233.

[67] A. Sirunyan, A. Tumasyan, W. Adam, and other, "Identification of heavy, energetic, hadronically decaying particles using machine learning techniques," *Journal of Instrumentation*, vol. 15, no. 06, P06005–P06005, 2020. DOI: 10.1088/1748-0221/15/06/p06005.

[68] J. Collins, K. Howe, and B. Nachman, "Anomaly detection for resonant new physics with machine learning," *Physical Review Letters*, vol. 121, no. 24, 2018, ISSN: 1079-7114. DOI: 10.1103/physrevlett.121.241803. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.121.241803.

[69] "Quark versus Gluon Jet Tagging Using Jet Images with the ATLAS Detector," CERN, Geneva, Tech. Rep., 2017. [Online]. Available: http://cds.cern.ch/record/2275641.

[70] "Convolutional Neural Networks with Event Images for Pileup Mitigation with the ATLAS Detector," CERN, Geneva, Tech. Rep., 2019. [Online]. Available: http://cds.cern.ch/record/2684070.

[71] O. Bakina *et al.*, "Deep learning for track recognition in pixel and strip-based particle detectors," *Journal of Instrumentation*, vol. 17, no. 12, P12023, 2022. DOI: 10.1088/1748-0221/17/12/p12023.

[72] ATLAS Collaboration, *Deep generative models for fast photon shower simulation in atlas*, 2022. DOI: 10.48550/ARXIV.2210.06204. [Online]. Available: https://arxiv.org/abs/2210.06204.

[73] M. Feickert and B. Nachman, *A living review of machine learning for particle physics*, 2021. DOI: 10.48550/ARXIV.2102.02770. [Online]. Available: https://arxiv.org/abs/2102.02770.

[74] L. M. Dery, B. Nachman, F. Rubbo, and A. Schwartzman, "Weakly supervised classification in high energy physics," *Journal of High Energy Physics*, vol. 2017, no. 5, 2017. DOI: 10.1007/jhep05(2017)145. [Online]. Available: https://doi.org/10.1007%2Fjhep05%282017%29145.

[75] J. S. H. Lee, S. M. Lee, Y. Lee, I. Park, I. J. Watson, and S. Yang, "Quark gluon jet discrimination with weakly supervised learning," *Journal of the Korean Physical Society*, vol. 75, no. 9, 652–659, Nov. 2019, ISSN: 1976-8524. DOI: 10.3938/jkps.75.652. [Online]. Available: http://dx.doi.org/10.3938/jkps.75.652.

[76] M. Kuusela, T. Vatanen, E. Malmi, T. Raiko, T. Aaltonen, and Y. Nagai, "Semi-supervised anomaly detection - towards model independent searches of new physics," *Journal of Physics: Conference Series*, vol. 368, no. 1, p. 012 032, 2012. DOI: 10.1088/1742-6596/368/1/012032. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/368/1/012032.

[77] E. M. Metodiev, B. Nachman, and J. Thaler, "Classification without labels: Learning from mixed samples in high energy physics," *Journal of High Energy Physics*, vol. 2017, no. 10, Oct. 2017, ISSN: 1029-8479. DOI: 10.1007/jhep10(2017)174. [Online]. Available: http://dx.doi.org/10.1007/JHEP10(2017)174.

[78] J. H. Collins, K. Howe, and B. Nachman, "Extending the search for new resonances with machine learning," *Physical Review D*, vol. 99, no. 1, 2019, ISSN: 2470-0029. DOI: 10.1103/physrevd.99.014038. [Online]. Available: http://dx.doi.org/10.1103/PhysRevD.99.014038.

[79] P. T. Komiske, E. M. Metodiev, and J. Thaler, "An operational definition of quark and gluon jets," *Journal of High Energy Physics*, vol. 2018, no. 11, 2018, ISSN: 1029-8479. DOI: 10.1007/jhep11(2018)059. [Online]. Available: http://dx.doi.org/10.1007/JHEP11(2018)059.

