Development of a Deep Neural Network Double b-tagger for Boosted Topologies using CMS Open Data

A Thesis submitted for the completion of requirements for the degree of

Master of Science

by

Dev Nair
(2023PGPHPH019)
National Institute of Technology
Jamshedpur

Under the supervision of

Prof. Jyothsna Rani Komaragiri
IISc Bangalore

Dr Abhishek Majhi
NIT Jamshedpur

DECLARATION

I, Dev Nair (Registration No. 2023PGPHPH019), hereby declare that the research work presented in this thesis, entitled **"Development of a Deep Neural Network Double b-tagger for Boosted Topologies using CMS Open Data"**, is my own original work conducted under the guidance of **Prof. Jyothsna Rani Komaragiri** (Indian Institute of Science, Bengaluru) and **Dr Abhishek Majhi** (National Institute of Technology, Jamshedpur).

I further declare that this work, in whole or in part, has not been previously submitted for any degree, diploma, or fellowship to this or any other university or institution. All sources have been appropriately acknowledged and referenced. This thesis represents a true record of the research undertaken by me.

Dev Nair
Reg. No. 2023PGPHPH019
Date: May 15, 2025

Department of Physics

National Institute of Technology, Jamshedpur

An Institute of National Importance under Ministry of Education, Govt. of India

Certificate

This is to certify that the thesis entitled **"Development of a Deep Neural Network Double b-tagger for Boosted Topologies using CMS Open Data"**, which is being submitted by **Mr. Dev Nair** with Registration No. **2023PGPHPH019**, to the Department of Physics, National Institute of Technology, Jamshedpur, for the award of degree of **Master of Science in Physics**, is a record of research work carried out by him under our guidance. He has fulfilled the requirements for the submission of thesis, which to our knowledge has reached the requisite standard. The results contained in this dissertation have not been submitted in part or in full to any university or institute for the award of any degree.

Prof. Jyothsna Rani Komaragiri
Associate Professor
IISc Bangalore

Dr Abhishek Majhi
Assistant Professor
NIT Jamshedpur

Date: May 15, 2025

Place: NIT Jamshedpur

Acknowledgements

I would like to express my sincere gratitude to my parents for their constant encouragement and unwavering support throughout my studies. Their belief in me has been invaluable in pursuing and completing this thesis.

I am profoundly grateful to my external supervisor, Prof. Jyothsna Rani Komaragiri from IISc Bangalore. I sincerely thank her for providing me with the invaluable opportunity to work under her guidance and for welcoming me into her research domain. Her expertise, insightful suggestions, and mentorship were instrumental in shaping this project.

My sincere thanks also go to my internal supervisor, Dr Abhishek Majhi from the Department of Physics, NIT Jamshedpur. I appreciate his valuable guidance within the institute's framework, and for taking me on as his student for my final semester.

I would also like to extend my gratitude to all the teachers who have guided me throughout my academic life. Their dedication to imparting knowledge has played a significant role in shaping my understanding and curiosity.

Finally, this research was made possible by the public release of simulation data from the CMS Collaboration through the CERN Open Data Portal. I extend my gratitude to CMS and CERN for their commitment to open science. My work also relied heavily on numerous open-source software packages developed and maintained by the scientific computing community, whose contributions are indispensable to modern research.

Abstract

The precise characterization of the Higgs boson is a cornerstone of the Large Hadron Collider (LHC) physics program. The dominant $H \to b\bar{b}$ decay channel provides a direct probe of the Higgs coupling to bottom quarks, but its identification, especially in boosted topologies crucial for certain production modes and Beyond the Standard Model searches, is challenged by enormous Quantum Chromodynamics (QCD) multijet backgrounds. This thesis presents the development and evaluation of a Deep Neural Network (DNN) designed as a double b-tagger to identify boosted $H \to b\bar{b}$ decays within large-radius (AK8) jets, utilizing publicly available CMS Run 2 simulation data (Record 12102, $\sqrt{s} = 13$ TeV). The DNN leverages a set of 28 engineered high-level features capturing tracking, vertexing, and jet substructure information to classify signal jets against a filtered QCD background.

The optimized DNN tagger achieves excellent discrimination performance on an independent test set, reaching a ROC Area Under the Curve (AUC) of **0.9441** and an Average Precision (AP) of **0.9004**. This significantly surpasses baseline benchmarks (**AUC $\approx$ 0.90**) associated with this dataset; the optimized architecture demonstrated strong performance (**AUC $\approx$ 0.92**) even with a reduced set of 27 features, with the final result reflecting the combined benefit of the network design and the inclusion of the crucial N-subjettiness (τ_{21}) variable. At a working point optimized for balanced precision and recall, the tagger achieves 83.3% signal efficiency with 81.9% precision. Feature importance analysis confirms the synergistic role of substructure, vertexing, and B-hadron lifetime information.

This work successfully demonstrates the effectiveness of applying optimized DNNs to engineered features for a complex particle physics classification task using CMS Open Data, providing a high-performance tagger and a strong foundation for future studies involving validation with collision data and integration into physics analyses.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Standard Model and the Higgs Boson

The Standard Model (SM) of particle physics provides an exceptionally successful description of the known fundamental particles and the forces governing their interactions, excluding gravity[1]. Within this framework, the mechanism proposed by Brout, Englert, Higgs, Guralnik, Hagen, and Kibble explains the origin of mass for elementary particles through their interaction with a scalar field permeating spacetime, known as the Higgs field [2]. The quantum excitation of this field is the Higgs boson, a scalar particle with a mass measured to be $m_H \approx 125$ GeV [1].

The discovery of the Higgs boson at the CERN Large Hadron Collider (LHC) in 2012 by the ATLAS and CMS experiments was a landmark achievement, confirming a key prediction of the SM [3, 4]. Subsequent research has focused intensely on characterizing this particle, measuring its production cross-sections (σ), decay branching ratios (BR), and couplings to other SM particles. Precise measurements of these properties are crucial tests of the SM's validity and serve as powerful indirect searches for physics Beyond the Standard Model (BSM), as deviations could indicate the influence of new, undiscovered phenomena.

1.2 The $H \to b\bar{b}$ Decay Channel in Boosted Topologies

Within the SM, the Higgs boson predominantly decays into the heaviest particle kinematically accessible. For $m_H \approx 125$ GeV, the largest branching ratio is predicted to be into a pair of bottom quarks, $BR(H \to b\bar{b}) \approx 58\%$ [5]. This decay mode offers the most statistically powerful channel to directly probe the Higgs boson's coupling to down-type quarks, a fundamental parameter of the SM.

Experimentally, observing the $H \to b\bar{b}$ decay at a hadron collider like the LHC is extremely challenging. The signal process, characterized by a rate proportional to $\sigma(pp \to H + X) \times BR(H \to b\bar{b})$, is swamped by an enormous background from Quantum Chromodynamics (QCD) multijet production, particularly processes involving gluon splitting $g \to b\bar{b}$ or direct $b\bar{b}$ production. The cross-section for producing jets containing b-quarks via QCD processes is many orders of magnitude larger than the Higgs signal rate.

In scenarios where the Higgs boson is produced with high transverse momentum (p_T) – referred to as **boosted topologies** – its decay products (b and $\bar{b}$) become highly collimated. Standard jet algorithms with small radius parameters (like $R = 0.4$) may fail to resolve the two b-quarks into separate jets. Instead, algorithms using a larger radius parameter (e.g., $R = 0.8$) are employed to reconstruct the $H \to b\bar{b}$ system as a single **fat jet**. Identifying such fat jets as originating from $H \to b\bar{b}$ requires sophisticated techniques, often referred to as **double b-tagging**, that exploit the jet's internal structure and the specific properties of b-hadrons within it. Isolating the signal requires exceptional background rejection capabilities.

1.3 Jet Physics and Double b-Tagging at the LHC

High-energy proton-proton collisions at the LHC produce quarks and gluons which shower and hadronize, forming collimated sprays of particles reconstructed as jets. Fat jets, typically reconstructed using the anti-k_T algorithm with $R = 0.8$ (AK8 jets) [6], are crucial for studying boosted heavy particles.

Identifying the flavour content of these jets is critical. Jets initiated by bottom quarks (b-jets) are distinguishable due to the properties of the B-hadrons they contain. B-hadrons possess a significant lifetime ($\tau_B \sim 1.5$ ps), leading to decay vertices measurably displaced from the primary interaction point (PV) [1]. This results in associated tracks having large impact parameters (d_0) relative to the PV. B-hadron decays also typically have higher track multiplicities and may contain leptons.

Standard b-tagging algorithms exploit these features using track impact parameter significance and secondary vertex (SV) reconstruction. In the context of boosted $H \to b\bar{b}$ decays reconstructed as a single fat jet, **double b-tagging** algorithms aim to identify the presence of *two* distinct b-hadrons within the jet. This often involves analysing the properties of sub-jets found within the fat jet, searching for multiple displaced vertices, or using advanced Machine Learning (ML) techniques that combine numerous input variables related to tracks, vertices, and the jet's substructure (e.g., N-subjettiness τ_{21}) [7].

Modern approaches increasingly utilize Deep Learning (DL) to learn complex correlations from these inputs, achieving substantial performance improvements [8].

1.4 CMS Open Data

The progress in particle physics relies on analysing large, complex datasets. The CMS Collaboration, alongside other LHC experiments and CERN, fosters open science through the CERN Open Data portal [9]. This initiative provides public access to significant fractions of the collision and simulation data collected by the experiments, complete with the necessary software environments and documentation. This open data policy enables data preservation, facilitates independent analyses and reproducibility, supports educational outreach, and empowers researchers globally to contribute to physics analysis and methods development. This thesis utilizes simulation data released through this portal, specifically corresponding to CMS Run 2 conditions at $\sqrt{s} = 13$ TeV.

1.5 Thesis Objective and Scope

The primary objective of this thesis is the **development and evaluation of a Deep Neural Network (DNN) designed as a double b-tagger** to identify boosted Higgs bosons decaying to bottom quark pairs ($H \to b\bar{b}$) within large-radius jets, discriminating them from Quantum Chromodynamics (QCD) multijet backgrounds.

The study leverages Monte Carlo simulation samples from the CMS experiment's Run 2 Open Data release. A specific focus is placed on utilizing a set of **28 high-level, engineered features** ($\mathbf{x} \in \mathbb{R}^{28}$) derived from reconstructed fat jet properties (AK8), encompassing tracking information, secondary vertex characteristics, impact parameters, and jet substructure variables. The DNN learns a mapping $f : \mathbf{x} \mapsto P(y = 1|\mathbf{x})$, where $y = 1$ represents the $H \to b\bar{b}$ class.

The scope is defined by this specific **binary classification task (boosted $H \to b\bar{b}$ vs. filtered QCD)** within the selected kinematic phase space relevant for boosted topologies, using the aforementioned engineered feature set. The performance of the developed DNN tagger is rigorously evaluated using standard classification metrics, achieving high discrimination performance characterized by an area under the ROC curve (AUC) of **0.9441** and an average precision (AP) of **0.9004** on an independent test set. This work serves as a case study in applying optimized DNNs to engineered features for a challenging particle physics classification problem using publicly available data.

1.6 Thesis Outline

The remainder of this thesis is structured as follows:

- **Chapter 2** provides a more detailed overview of the relevant physics background, including the Standard Model, Higgs boson physics, jet reconstruction in boosted topologies, jet substructure, and the principles underlying b-tagging and double b-tagging techniques.
- **Chapter 3** describes the specific CMS Open Data simulation samples used, details the data preprocessing workflow including the kinematic filtering applied, and lists the 28 engineered features selected as inputs for the model.
- **Chapter 4** outlines the machine learning methodology, covering the fundamentals of Deep Neural Networks, the specific architecture implemented (including hyperparameter optimization details), the training procedure, and the metrics used for performance evaluation.
- **Chapter 5** presents the results of the study, including analysis of input features, model training details, comprehensive performance evaluation on the test dataset (AUC, AP, rejection factors, etc.), and potentially additional analyses such as feature importance or performance dependencies.
- **Chapter 6** discusses the interpretation of the results, compares the achieved performance with relevant benchmarks, contextualizes the findings within the broader field of jet tagging, acknowledges the limitations of the study, and suggests potential avenues for future work.
- **Chapter 7** concludes the thesis by summarizing the key findings and contributions of this work.

Chapter 2

Physics Background

This chapter provides the necessary physics context for understanding the identification of Higgs boson decays to bottom quarks ($H \to b\bar{b}$) at the Large Hadron Collider (LHC), particularly within the challenging environment of boosted topologies. We will briefly review relevant aspects of the Standard Model, discuss the production of jets and the specific properties of b-quark jets, describe the dominant background processes, outline the principles underlying b-tagging algorithms, introduce concepts of jet substructure relevant for boosted objects, and finally, discuss the motivation for using high-level features as inputs for the machine learning model developed in this thesis.

2.1 The Standard Model Context

The Standard Model (SM) of particle physics is the theoretical framework describing the fundamental constituents of matter—quarks and leptons—and their interactions via force carriers (gauge bosons) [1]. The particles relevant to this analysis include the six quarks (up, down, charm, strange, top, bottom), the six leptons (electron, muon, tau, and their corresponding neutrinos), the gauge bosons mediating the fundamental forces (photon for electromagnetism, W and Z bosons for the weak force, gluons for the strong force), and the Higgs boson (H).

A key feature of the SM is the Higgs mechanism, responsible for generating the masses of the W and Z bosons and the fundamental fermions through their interaction with the Higgs field. The Higgs boson itself is a scalar particle with a mass measured precisely by the ATLAS and CMS experiments to be $m_H \approx 125$ GeV[1]. Within the SM, the Higgs boson couples to other particles with a strength proportional to their mass. This implies a significant coupling to the heavy bottom quark ($m_b \approx 4.2$ GeV), making the $H \to b\bar{b}$ decay the most probable one ($BR(H \to b\bar{b}) \approx 58\%$) [5]. Studying this decay channel is therefore crucial for probing the Higgs coupling to down-type quarks.

2.2 Jet Production and Reconstruction at the LHC

In high-energy proton-proton (pp) collisions at the LHC, the fundamental interactions occur between the constituent partons (quarks and gluons) within the protons. When high-momentum partons are produced in the hard scattering process, they cannot exist as free coloured objects due to Quantum Chromodynamics (QCD) confinement. Instead, they initiate a complex sequence of events:

- **Parton Showering:** The initial high-energy parton radiates gluons, which can subsequently split into quark-antiquark pairs ($g \to q\bar{q}$) or further gluons ($g \to gg$). Quarks can also radiate gluons ($q \to qg$). This process creates a cascade of lower-energy partons moving in roughly the same direction.

- **Hadronization:** As the energy scale decreases during the shower, the partons combine to form colour-neutral bound states called hadrons (mesons like pions and kaons, and baryons like protons and neutrons).

The result of this cascade is a collimated spray of dozens or even hundreds of detectable particles (mostly hadrons, but also photons and leptons from decays) traveling in approximately the same direction as the initial high-energy parton. Experimentally, these sprays are reconstructed as **jets**.

Jets are typically reconstructed by clustering energy deposits in calorimeter cells or, more commonly at CMS, by clustering reconstructed particles identified by the Particle Flow algorithm [10] using sequential recombination algorithms. The standard algorithm used at the LHC is the **anti-k_T algorithm** [6]. This algorithm takes a radius parameter $R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$, where $\Delta\eta$ and $\Delta\phi$ are the distances in pseudorapidity and azimuthal angle, respectively. Standard analyses often use $R = 0.4$ (AK4 jets).

2.3 Boosted Topologies and Fat Jets

In many physics processes at the LHC, including the production of Higgs bosons, the particle of interest can be produced with very high transverse momentum (p_T). When such a heavy particle (like H, W, Z, or top quark) decays hadronically, its decay products are kinematically forced (”boosted”) into a narrow cone in the detector. The typical angular separation ΔR between the decay products scales approximately as $\Delta R \sim 2m/p_T$, where m is the mass of the decaying particle.

For a Higgs boson with $m_H \approx 125$ GeV, if its p_T is sufficiently high (e.g., $p_T \gtrsim 250 - 300$ GeV), the ΔR between the b and $\bar{b}$ quarks becomes small enough ($\Delta R \lesssim 1.0$) that their subsequent parton showers and hadronization streams significantly overlap.

This transition from a resolved two-jet topology to a merged single-jet topology as a function of the Higgs boson's p_T is illustrated in Fig. 2.1. In such **boosted topologies**, a standard small-radius jet algorithm (like AK4) might reconstruct the decay products as two separate, nearby jets, or might fail to efficiently capture the full system.

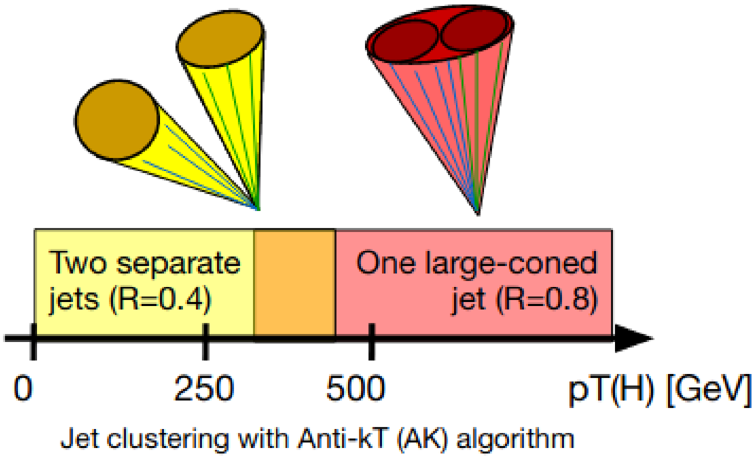
Two separate jets (R=0.4)
One large-coned jet (R=0.8)
0
250
500
pT(H) [GeV]
Jet clustering with Anti-kT (AK) algorithm

Figure 2.1: Illustration of jet clustering for $H \to b\bar{b}$ decays as a function of the Higgs boson transverse momentum ($p_T(H)$). At lower $p_T(H)$, the decay products form two separate jets (e.g., with $R = 0.4$). As $p_T(H)$ increases, the decay products become more collimated, eventually merging into a single large-radius jet (e.g., with $R = 0.8$). (Adapted from [11])

To address this, analyses targeting boosted particles employ **fat jets**, which are reconstructed using the same algorithm (typically anti-k_T) but with a larger radius parameter, commonly **R=0.8 (AK8 jets)** or sometimes R=1.0. An AK8 jet has a much higher probability of containing all the decay products of a boosted heavy particle like the Higgs boson. This thesis focuses on the identification of $H \to b\bar{b}$ decays reconstructed within such AK8 fat jets.

2.4 Properties of b-Quark Jets

Jets originating from the hadronization of a bottom quark (b-jets) possess distinct properties compared to jets from light quarks (u, d, s) or gluons. These differences arise primarily from the characteristics of the B-hadrons (mesons such as B^0, B^+, B_s^0 and baryons such as Λ_b^0) formed during hadronization. Key properties include [1]:

- **Long Lifetime:** B-hadrons have a relatively long average proper lifetime, $\tau_B \approx 1.5$ ps. This corresponds to a proper decay length of $c\tau_B \approx 450$ μm. In the laboratory frame, highly energetic B-hadrons produced in LHC collisions travel a measurable distance before decaying, due to relativistic time dilation ($\gamma = E/m_B$). This lab-frame decay length ($\sim \gamma c\tau_B$) often results in decay vertices displaced by millimeters from the primary pp interaction vertex (PV).

- **High Mass:** The mass of B-hadrons ($m_B \approx 5$ GeV) is significantly larger than that of light hadrons or D-hadrons (containing charm quarks, $m_D \approx 1.9$ GeV). This influences the kinematics of their decay products.

- **Decay Characteristics:** B-hadron decays typically result in multiple charged particles (average charged multiplicity ~ 5). Furthermore, semi-leptonic decays (e.g., $b \to c\ell^-\bar{\nu}_\ell$) occur with a branching ratio of approximately 10–11% for each lepton flavour ($\ell = e, \mu$). These decays produce relatively low-p_T leptons within the jet cone.