[80] P. T. Komiske, S. Kryhin, and J. Thaler, "Disentangling quarks and gluons in cms open data," *Physical Review D*, vol. 106, no. 9, 2022,

ISSN: 2470-0029. DOI: 10.1103/physrevd.106.094021. [Online]. Available: http://dx.doi.org/10.1103/PhysRevD.106.094021.

[81] T. Finke, M. Krämer, M. Lipp, and A. Mück, "Boosting mono-jet searches with model-agnostic machine learning," *Journal of High Energy Physics*, vol. 2022, no. 8, 2022, ISSN: 1029-8479. DOI: 10.1007/jhep08(2022)015. [Online]. Available: http://dx.doi.org/10.1007/JHEP08(2022)015.

[82] G. Aad, B. Abbott, D. C. Abbott, A. Abed Abud, *et al.*, "Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the atlas detector," *Phys. Rev. Lett.*, vol. 125, p. 131 801, 13 2020. DOI: 10.1103/PhysRevLett.125.131801. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.125.131801.

[83] D. Bardhan, Y. Kats, and N. Wunch, "Searching for dark jets with displaced vertices using weakly supervised machine learning," *Physical Review D*, vol. 108, no. 3, 2023, ISSN: 2470-0029. DOI: 10.1103/physrevd.108.035036. [Online]. Available: http://dx.doi.org/10.1103/PhysRevD.108.035036.

[84] M. A. Md Ali, N. Badrud'din, H. Abdullah, and F. Kemi, "Alternate methods for anomaly detection in high-energy physics via semi-supervised learning," *International Journal of Modern Physics A*, vol. 35, no. 23, p. 2 050 131, 2020. DOI: 10.1142/S0217751X20501316. eprint: https://doi.org/10.1142/S0217751X20501316. [Online]. Available: https://doi.org/10.1142/S0217751X20501316.

[85] P. T. Komiske, E. M. Metodiev, B. Nachman, and M. D. Schwartz, "Learning to classify from impure samples with high-dimensional data," *Physical Review D*, vol. 98, no. 1, 2018, ISSN: 2470-0029. DOI: 10.1103/physrevd.98.011502. [Online]. Available: http://dx.doi.org/10.1103/PhysRevD.98.011502.

[86] O. Amram and C. M. Suarez, "Tag n' train: A technique to train improved classifiers on unlabeled data," *Journal of High Energy Physics*, vol. 2021, no. 1, Jan. 2021, ISSN: 1029-8479. DOI: 10.1007/jhep01(2021)153. [Online]. Available: http://dx.doi.org/10.1007/JHEP01(2021)153.

[87] T. Li *et al.*, "Semi-supervised graph neural networks for pileup noise removal," *The European Physical Journal C*, vol. 83, no. 1, 2023, ISSN:

1434-6052. DOI: 10.1140/epjc/s10052-022-11083-5. [Online]. Available: http://dx.doi.org/10.1140/epjc/s10052-022-11083-5.

[88] H. Beauchesne, Z.-E. Chen, and C.-W. Chiang, *Improving the performance of weak supervision searches using transfer and meta-learning*, 2023. arXiv: 2312.06152 [hep-ph].

[89] T. Cohen, M. Freytsis, and B. Ostdiek, "(machine) learning to do more with less," *Journal of High Energy Physics*, vol. 2018, no. 2, Feb. 2018, ISSN: 1029-8479. DOI: 10.1007/jhep02(2018)034. [Online]. Available: http://dx.doi.org/10.1007/JHEP02(2018)034.

[90] S.-E. Dahbi *et al.*, "Machine learning approach for the search of resonances with topological features at the large hadron collider," *International Journal of Modern Physics A*, vol. 37, no. 03, p. 2 150 241, 2022. DOI: 10.1142/S0217751X21502419. eprint: https://doi.org/10.1142/S0217751X21502419. [Online]. Available: https://doi.org/10.1142/S0217751X21502419.

[91] Y. Coadou, "Boosted decision trees," in *Artificial Intelligence for High Energy Physics*, WORLD SCIENTIFIC, 2022, pp. 9–58. DOI: 10.1142/9789811234033_0002.

[92] B. Carlson, Q. Bayer, T. Hong, and S. Roche, "Nanosecond machine learning regression with deep boosted decision trees in FPGA for high energy physics," *Journal of Instrumentation*, vol. 17, no. 09, P09039, 2022. DOI: 10.1088/1748-0221/17/09/p09039.

[93] H.-J. Yang, B. P. Roe, and J. Zhu, "Studies of boosted decision trees for MiniBooNE particle identification," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 555, no. 1-2, pp. 370–385, 2005. DOI: 10.1016/j.nima.2005.09.022.

[94] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997, ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.1997.1504. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S002200009791504X.

[95] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001, ISSN: 00905364. [Online]. Available: http://www.jstor.org/stable/2699986 (visited on 01/24/2023).

[96] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016. DOI: 10.1145/2939672.2939785.

[97] W. J. Nauta, M. Feirtag, and C. Donner, *Fundamental neuroanatomy*. Freeman, 1986.

[98] StudyGlance, *The biological neuron*. [Online]. Available: https://studyglance.in/nn (visited on 01/24/2023).

[99] F. A. Azevedo *et al.*, "Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain," *Journal of Comparative Neurology*, vol. 513, no. 5, pp. 532–541, 2009. DOI: https://doi.org/10.1002/cne.21974. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cne.21974. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.21974.

[100] S. Narayan, "The generalized sigmoid activation function: Competitive supervised learning," *Information Sciences*, vol. 99, no. 1, pp. 69–82, 1997, ISSN: 0020-0255. DOI: https://doi.org/10.1016/S0020-0255(96)00200-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025596002009.

[101] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *The Annals of Statistics*, vol. 48, no. 4, pp. 1875 –1897, 2020. DOI: 10.1214/19-AOS1875. [Online]. Available: https://doi.org/10.1214/19-AOS1875.

[102] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[103] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, 436–444, 2015. DOI: 10.1038/nature14539.

[104] D. Guest, K. Cranmer, and D. Whiteson, "Deep learning and its application to lhc physics," *Annual Review of Nuclear and Particle Science*, vol. 68, no. 1, pp. 161–181, 2018. DOI: 10.1146/annurev-nucl-

101917-021019. eprint: https://doi.org/10.1146/annurev-nucl-101917-021019. [Online]. Available: https://doi.org/10.1146/annurev-nucl-101917-021019.

[105] G Apollinari, I Béjar Alonso, O Brüning, M Lamont, and L Rossi, *High-Luminosity Large Hadron Collider (HL-LHC): Preliminary Design Report* (CERN Yellow Reports: Monographs). Geneva: CERN, 2015. DOI: 10.5170/CERN-2015-005. [Online]. Available: https://cds.cern.ch/record/2116337.

[106] N. Metropolis and S. Ulam, "The monte carlo method," *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, 1949, PMID: 18139350. DOI: 10.1080/01621459.1949.10483310. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1949.10483310. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1949.10483310.

[107] A. Buckley *et al.*, "General-purpose event generators for LHC physics," *Physics Reports*, vol. 504, no. 5, pp. 145–233, 2011. DOI: 10.1016/j.physrep.2011.03.005. [Online]. Available: https://doi.org/10.1016%2Fj.physrep.2011.03.005.

[108] T. Sjöstrand, S. Mrenna, and P. Skands, "A brief introduction to pythia 8.1," *Computer Physics Communications*, vol. 178, no. 11, pp. 852–867, 2008. DOI: 10.1016/j.cpc.2008.01.036. [Online]. Available: https://doi.org/10.1016%2Fj.cpc.2008.01.036.

[109] M. Bähr *et al.*, "Herwig++ physics and manual," *The European Physical Journal C*, vol. 58, no. 4, pp. 639–707, 2008. DOI: 10.1140/epjc/s10052-008-0798-9. [Online]. Available: https://doi.org/10.1140%2Fepjc%2Fs10052-008-0798-9.