These properties, particularly the long lifetime leading to displaced decay vertices and tracks with large impact parameters, are crucial for b-tagging, as illustrated schematically in Fig. 2.2. They form the basis for experimentally identifying, or ”tagging,” b-jets.

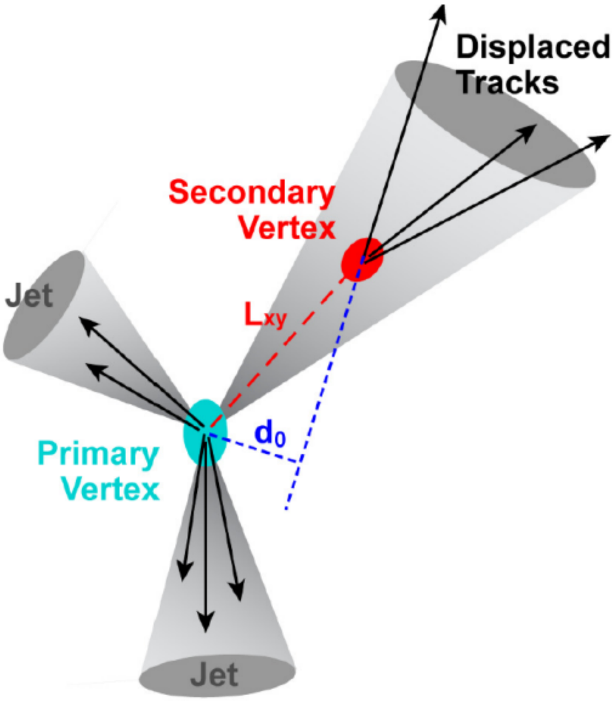

Figure 2.2: Schematic view of a b-jet illustrating key properties for b-tagging: the primary vertex (PV), a secondary vertex (SV) displaced from the PV due to the B-hadron’s lifetime, the transverse decay length (L_{xy}), and the transverse impact parameter (d_0) of a track originating from the SV. (Adapted from [12])

2.5 Principles of b-Tagging and Double b-Tagging

b-Tagging algorithms aim to distinguish b-jets from c-jets (originating from charm quarks) and light-flavour/gluon jets by exploiting the unique properties of B-hadrons. Key experimental signatures leveraged include:

- **Track Impact Parameter (IP):** Charged particles produced in B-hadron decays often originate from a displaced vertex. This results in their reconstructed tracks having large impact parameters (transverse d_0 and longitudinal z_0) relative to the PV. The significance of the impact parameter, typically defined as $S = d_0/\sigma_{d_0}$ where σ_{d_0} is the measurement uncertainty on d_0, is a powerful variable, as tracks from b-decays tend to have large positive S values. Algorithms often combine the significance information from multiple tracks within a jet.

- **Secondary Vertex (SV) Reconstruction:** The displaced decay point of a B-hadron can often be reconstructed as a secondary vertex (SV), separate from the PV. Algorithms search for clusters of tracks consistent with originating from a common displaced point. Properties of reconstructed SVs, such as their flight distance significance (distance from PV divided by its uncertainty), invariant mass, track multiplicity, and energy fraction, serve as strong indicators of a b-jet.

- **Soft Lepton Identification:** The presence of a relatively low-p_T electron or muon within the jet cone, consistent with a semi-leptonic B-decay, provides an additional discriminating handle.

In the context of boosted $H \rightarrow b\bar{b}$ decays captured within a single fat jet (like AK8), the task becomes **double b-tagging**: identifying the simultaneous presence of *two* b-quarks originating from the Higgs decay. This requires analysing the internal structure of the fat jet. Techniques include:

- Identifying two distinct sub-jets within the fat jet and applying standard b-tagging algorithms to each sub-jet.

- Searching for multiple displaced secondary vertices within the fat jet cone.

- Using multivariate algorithms (like DNNs) trained specifically to recognize the pattern of two b-hadron decays within the fat jet, using inputs related to tracks, vertices, and overall jet substructure.

The DNN developed in this thesis falls into the latter category, aiming to perform this double b-tagging task.

2.6 Jet Substructure

The internal structure of jets, particularly fat jets, carries valuable information about the particle that initiated the jet. **Jet substructure** techniques aim to exploit this information [7]. Two concepts relevant to this work are:

- **Jet Grooming:** These algorithms aim to remove contamination from soft, wide-angle radiation associated with the initial hard scatter, as well as contributions from pileup (additional pp interactions occurring in the same bunch crossing). Grooming provides a more stable measurement of the jet's mass and internal structure, connecting it more closely to the hard decay products. **Soft Drop** [13] is a widely used grooming algorithm that recursively removes soft, wide-angle branches from the jet's clustering history. The mass calculated after applying Soft Drop (m_{SD}) is often used in boosted object searches (and was used for the pre-selection cuts in this thesis, Sec. 3).

- **N-subjettiness (τ_N):** This variable quantifies how consistent a jet's energy distribution is with having N distinct sub-jets (or "prongs") compared to $N-1$ sub-jets [14]. The ratio $\tau_{21} = \tau_2/\tau_1$ is particularly useful for discriminating 2-prong decays (like $H \to b\bar{b}$, W$\to q\bar{q}'$, Z$\to q\bar{q}$) from 1-prong QCD jets (initiated by single quarks or gluons). Jets originating from 2-prong decays are expected to have smaller values of τ_{21}. This variable is one of the inputs to the DNN developed here.

2.7 High-Level Features for Machine Learning

While state-of-the-art taggers increasingly use low-level inputs like lists of tracks and calorimeter clusters ("constituents"), a powerful and widely used approach involves defining a set of **high-level or engineered features**. These features are calculated based on the reconstructed objects (jets, tracks, vertices) and are designed to capture the key physical characteristics relevant for discrimination (lifetime information, decay kinematics, substructure patterns) in a fixed-size vector $\mathbf{x} \in \mathbb{R}^N$.

Examples include:

- Counts of objects (e.g., number of tracks, number of secondary vertices).
- Kinematic properties of tracks or vertices (e.g., impact parameter significances, vertex mass, flight distance significance).
- Jet substructure variables (e.g., N-subjettiness ratios like τ_{21}).
- Angular information (e.g., ΔR between objects).

These feature vectors can then be fed into standard machine learning algorithms like Boosted Decision Trees (BDTs) or, as in this thesis, Deep Neural Networks (DNNs). The ML algorithm learns the optimal way to combine these physically motivated features to perform the classification task. The specific set of 28 high-level features used in this work is detailed in Sec. 3.

Chapter 3

Dataset and Simulation

This chapter details the origin and preparation of the dataset used for training and evaluating the Deep Neural Network (DNN) double b-tagger developed in this thesis. It begins with an overview of the Large Hadron Collider (LHC) Run 2 conditions and the Compact Muon Solenoid (CMS) detector, followed by a description of the specific Monte Carlo simulation samples obtained from the CERN Open Data Portal. The core of the chapter describes the multi-step workflow employed to process these samples, define the signal and background categories, filter the data based on kinematic requirements, and select the final set of input features. Finally, the characteristics of the resulting dataset used for the machine learning task are presented.

3.1 LHC Run 2 and the CMS Detector

The data used in this study originates from simulations corresponding to the operating conditions of the Large Hadron Collider (LHC) during its second major run (Run 2), specifically simulating 2016 conditions. During this period, the LHC collided protons (pp) at a center-of-mass energy of $\sqrt{s} = 13$ TeV. The Compact Muon Solenoid (CMS) experiment is a general-purpose particle detector situated at one of the LHC's interaction points [15]. Its design features a high-field superconducting solenoid magnet (3.8 T) encompassing various subdetectors crucial for reconstructing the products of pp collisions. These include a high-granularity silicon tracker for precise charged particle trajectory and vertex reconstruction, electromagnetic and hadronic calorimeters for energy measurements, and extensive muon chambers. The Particle Flow (PF) algorithm integrates information from all subsystems to provide a comprehensive reconstruction of individual particles (photons, electrons, muons, charged and neutral hadrons) within each event [10], which are then used as inputs for jet clustering.

3.2 Monte Carlo Simulation Samples

This analysis relies exclusively on Monte Carlo (MC) simulated datasets obtained from the CERN Open Data Portal [9]. These datasets simulate pp collisions at $\sqrt{s} = 13$ TeV and the subsequent response of the CMS detector, providing the necessary ground truth information required for supervised machine learning studies.

The specific dataset record used is titled "Sample with jet, track and secondary vertex properties for Hbb tagging ML studies" (Record 12102) [16]. This record contains NTuples derived from the **RunIISummer16MiniAODv2** simulation campaign, corresponding to 2016 data-taking conditions (processed with CMSSW 80X). According to the record's description page, it includes jets identified as originating from Higgs bosons decaying to $b\bar{b}$ ($H \to b\bar{b}$) for the signal component, and jets from QCD multijet processes for the background component.

- **Signal ($H \to b\bar{b}$):** The record description page lists Beyond the Standard Model `/BulkGravTohhTohbbhbb...` samples (generated using MadGraph [17]) as source datasets containing $H \to b\bar{b}$ decays.
- **Background (QCD):** The background component is confirmed to originate from QCD multijet events simulated across various transverse momentum (p_T) bins using the `PYTHIA8` generator [18]. The specific source datasets listed on the record page correspond to entries such as `QCD_Pt_XXXtoYYY_TuneCUETP8M1_13TeV_pythia8`.

The NTuples contain a pre-processed selection of jets stored in a flat TTree structure named `deepntuplizer/tree`. The variables relevant for this analysis, including high-level jet features prefixed with `fj_` (indicating fat jets, specifically AK8 jets as discussed in Sec. 2) and truth information (such as flags identifying jets matched to $H \to b\bar{b}$ decays or QCD processes), are available within this tree for training and evaluation.

3.3 Data Preparation Workflow

A dedicated workflow was implemented using Python, leveraging libraries such as `uproot`, `awkward-array`, `pyarrow`, and `pandas`, to transform the data from the downloaded ROOT NTuples into the final format used for DNN training and evaluation. The key steps are outlined below.