[110] T. Gleisberg *et al.*, "Event generation with sherpa 1.1," *Journal of High Energy Physics*, vol. 2009, no. 02, p. 007, 2009. DOI: 10.1088/1126-6708/2009/02/007. [Online]. Available: https://dx.doi.org/10.1088/1126-6708/2009/02/007.

[111] J. Alwall *et al.*, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *Journal of High Energy Physics*, vol. 2014, no. 7, 2014. DOI: 10.1007/jhep07(2014)079. [Online]. Available: https://doi.org/10.1007%2Fjhep07%282014%29079.

[112] R. A. Davis, K.-S. Lii, and D. N. Politis, *Remarks on Some Nonparametric Estimates of a Density Function*, K.-S. Davis Richard A.and Lii and D. N. Politis, Eds. New York, NY: Springer New York, 2011, pp. 95–100, ISBN: 978-1-4419-8339-8. DOI: 10.1007/978-1-4419-8339-8_13. [Online]. Available: https://doi.org/10.1007/978-1-4419-8339-8_13.

[113] E. Parzen, "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962. DOI: 10.1214/aoms/1177704472. [Online]. Available: https://doi.org/10.1214/aoms/1177704472.

[114] I. J. Goodfellow *et al.*, *Generative adversarial networks*, 2014. arXiv: 1406.2661 [stat.ML].

[115] A. Butter, T. Plehn, and R. Winterhalder, "How to GAN LHC events," *SciPost Physics*, vol. 7, no. 6, 2019. DOI: 10.21468/scipostphys.7.6.075. [Online]. Available: https://doi.org/10.21468%2Fscipostphys.7.6.075.

[116] A. Butter, T. Plehn, and R. Winterhalder, "How to GAN event subtraction," *SciPost Physics Core*, vol. 3, no. 2, 2020. DOI: 10.21468/scipostphyscore.3.2.009. [Online]. Available: https://doi.org/10.21468%2Fscipostphyscore.3.2.009.

[117] A. Butter, S. Diefenbacher, G. Kasieczka, B. Nachman, and T. Plehn, "GANplifying event samples," *SciPost Physics*, vol. 10, no. 6, 2021. DOI: 10.21468/scipostphys.10.6.139. [Online]. Available: https://doi.org/10.21468%2Fscipostphys.10.6.139.

[118] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, *Improved training of wasserstein gans*, 2017. DOI: 10.48550/ARXIV.1704.00028. [Online]. Available: https://arxiv.org/abs/1704.00028.

[119] R. D. Ball *et al.*, "Parton distributions with LHC data," *Nuclear Physics B*, vol. 867, no. 2, pp. 244–289, 2013. DOI: 10.1016/j.nuclphysb.2012.10.003. [Online]. Available: https://doi.org/10.1016%2Fj.nuclphysb.2012.10.003.

[120] "ATLAS Pythia 8 tunes to 7 TeV data," CERN, Geneva, Tech. Rep., 2014. [Online]. Available: https://cds.cern.ch/record/1966419.

[121] and 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," *Journal of High Energy Physics*, vol. 2018, no. 10, 2018. DOI: 10.1007/jhep10(2018)117. [Online]. Available: https://doi.org/10.1007%2Fjhep10%282018%29117.

[122] Y. Sumino and H. Yokoya, "Bound-state effects on kinematical distributions of top quarks at hadron colliders," *Journal of High Energy Physics*, vol. 2010, no. 9, 2010. DOI: 10.1007/jhep09(2010)034. [Online]. Available: https://doi.org/10.1007%2Fjhep09%282010%29034.

[123] A. Hoecker *et al.*, *Tmva - toolkit for multivariate data analysis*, 2009. arXiv: physics/0703039 [physics.data-an].

[124] M. Aaboud *et al.*, "Searches for the zγ decay mode of the higgs boson and for new high-mass resonances in pp collisions at s = 13 TeV with the ATLAS detector," *Journal of High Energy Physics*, vol. 2017, no. 10, 2017. DOI: 10.1007/jhep10(2017)112. [Online]. Available: https://doi.org/10.1007%2Fjhep10%282017%29112.