3.3.1 ROOT to Parquet Conversion

The initial step involved converting the data stored in the `deepntuplizer/tree` TTree within the downloaded `.root` files into the Apache Parquet format [19]. The `uproot` library [20] was used to read the ROOT files efficiently, and the data for relevant branches was converted into Apache Arrow arrays, which were then written to Parquet files using `pyarrow`. This conversion facilitates faster data loading and manipulation in the subsequent Python-based processing steps due to Parquet's columnar storage structure and potential for type optimization.

3.3.2 Signal, Background Definition and Filtering

The Parquet files generated in the previous step were processed individually using a script based on the `pandas` library. This script performed the following crucial steps to define the analysis dataset:

1. **Label Definition:** Intermediate boolean flags were created based on Monte Carlo truth variables present in the NTuples:
   - Signal jets (`isHbb`) were defined by requiring the jet to be matched to a Higgs boson (`fj_isH == 1`) and to contain two b-quarks (`fj_isBB == 1`).
   - Background jets (`isQCD`) were defined by requiring the jet to originate from a QCD process (`fj_isQCD == 1`) and confirming that the source sample was indeed QCD (using flags like `sample_isQCD == 1` if available, or implicitly by processing only QCD source files for background).
2. **Kinematic Cuts:** Jets were required to satisfy kinematic selections relevant for boosted Higgs analyses, targeting a specific phase space:

   $$40 < \texttt{fj\_sdmass} < 200 \text{ GeV}, \quad 300 < \texttt{fj\_pt} < 2000 \text{ GeV}.$$

   Here, `fj_sdmass` refers to the Soft Drop groomed mass of the AK8 jet, and `fj_pt` is its transverse momentum. As discussed in Sec. 2, the p_T cut selects the boosted regime, while the mass cut defines a window around the expected Higgs mass ($m_H \approx 125$ GeV), albeit a relatively wide one for this pre-selection stage.
3. **Binary Task Filter:** Only jets satisfying exactly one of the above signal or background conditions were kept, by applying the filter

   $$\texttt{isHbb} + \texttt{isQCD} == 1.$$

This step ensures a clean dataset containing only jets unambiguously belonging either to the target signal ($H \to b\bar{b}$) or the defined QCD background category for this specific study.

4. **Final Label Assignment:** A final binary column `label` was created, where a value of 1 denotes signal jets (where `isHbb` is 1) and 0 denotes background jets (where `isQCD` is 1).

5. **Feature Selection & NaN/Inf Handling:** Only the predefined input features (listed in Sec. 3.4) and the final `label` were retained. Any remaining missing (`NaN`) or infinite values in the feature columns, resulting from reconstruction failures or undefined quantities, were handled by imputing with the median value of the respective feature column calculated from the training dataset.

The fully processed and filtered DataFrames were saved as new Parquet files, which were then concatenated to form the final dataset used for training and evaluation.

3.4 Input Features

The DNN model developed in this thesis utilizes a set of $N = 28$ high-level, engineered features derived from the AK8 fat jet properties as available in the processed dataset from Record 12102. These features capture information related to tracking, secondary vertices, track impact parameters, and jet substructure. The specific features used are:

- `fj_jetNTracks`: Number of tracks associated with the jet.
- `fj_nSV`: Number of reconstructed secondary vertices within the jet.
- `fj_tau0_trackEtaRel_0`, _1, _2: Relative pseudorapidity (η_{rel}) of the leading three tracks associated with the leading secondary vertex candidate's subjet axis.
- `fj_tau1_trackEtaRel_0`, _1, _2: Relative pseudorapidity (η_{rel}) of the leading three tracks associated with the sub-leading secondary vertex candidate's subjet axis.
- `fj_tau_flightDistance2dSig_0`, _1: 2D flight distance significance of the leading and sub-leading secondary vertices (SVs).
- `fj_tau_vertexDeltaR_0`: ΔR between the leading SV direction and the jet axis.
- `fj_tau_vertexEnergyRatio_0`, _1: Energy fraction carried by tracks associated with the leading and sub-leading SVs.

- `fj_tau_vertexMass_0`, `_1`: Invariant mass of the tracks associated with the leading and sub-leading SVs.
- `fj_trackSip2dSigAboveBottom_0`, `_1`: 2D impact parameter significance of the leading two tracks, computed relative to a B-hadron lifetime hypothesis.
- `fj_trackSip2dSigAboveCharm_0`: 2D impact parameter significance of the leading track, computed relative to a C-hadron lifetime hypothesis.
- `fj_trackSipdSig_0`, `_1`, `_2`, `_3`: 3D impact parameter significance (IP/σ_{IP}) of the 1st through 4th leading p_T tracks associated with the jet.
- `fj_trackSipdSig_0_0`, `_0_1`: Components related to the 3D impact parameter significance calculation for the leading track.
- `fj_trackSipdSig_1_0`, `_1_1`: Components related to the 3D impact parameter significance calculation for the second leading track.
- `fj_z_ratio`: A momentum-sharing variable related to subjet kinematics, potentially sensitive to the symmetric splitting in $H \to b\bar{b}$.
- `fj_tau21`: N-subjettiness ratio τ_2/τ_1, sensitive to the 2-prong versus 1-prong structure of the jet.

These 28 features constitute the input vector $\mathbf{x}$ for the DNN model. Note that the jet p_T and Soft Drop mass, while used for the initial kinematic selection, were not included as input features to the DNN itself.

3.5 Final Dataset Characteristics

After applying the full data preparation workflow described above to the selected ROOT files from the CMS Open Data Record 12102, the final combined dataset used for training, validation, and testing contained a total of 629,619 jets.

The class distribution in the final dataset is as follows:

- **Signal ($H \to b\bar{b}$, label=1):** 209,305 jets ($\approx$ 33.2%).
- **Background (Filtered QCD, label=0):** 420,314 jets ($\approx$ 66.8%).

This dataset exhibits a moderate class imbalance, with approximately twice as many background jets as signal jets. The dataset was subsequently split using stratified sampling into training (64%), validation (16%), and test (20%) subsets, preserving the class proportions in each. The specific number of jets in each subset is detailed in Sec. 4 and Sec. 5.

Chapter 4

Methodology

This chapter details the machine learning methodology employed in this thesis to develop the double b-tagger, discriminating between jets originating from Higgs boson decays to bottom quarks ($H \to b\bar{b}$) and those from Quantum Chromodynamics (QCD) multijet processes. It begins with fundamental concepts of supervised learning and Deep Neural Networks (DNNs), describes the specific components and architecture of the DNN model used, outlines the hyperparameter optimization process, details the training procedure including the choice of loss function and optimizer, and finally defines the metrics used for evaluating the model's performance.

4.1 Machine Learning Fundamentals

The task addressed in this thesis falls under the category of **supervised learning**, specifically **binary classification**. Given a dataset of input examples, each characterized by a set of features $\mathbf{x} \in \mathbb{R}^N$ (where $N = 28$ is the number of high-level features described in Sec. 3) and associated with a known true label $y \in \{0, 1\}$ (where $y = 1$ represents the signal class, $H \to b\bar{b}$, and $y = 0$ represents the background class, filtered QCD), the goal is to train a model f that learns a mapping $f : \mathbf{x} \mapsto \hat{y}$. Here, $\hat{y}$ is typically the predicted probability $P(y = 1|\mathbf{x})$ that the input jet belongs to the signal class. The model should generalize well, meaning it should accurately classify previously unseen examples from a test dataset drawn from the same underlying distribution.

4.2 Deep Neural Networks (DNNs)

Deep Neural Networks are powerful function approximators inspired by the structure of biological neural networks, capable of learning complex patterns in data. The specific type used in this work is a feedforward DNN, also known as a Multi-Layer Perceptron (MLP). These networks consist of interconnected layers of artificial neurons (or nodes).

A typical neuron computes a weighted sum of its inputs from the previous layer, adds a bias term, and then applies a non-linear **activation function** $\sigma(\cdot)$ to the result. For a neuron receiving inputs $\mathbf{z}_{in}$ with weights $\mathbf{w}$ and bias b, the output is $z_{out} = \sigma(\mathbf{w} \cdot \mathbf{z}_{in} + b)$. Stacking layers of these neurons allows the network to learn increasingly complex hierarchical representations of the input data.

The choice of activation function introduces non-linearity, enabling the network to learn relationships beyond simple linear combinations. This work primarily utilizes the Sigmoid-weighted Linear Unit (SiLU), also known as Swish [21], defined as:

$$\text{SiLU}(x) = x \cdot \text{sigmoid}(x) = \frac{x}{1 + e^{-x}} \tag{4.1}$$

SiLU has been shown to perform well in deep networks, often outperforming traditional functions like ReLU in certain tasks. The final output layer uses the standard sigmoid (logistic) function, $\sigma(x) = 1/(1 + e^{-x})$, to produce a probability estimate $P(y = 1|\mathbf{x})$ bounded between 0 and 1.

4.3 DNN Components Used

The implemented DNN architecture incorporates several standard components to improve training stability, generalization, and performance:

- **Batch Normalization:** Applied after dense layers (before activation), Batch Normalization [22] normalizes the activations within each mini-batch during training. This helps to stabilize the training process by reducing internal covariate shift, allows for potentially higher learning rates, reduces sensitivity to weight initialization, and can have a regularizing effect.

- **Dropout:** Dropout [23] is a regularization technique used to prevent overfitting. During training, it randomly sets a fraction of neuron outputs in a layer to zero for each training example (based on the dropout rate). This forces the network to learn more robust representations that do not overly rely on any single neuron or feature pathway. The specific dropout rates used in this model were determined via hyperparameter optimization (Sec. 4.4).

- **L2 Regularization (Weight Decay):** Added as a kernel regularizer to the Dense layers, L2 regularization penalizes large weight values by adding a term proportional to the squared magnitude of the weights to the loss function ($\lambda\|\mathbf{w}\|_2^2$). This encourages smaller weights, leading to simpler models that often generalize better. The regularization strength (λ) was also optimized.

- **Residual Connections:** Inspired by ResNet architectures [24], a residual (or skip) connection was implemented within the network. The output of one block of layers is added element-wise to the output of a subsequent block before the final activation of that combined block. This allows gradients to propagate more easily through deeper networks, mitigating the vanishing gradient problem and facilitating the training of deeper architectures by allowing the network to easily learn identity mappings if needed.

4.4 Hyperparameter Optimization with Keras Tuner

Finding optimal hyperparameters for a Deep Neural Network (DNN) is crucial for achieving the best possible performance. Manual tuning can be time-consuming and inefficient. Therefore, this work utilized the **Keras Tuner** library [25], a framework specifically designed for optimizing hyperparameters of Keras models. Keras Tuner provides several search algorithms to explore the defined hyperparameter space and identify combinations that maximize (or minimize) a chosen objective metric evaluated on a validation dataset.

In this thesis, Keras Tuner, using the **Hyperband** algorithm (`kt.Hyperband`), was employed to optimize key hyperparameters of the DNN architecture and the training process. The search space included:

- Number of units in the first, third, and fourth dense blocks (denoted as `units_1`, `units_3`, `units_4`).
- Dropout rates following each of the four main dense layers (`dropout_1`, `dropout_2`, `dropout_3`, `dropout_4`).
- L2 regularization strength (`l2_reg`) applied to the dense layers.
- Learning rate (`lr`) for the Adam optimizer.

The Hyperband tuner efficiently searched this space by adaptively allocating resources, training promising configurations for more epochs while quickly discarding less promising ones. Each trial involved building a model with a specific hyperparameter combination, training it for a variable number of epochs (up to a maximum of 50 in this case) using early stopping within the trial (`patience=8`) based on validation performance, and evaluating it on the validation set. The objective function maximized during the optimization process was the Area Under the ROC Curve (`val_auc`) calculated on the validation set. The set of hyperparameters corresponding to the trial that yielded the best validation AUC (specifically: `units_1=320`, `units_3=192`, `units_4=96`, `dropout_1=0.35`, `dropout_2=0.40`, `dropout_3=0.25`, `dropout_4=0.25`, `l2_reg=1e-6`, `lr=0.000607`) was selected and used for the final model architecture and training detailed in Sec. 4.5 and Sec. 4.6.

4.5 Implemented Model Architecture

The final DNN architecture, incorporating the components described in Sec. 4.3 and using the optimal hyperparameters selected via Keras Tuner (as detailed in Sec. 4.4), was implemented using the Keras API [26] within TensorFlow [27]. It takes the $N = 28$ input features (Sec. 3.4) and processes them through a series of blocks, as illustrated schematically in Fig. 4.1.

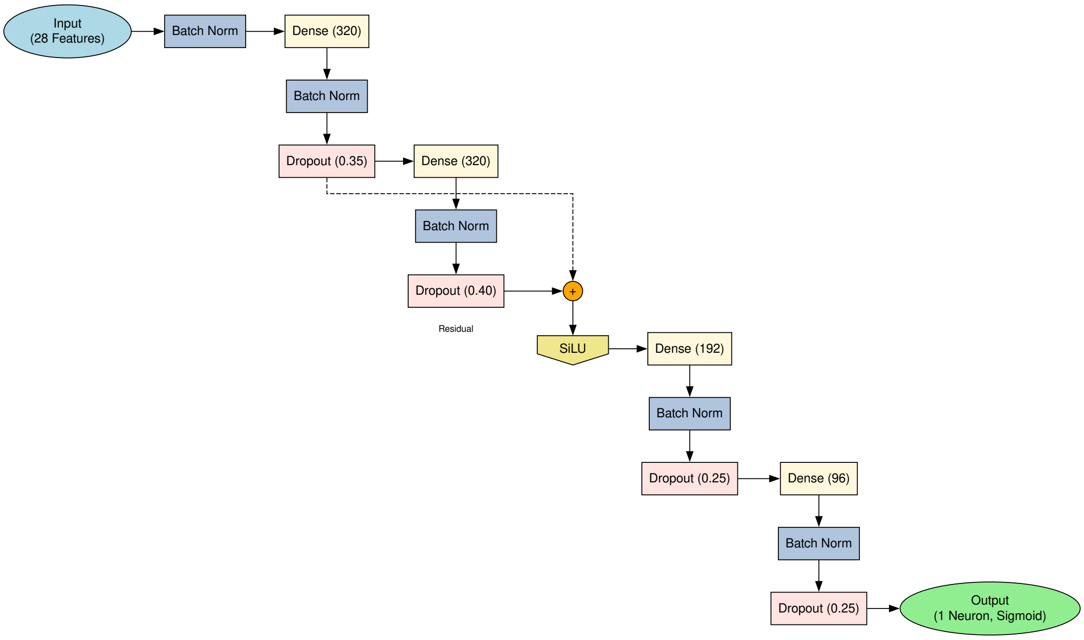

Figure 4.1: Schematic diagram of the implemented Deep Neural Network (DNN) architecture. It shows the input layer, batch normalization, dense layers with SiLU activation, dropout layers, the residual connection adding outputs from Block 1 and Block 2, and the final sigmoid output layer.

The architecture can be summarized as follows (referencing optimized hyperparameters `HP_*`):

1. Input Layer (shape: 28 features)
2. Batch Normalization (applied directly to inputs)
3. **Block 1:**

   - Dense Layer (`HP_UNITS_1` $=$ 320 neurons, L2 reg $=$ `HP_L2_REG` $= 10^{-6}$, SiLU activation)
   - Batch Normalization
   - Dropout (rate $=$ `HP_DROPOUT_1` $= 0.35$)

4. **Block 2:**
   - Dense Layer (`HP_UNITS_1` $=$ 320 neurons, L2 reg $= 10^{-6}$, SiLU activation)
   - Batch Normalization
   - Dropout (rate $=$ `HP_DROPOUT_2` $= 0.40$)
5. **Residual Connection:** Output of Block 1 is added element-wise to the output of Block 2.
6. Activation (SiLU) applied to the sum from the residual connection.
7. **Block 3:**
   - Dense Layer (`HP_UNITS_3` $=$ 192 neurons, L2 reg $= 10^{-6}$, SiLU activation)
   - Batch Normalization
   - Dropout (rate $=$ `HP_DROPOUT_3` $= 0.25$)
8. **Block 4:**
   - Dense Layer (`HP_UNITS_4` $=$ 96 neurons, L2 reg $= 10^{-6}$, SiLU activation)
   - Batch Normalization
   - Dropout (rate $=$ `HP_DROPOUT_4` $= 0.25$)
9. **Output Layer:** Dense Layer (1 neuron, Sigmoid activation)

This architecture, determined through optimization, provides a specific balance between network capacity and regularization tailored to this dataset and feature set.

4.6 Training Procedure

The final DNN model with the optimized architecture was trained using the processed dataset described in Sec. 3. Key aspects of the training procedure include:

- **Data Split:** The final dataset (containing 629,619 jets, see Sec. 3.5) was split into training (64%), validation (16%), and test (20%) sets using stratified sampling to preserve the approximate background-to-signal ratio in each subset. The training set is used to update model weights, the validation set is used for monitoring training progress and tuning the final classification threshold (Sec. 5.4), and the test set is held out for final unbiased performance evaluation.

- **Feature Scaling:** Before training, the 28 input features in the training set were scaled using `sklearn.preprocessing.StandardScaler` (removing the mean and scaling to unit variance). The *same* fitted scaler object was then applied to the validation and test sets to ensure consistent scaling based only on training set information.

- **Loss Function:** Preliminary data exploration revealed that the class imbalance between signal ($H \to b\bar{b}$) and background (QCD) could vary significantly (from approximately 1:2 up to 1:9) depending on the specific kinematic selections applied (e.g., different jet p_T or mass ranges). Although the final dataset used for training exhibited a moderate imbalance (roughly 1:2 signal-to-background, see Sec. 3.5), the potential for larger imbalances motivated the choice of the **Focal Loss** [28]. This loss function is designed to be robust against class imbalance by down-weighting easy-to-classify examples (typically the abundant background class) and focusing the training effort on harder-to-classify examples. The Focal Loss is defined as:

$$L_{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{4.2}$$

  where $p_t = p$ if $y = 1$ and $p_t = 1 - p$ if $y = 0$, with $p = P(y = 1|\mathbf{x})$ being the model's predicted probability for the positive class. The parameters α_t (balancing factor, α for class 1, $1 - \alpha$ for class 0) and γ (focusing parameter) control the weighting. Standard values of $\alpha = 0.5$ and $\gamma = 2.0$ were used in this work, providing a balance between addressing imbalance and focusing on difficult examples.

- **Optimizer:** The **Adam optimizer** [29] was used for updating the model weights, utilizing the learning rate which was found during hyperparameter optimization (`HP_LEARNING_RATE` = 0.000607) and default β_1, β_2 parameters. Adam is an adaptive learning rate optimization algorithm well-suited for training deep networks.

- **Training Parameters:** The model was trained for a maximum of **100 epochs** with a **batch size of 1024**.

- **Callbacks:** Several Keras callbacks were used during training to manage the process:

  - `EarlyStopping`: Monitored the validation AUC (`val_auc`) and stopped training if it did not improve for 15 consecutive epochs (`patience=15`), restoring the weights from the epoch with the best `val_auc` (`restore_best_weights=True`).

- `ReduceLROnPlateau`: Monitored `val_auc` and reduced the learning rate by a factor of 0.5 if no improvement was seen for 5 epochs (`patience=5`), down to a minimum learning rate of 10^{-6}.
- `ModelCheckpoint`: Saved the model weights corresponding to the best `val_auc` observed during training to a file (`hbb_tagger.keras`). This ensures the best performing model state is preserved independently of early stopping.