[125] "Measurement of $Z\gamma \to \ell^+\ell^-\gamma$ differential cross-sections in pp collisions at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS detector," CERN, Geneva, Tech. Rep., 2019. [Online]. Available: https://cds.cern.ch/record/2682846.

[126] T. Sjöstrand, S. Mrenna, and P. Skands, "PYTHIA 6.4 physics and manual," *Journal of High Energy Physics*, vol. 2006, no. 05, pp. 026–026, 2006. DOI: 10.1088/1126-6708/2006/05/026. [Online]. Available: https://doi.org/10.1088%2F1126-6708%2F2006%2F05%2F026.

[127] R. D. Ball *et al.*, "Parton distributions from high-precision collider data. Parton distributions from high-precision collider data," *Eur. Phys. J. C*, vol. 77, no. 10, p. 663, 2017. DOI: 10.1140/epjc/s10052-017-5199-5. arXiv: 1706.00428. [Online]. Available: https://cds.cern.ch/record/2267455.

[128] J. de Favereau *et al.*, "Delphes 3: A modular framework for fast simulation of a generic collider experiment," *Journal of High Energy Physics*, vol. 2014, no. 2, 2014. DOI: 10.1007/jhep02(2014)057.

[129] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm," *JHEP*, vol. 04, p. 063, 2008. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].

[130] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual," *The European Physical Journal C*, vol. 72, no. 3, 2012. DOI: 10.1140/epjc/s10052-012-1896-2. [Online]. Available: https://doi.org/10.1140%2Fepjc%2Fs10052-012-1896-2.

[131] L. Vacavant, "b-tagging algorithms and performance in ATLAS," *PoS*, vol. 2008LHC, K. Jacobs, D. Denegri, and I. Puljak, Eds., p. 64, 2008. DOI: 10.22323/1.055.0064.

[132] A. Crivellin *et al.*, *Accumulating evidence for the associate production of a neutral scalar with mass around 151 gev*, 2021. arXiv: 2109.02650 [hep-ph].

[133] J. Albrecht *et al.*, "A roadmap for HEP software and computing r&d for the 2020s," *Computing and Software for Big Science*, vol. 3, no. 1, 2019. DOI: 10.1007/s41781-018-0018-8. [Online]. Available: https://doi.org/10.1007%2Fs41781-018-0018-8.

[134] A. Ghosh, "Deep generative models for fast shower simulation in ATLAS," CERN, Geneva, Tech. Rep., 2020. DOI: 10.1088/1742-6596/1525/1/012077. [Online]. Available: http://cds.cern.ch/record/2680531.

[135] S. Otten *et al.*, "Event generation and statistical sampling for physics with deep generative models and a density information buffer," *Nature Communications*, vol. 12, no. 1, 2021. DOI: 10.1038/s41467-021-22616-z.

[136] J. Eschle, A. Puig Navarro, R. Silva Coutinho, and N. Serra, "Zfit: Scalable pythonic fitting," *SoftwareX*, vol. 11, p. 100 508, 2020, ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2020.100508. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352711019303851.

[137] J. J. Hodges, "The significance probability of the smirnov two-sample test," *Arkiv fiur Matematik*, vol. 3, pp. 469–86, 1958.

[138] "Spearman rank correlation coefficient," in *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, 2008, pp. 502–505, ISBN: 978-0-387-32833-1. DOI: 10.1007/978-0-387-32833-1_379.

[Online]. Available: https://doi.org/10.1007/978-0-387-32833-1_379.

[139] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon *et al.*, Eds., Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[140] F. Stevenson, B. Lieberman, A. Swain, and B. Mellado, "Evaluation and optimisation of a generative-classification hybrid variational autoencoder in the search for resonances at the lhc," *Journal of Physics: Conference Series*, vol. 2586, no. 1, p. 012 160, 2023. DOI: 10.1088/1742-6596/2586/1/012160. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/2586/1/012160.

[141] P. Virtanen *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

[142] A. Paszke *et al.*, "Automatic differentiation in pytorch," 2017.

[143] M. Abadi *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/.