4.7 Evaluation Metrics

To assess the performance of the trained DNN double b-tagger on the independent test set, several standard metrics for binary classification are used:

- **Confusion Matrix:** A table summarizing the counts of True Positives (TP, correctly identified signal), True Negatives (TN, correctly identified background), False Positives (FP, background misidentified as signal, Type I error), and False Negatives (FN, signal misidentified as background, Type II error). Calculated at specific operating points (thresholds).
- **Accuracy:** The overall fraction of correct predictions: $\text{Acc} = (TP + TN)/(TP + TN + FP + FN)$. Can be misleading for imbalanced datasets.
- **Precision (Purity):** The fraction of positive predictions that are actually correct: $P = TP/(TP + FP)$.
- **Recall (Sensitivity, True Positive Rate, TPR, Signal Efficiency ϵ_S):** The fraction of actual positive instances that are correctly identified: $R = TPR = \epsilon_S = TP/(TP + FN)$.
- **F1 Score:** The harmonic mean of Precision and Recall, providing a single metric that balances both: $F1 = 2 \times (P \times R)/(P + R)$. The optimal classification threshold is determined by maximizing this metric on the validation set.
- **False Positive Rate (FPR, Mistag Rate ϵ_B):** The fraction of actual negative instances that are incorrectly identified as positive: $FPR = \epsilon_B = FP/(FP + TN)$. Note: Sometimes referred to as background efficiency.
- **Receiver Operating Characteristic (ROC) Curve:** A plot of TPR (Recall, ϵ_S) versus FPR (ϵ_B) at various classification thresholds.

- **Area Under the ROC Curve (AUC or ROC AUC):** A scalar value representing the overall discrimination ability of the model across all thresholds. AUC = 1 indicates a perfect classifier, while AUC = 0.5 indicates performance no better than random guessing.

- **Precision-Recall (PR) Curve:** A plot of Precision versus Recall (TPR) at various thresholds. Particularly informative for imbalanced datasets where the baseline (random guessing) is not 0.5.

- **Average Precision (AP or PR AUC):** The area under the PR curve, computed as a weighted mean of precisions achieved at each threshold, summarizing performance with a focus on positive class identification.

- **Background Rejection:** Defined as $1/\epsilon_B$ (or $1/FPR$), representing how effectively the background is suppressed. It is often evaluated at specific fixed values of signal efficiency (ϵ_S).

These metrics provide a comprehensive evaluation of the tagger's performance, presented in Sec. 5.

Chapter 5

Results

This chapter presents the results obtained from training and evaluating the Deep Neural Network (DNN) double b-tagger developed in this thesis. The primary goal was to discriminate boosted Higgs boson decays to bottom quark pairs ($H \to b\bar{b}$) from QCD multijet backgrounds using a set of 28 engineered features derived from CMS Open Data simulations. We will analyse the input features, detail the model's training process and convergence, evaluate its performance on an independent test set using various metrics, investigate the importance of individual input features, and discuss the characteristics of the most and least discriminating variables.

5.1 Dataset Overview

As described in Sec. 3, the dataset was derived from CMS Run 2 Open Data simulation samples (Record 12102) processed to select AK8 jets satisfying kinematic requirements relevant for boosted $H \to b\bar{b}$ searches ($40 < m_{\mathrm{SD}} < 200$ GeV, $300 < p_{\mathrm{T}} < 2000$ GeV).

The final dataset used for this study consists of 629,619 jets, with the following class distribution:

- **Signal ($H \to b\bar{b}$, label=1):** 209,305 jets (33.2%).
- **Background (Filtered QCD, label=0):** 420,314 jets (66.8%).

This dataset was split using stratified sampling into:

- Training set: 402,956 jets (64%)
- Validation set: 100,739 jets (16%)
- Test set: 125,924 jets (20%)

The validation set was used for hyperparameter optimization (Sec. 4) and determining the optimal classification threshold, while the test set was held out for the final unbiased performance evaluation presented in this chapter.

5.2 Input Feature Analysis

Before training the model, the correlations between the 28 input features listed in Sec. 3.4 were examined. The correlation matrix, calculated on the full dataset before splitting, is shown in Fig. 5.1. The matrix reveals varying degrees of correlation between features.

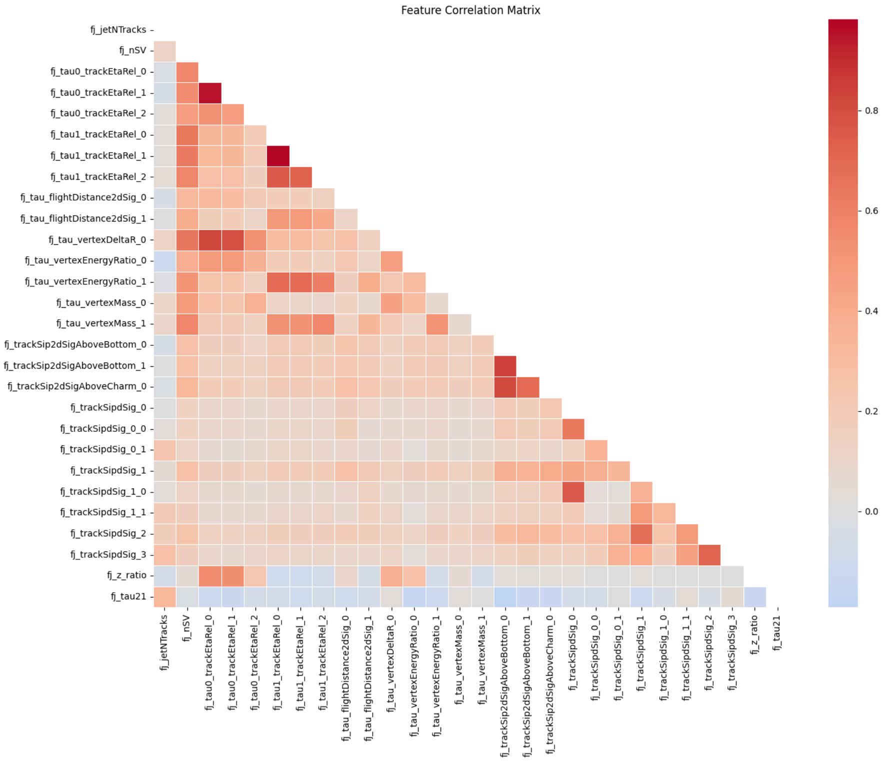

Figure 5.1: Correlation matrix for the 28 input features used in the DNN tagger. The colour scale indicates the Pearson correlation coefficient.

Some groups of features, such as those related to the properties of the same secondary vertex (e.g., fj_tau_vertexMass_0, fj_tau_vertexEnergyRatio_0) or impact parameters of the same track (e.g., fj_trackSipdSig_0, fj_trackSipdSig_0_0, fj_trackSipdSig_0_1), exhibit expected moderate to high correlations. However, many features show low correlation with each other, suggesting they provide complementary information. The DNN architecture is well-suited to handle such correlated inputs and learn the optimal way to combine them. Features with very high correlation (close to ±1) might indicate redundancy, but no strong redundancies demanding feature removal were observed at this stage.

5.3 Model Training and Convergence

The DNN model, with the architecture specified in Sec. 4.5, was trained using the procedure detailed in Sec. 4.6 for a total of 100 epochs. The training process was monitored using the validation set, employing callbacks for learning rate reduction and saving the best model based on validation AUC. Early stopping based on validation AUC (patience=15) was configured but did not trigger before the completion of 100 epochs, indicating continuous (though potentially small) improvements or stability in validation performance towards the end of training. The model weights corresponding to the epoch with the highest validation AUC were saved and used for the final evaluation.

The training history, showing the evolution of the loss function (Focal Loss), Area Under the ROC Curve (AUC), and F1 score over the training epochs for both the training and validation sets, is presented in Fig. 5.2. The plots demonstrate stable training convergence over the 100 epochs. The loss decreases steadily for both training and validation sets, while the AUC and F1 score increase, indicating successful learning. The validation curves track the training curves closely throughout the training process, suggesting that the regularization techniques employed (Dropout, L2 regularization, Batch Normalization) were effective in preventing significant overfitting even during extended training.

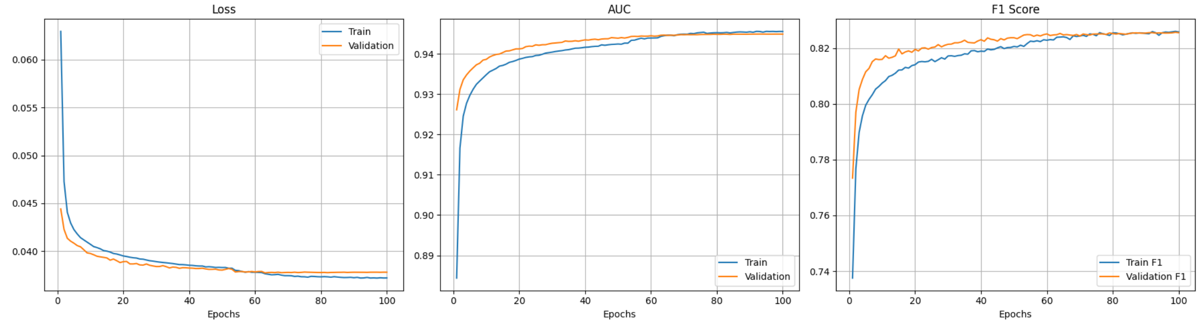

Figure 5.2: Training history of the DNN model over 100 epochs: Loss (left), AUC (center), and F1 Score (right) as a function of training epochs for the training set (blue) and validation set (orange).

5.4 Threshold Optimization

The output of the DNN is a continuous score between 0 and 1, representing the predicted probability of the jet being signal ($H \to b\bar{b}$). To make a binary classification decision, a threshold must be applied to this score. The optimal threshold was determined by maximizing the F1 score on the validation set predictions, balancing precision and recall. Fig. 5.3 illustrates how various metrics (Accuracy, Precision, Recall, F1

Score) vary with the chosen threshold on the validation set. The maximum F1 score on the validation set (0.8279) was achieved at a threshold of 0.4800. This value, saved to `hbb_tagger_optimal_threshold.txt`, is used as the primary operating point for evaluating the tagger's performance on the test set in subsequent sections.

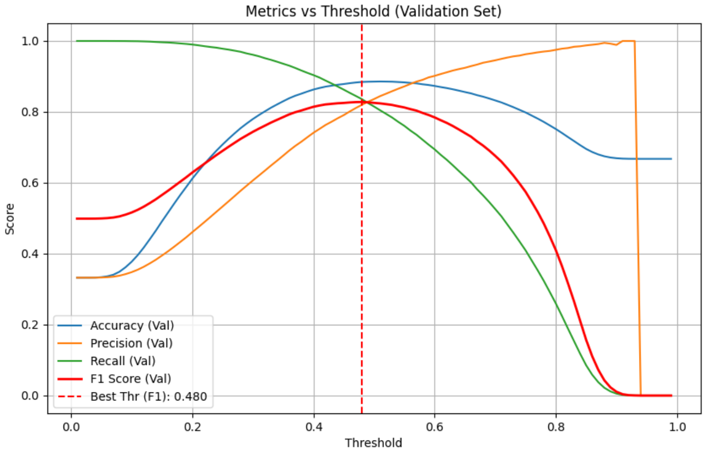

Figure 5.3: Performance metrics (Accuracy, Precision, Recall, F1 Score) on the validation set as a function of the classification threshold applied to the DNN output score. The vertical dashed line indicates the optimal threshold chosen to maximize the F1 score.

5.5 Overall Performance Evaluation on Test Set

The final performance of the trained DNN tagger was evaluated on the independent test set (125,924 jets), which was not used during training or threshold optimization.

5.5.1 ROC and Precision-Recall Curves

The Receiver Operating Characteristic (ROC) curve and the Precision-Recall (PR) curve provide comprehensive views of the tagger's discrimination performance across all possible operating points. These curves for the test set are shown in Fig. 5.4. The Area Under the ROC Curve (AUC) quantifies the overall ability of the model to distinguish between signal and background jets. The calculated AUC on the test set using all 28 features is **0.9441**. This high value, close to 1, confirms the excellent discrimination power of the DNN tagger.

The Area Under the PR Curve, also known as Average Precision (AP), summarizes the trade-off between precision and recall. The calculated AP on the test set is **0.9004**. This strong AP score indicates high performance, especially relevant given the class imbalance (approx. 1:2 signal:background) in the dataset.

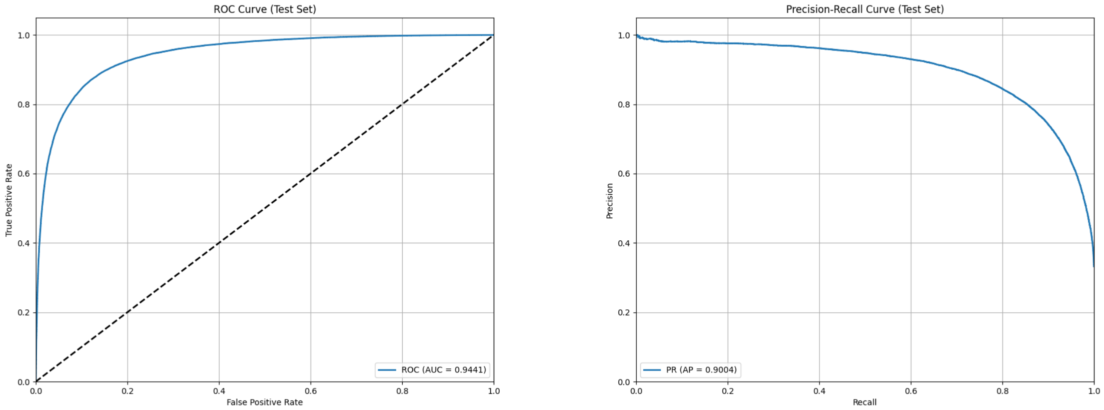

Figure 5.4: Performance curves on the test set. Left: Receiver Operating Characteristic (ROC) curve plotting True Positive Rate (Signal Efficiency) vs. False Positive Rate (Background Mistag Rate). Right: Precision-Recall (PR) curve.

5.5.2 Discriminator Output Distribution

Fig. 5.5 presents the distribution of the final DNN output score for signal and background jets in the independent test set, providing a visual representation of the tagger's separation power. The figure displays the distributions side-by-side using both a linear y-axis scale (left panel) and a logarithmic y-axis scale (right panel), with all distributions normalized to unit area. The linear scale view highlights the shape and peak locations of the signal ($H \to b\bar{b}$, orange) and background (QCD, blue) components, while the logarithmic scale enhances the visibility of the separation in the low-statistics tails, which is crucial for evaluating performance in high-purity or high-rejection regions.

A clear separation between the classes is evident, with signal jets predominantly receiving scores close to 1 and background jets concentrated near 0. The vertical dashed line indicates the optimal classification threshold (0.4800), selected to maximize the F1 score on the validation set. The text box overlaid on the linear plot quantifies the performance at this specific operating point on the test set, showing high Signal Efficiency (Recall: 83.30%), Precision (81.85%), F1 Score (0.8257), and Background Rejection (True Negative Rate: 90.80%). This visualization confirms the strong discriminating capability of the developed DNN tagger, complementing the integrated performance metrics like AUC and Average Precision discussed previously.

5.5.3 Performance Metrics at Specific Thresholds

To understand the tagger's performance at specific operating points, we evaluate standard classification metrics on the test set using both the optimal threshold determined from the validation set (0.4800) and the default threshold of 0.5.

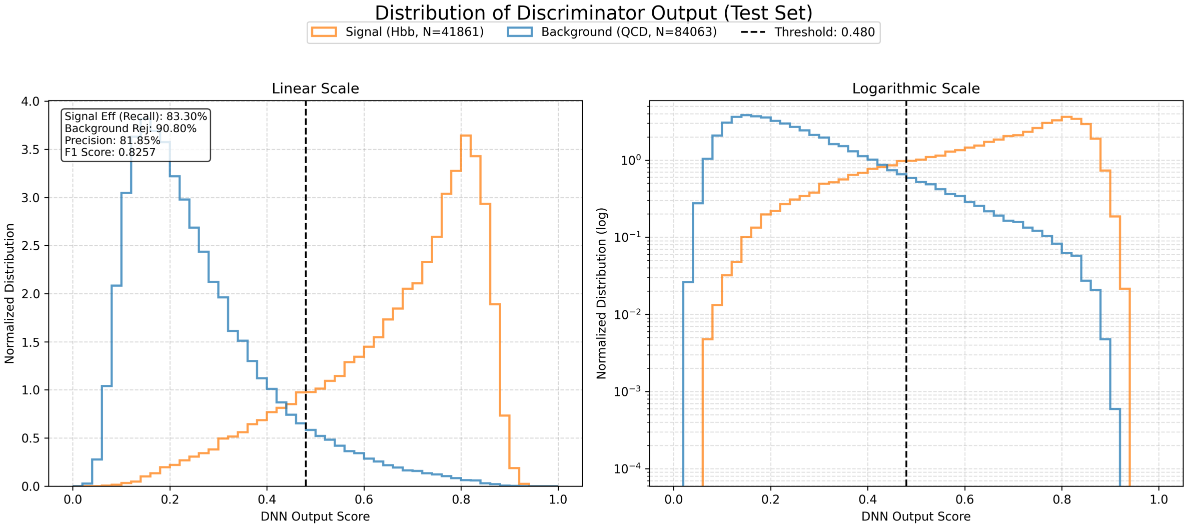

Figure 5.5: Distribution of the DNN output score for signal ($H \to b\bar{b}$, orange) and background (QCD, blue) jets in the test set, normalized to unit area. The left panel displays the distribution with a linear y-axis scale, while the right panel uses a logarithmic scale. The vertical dashed line indicates the optimal classification threshold (0.4800). The info box on the linear plot shows performance metrics at this threshold.

The confusion matrices for these two thresholds are shown in Fig. 5.6. Table 5.1 summarizes the key performance metrics calculated on the test set for both the optimal and default thresholds.

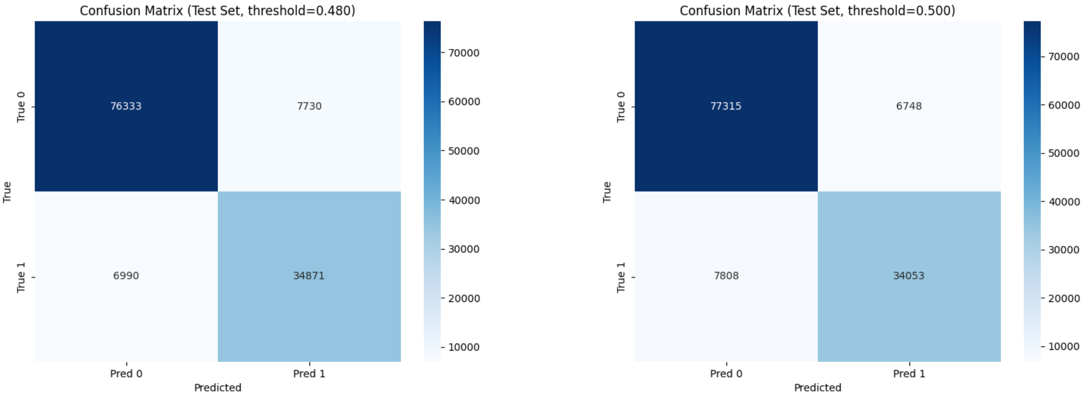

Figure 5.6: Confusion matrices on the test set using the optimal threshold (0.4800, left) and the default threshold (0.5, right).

At the optimal threshold of 0.4800, the tagger achieves a signal efficiency (Recall) of 83.3% and a background rejection factor of 11.1, corresponding to a background mistag rate (ϵ_B) of 9.0%. The F1 score at this point is 0.8257. Using the default threshold of 0.5 results in slightly different trade-offs, notably slightly higher precision and background rejection, but at the cost of lower signal efficiency. This comparison underscores the benefit of tuning the classification threshold based on the specific analysis goal.

Table 5.1: Performance metrics on the test set at the optimal threshold (max F1 on validation) and the default 0.5 threshold.

Metric	Optimal Threshold (0.4800)	Default Threshold (0.5000)
Accuracy	0.8831	0.8844
Precision	0.8185	0.8346
Recall (ϵ_S)	0.8330	0.8135
F1 Score	0.8257	0.8239
True Positives (TP)	34,867	34,051
False Positives (FP)	7,563	6,725
True Negatives (TN)	76,500	77,338
False Negatives (FN)	6,994	7,810
False Positive Rate (ϵ_B)	0.090	0.080
Background Rejection ($1/\epsilon_B$)	11.1	12.5

5.5.4 Background Rejection vs. Signal Efficiency

A common way to characterize tagger performance in particle physics is to evaluate the background rejection ($1/\epsilon_B$) achieved at specific target signal efficiencies (ϵ_S). Table 5.2 presents these values, calculated on the test set.