[144] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[145] J. Sun *et al.*, "Covid-19: Epidemiology, evolution, and cross-disciplinary perspectives," *Trends in molecular medicine*, vol. 26, no. 5, pp. 483–495, 2020.

[146] D. Cucinotta and M. Vanelli, "Who declares covid-19 a pandemic," *Acta Biomedica Atenei Parmensis*, vol. 91, no. 1, 157–160, 2020. DOI: 10.23750/abm.v91i1.9397. [Online]. Available: https://www.mattioli1885journals.com/index.php/actabiomedica/article/view/9397.

[147] Z. Zhu, X. Lian, X. Su, W. Wu, G. A. Marraro, and Y. Zeng, "From sars and mers to covid-19: A brief summary and comparison of severe acute respiratory infections caused by three highly pathogenic human coronaviruses," *Respiratory Research*, vol. 21, no. 1, 2020. DOI: 10.1186/s12931-020-01479-w.

[148] C. M. M. A. S. N. R; *Features, evaluation, and treatment of coronavirus (covid-19)*. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32150360/.

[149] B. Mellado *et al.*, "Leveraging artificial intelligence and big data to optimize covid-19 clinical public health and vaccination roll-out strategies in africa," *Available at SSRN 3787748*, 2021.

[150] J. Duhon *et al.*, "The impact of non-pharmaceutical interventions, demographic, social, and climatic factors on the initial growth rate of covid-19: A cross-country study," *Science of The Total Environment*, vol. 760, p. 144 325, 2020.

[151] J. D. Kong, E. Tekwa, and S. Gignoux-Wolfsohn, "Social, economic, and environmental factors influencing the basic reproduction number of covid-19 across countries," *medRxiv*, 2021.

[152] S. A. Government, *South africa corona virus online portal 2020*. [Online]. Available: https://sacoronavirus.co.za/covid-19-risk-adjusted-strategy/.

[153] C. Ramaphosa, *South africa's response to coronavirus covid-19 pandemic*, 2021. [Online]. Available: https://tinyurl.com/2hbrby83.

[154] W. C. Roda, M. B. Varughese, D. Han, and M. Y. Li, "Why is it difficult to accurately predict the covid-19 epidemic?" *Infectious Disease Modelling*, vol. 5, pp. 271–281, 2020, ISSN: 2468-0427. DOI: https://doi.org/10.1016/j.idm.2020.03.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468042720300075.

[155] J. Choma *et al.*, "Worldwide effectiveness of various non- pharmaceutical intervention control strategies on the global covid-19 pandemic: A linearised control model," *medRxiv*, 2020. DOI: 10.1101/2020.04.30.20085316. eprint: https://www.medrxiv.org/content/early/2020/05/12/2020.04.30.20085316.full.pdf. [Online]. Available: https://www.medrxiv.org/content/early/2020/05/12/2020.04.30.20085316.

[156] R. Verity *et al.*, "Estimates of the severity of covid-19 disease," *medRxiv*, 2020. DOI: 10.1101/2020.03.09.20033357. eprint: https://www.medrxiv.org/content/early/2020/03/13/2020.03.09.20033357.full.pdf. [Online]. Available: https://www.medrxiv.org/content/early/2020/03/13/2020.03.09.20033357.

[157] J. Choma *et al.*, "Risk adjusted non-pharmaceutical interventions for the management of covid-19 in south africa," *medRxiv*, 2020. DOI: 10.1101/2020.07.15.20149559. eprint: https://www.medrxiv.org/content/early/2020/07/17/2020.07.15.20149559.full.pdf. [Online]. Available: https://www.medrxiv.org/content/early/2020/07/17/2020.07.15.20149559.

[158] B. Lieberman *et al.*, "Big data- and artificial intelligence-based hotspot analysis of covid-19: Gauteng, south africa, as a case study," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, 2023. DOI: 10.1186/s12911-023-02098-3.

[159] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," 2012. DOI: 10.48550/ARXIV.1201.0490. [Online]. Available: https://arxiv.org/abs/1201.0490.