Table 5.2: Background rejection and mistag rate (ϵ_B) achieved on the test set for various target signal efficiencies (ϵ_S).

Signal Efficiency (ϵ_S)	Threshold	Mistag Rate (ϵ_B)	Background Rejection ($1/\epsilon_B$)
0.30	0.7871	0.0046	217.8
0.50	0.7110	0.0136	73.7
0.70	0.5965	0.0388	25.8
0.90	0.4015	0.1559	6.4

The results show that the DNN tagger achieves significant background rejection across a range of signal efficiencies. For instance, at ϵ_S=50%, the background mistag rate is only 1.36% (rejection 74), while at ϵ_S=70%, the rejection is 26. This demonstrates the tagger's effectiveness for physics analyses operating at different working points.

5.6 Feature Importance Analysis

To understand which input features contribute most to the DNN's discrimination power, permutation feature importance was calculated on the validation set. This method measures the decrease in AUC when a single feature's values are randomly shuffled. The results are shown in Fig. 5.7.

The plot reveals a clear hierarchy. Variables related to jet substructure (`fj_tau21`), secondary vertex properties (`fj_tau_vertexEnergyRatio_0`), and displaced track impact parameters (`fj_trackSip2dSigAboveBottom_0`) dominate the ranking. This aligns well with the expected physics signatures of boosted $H \to b\bar{b}$ decays discussed in Sec. 2.

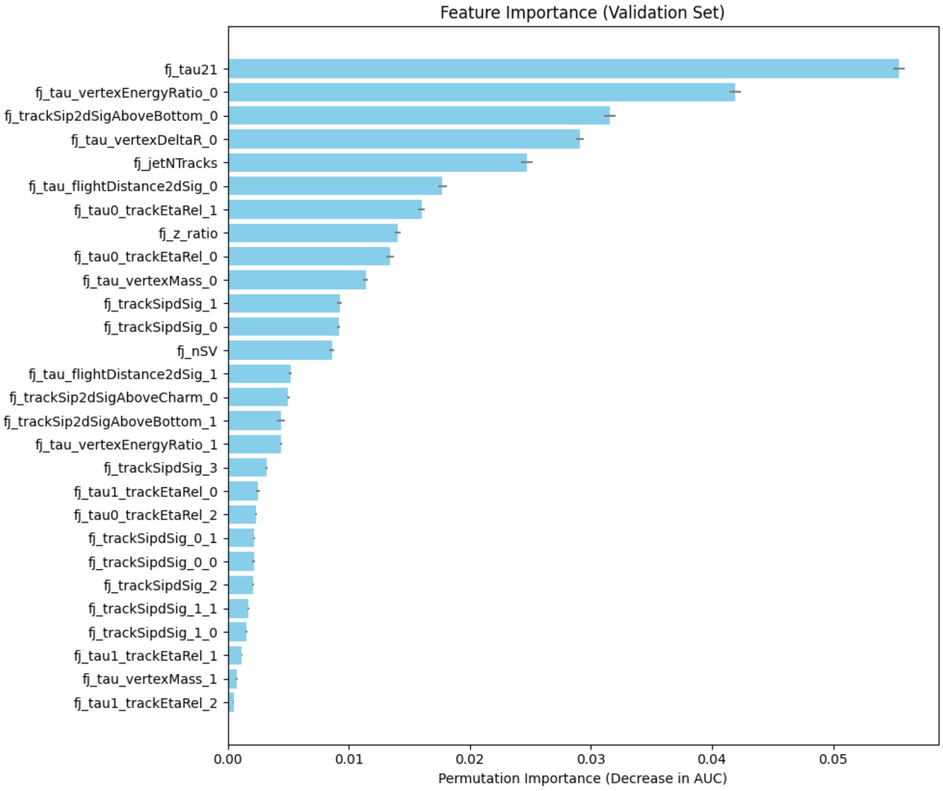

Figure 5.7: Permutation feature importance calculated on the validation set. Features are ranked by the mean decrease in AUC when the feature is shuffled. Error bars represent the standard deviation over 10 repeats.

5.6.1 Most Important Features

Based on Fig. 5.7, the three most important features are:

1. `fj_tau21`
2. `fj_tau_vertexEnergyRatio_0`
3. `fj_trackSip2dSigAboveBottom_0`

Fig. 5.8, Fig. 5.9, and Fig. 5.10 show the distributions of these features for signal and background jets, comparing true labels and predicted labels at the optimal threshold (0.4800).

`fj_tau21` (N-subjettiness Ratio τ_2/τ_1): As discussed in Sec. 2.6, τ_{21} measures the compatibility of the jet's energy distribution with a two-prong versus a one-prong structure. Fig. 5.8 (left) shows that true signal jets ($H \to b\bar{b}$, inherently two-prong) exhibit significantly lower τ_{21} values compared to background QCD jets, which are often initiated by single quarks or gluons (one-prong). The DNN effectively learns this, as the predicted signal jets (Fig. 5.8 (right)) are predominantly those with low τ_{21}. This confirms the crucial role of jet substructure in identifying boosted, hadronically decaying heavy particles.

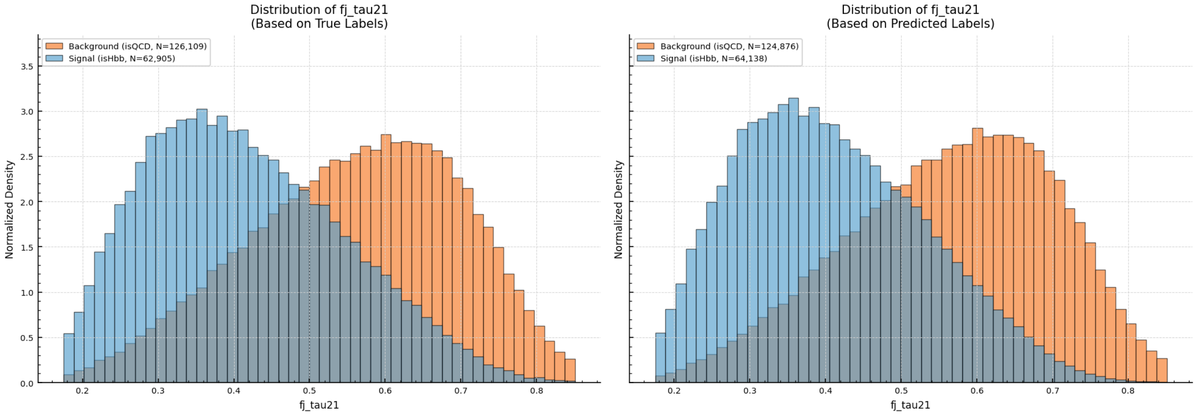

Figure 5.8: Distribution of the feature `fj_tau21`. Left: Separated by true labels. Right: Separated by predicted labels using the optimal threshold (0.4800).

`fj_tau_vertexEnergyRatio_0` (Energy Ratio of Leading SV): This variable represents the energy fraction carried by tracks associated with the leading (highest flight distance significance) secondary vertex (SV) within the jet. Fig. 5.9 (left) shows that true signal jets tend to have higher values for this ratio compared to background jets. This likely reflects that in $H \to b\bar{b}$ decays, a significant portion of the jet energy originates from the B-hadron decays captured by the leading SV. Background jets may have SVs from gluon splitting or light hadron decays carrying a smaller fraction of the total jet energy. The prediction plot (Fig. 5.9 (right)) indicates the DNN uses this feature effectively to identify signal events.

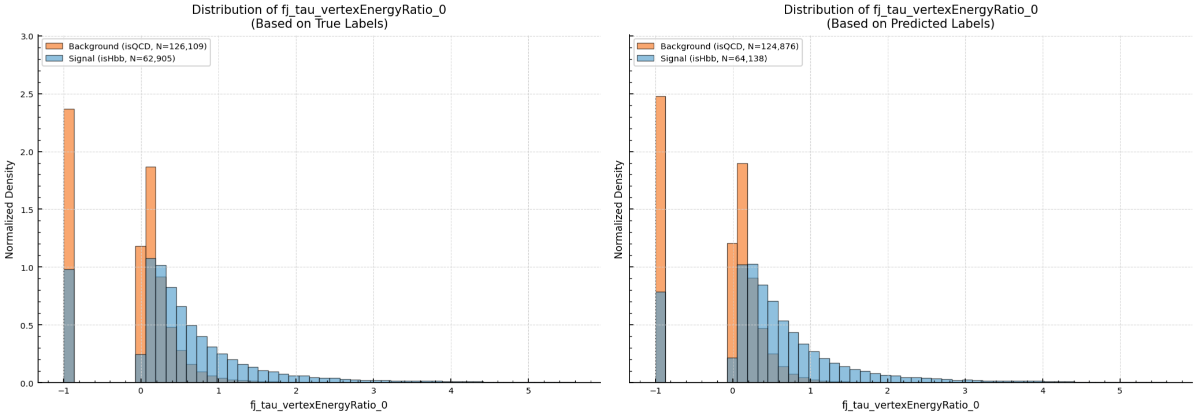

Figure 5.9: Distribution of the feature `fj_tau_vertexEnergyRatio_0`. Left: Separated by true labels. Right: Separated by predicted labels using the optimal threshold (0.4800).

`fj_trackSip2dSigAboveBottom_0` (Track 2D IP Significance relative to B hypothesis): This feature measures the 2D impact parameter significance of the leading track associated with the leading secondary vertex candidate, calculated relative to a B-hadron lifetime hypothesis. As expected from Sec. 2, B-hadron decays produce tracks with large impact parameters. Fig. 5.10 (left) demonstrates that true signal jets have a distribution skewed towards larger positive values for this significance variable compared to background jets. This provides strong evidence for the presence of a B-hadron. The DNN leverages this key lifetime information, as shown by the separation in the predicted label plot (Fig. 5.10 (right)).

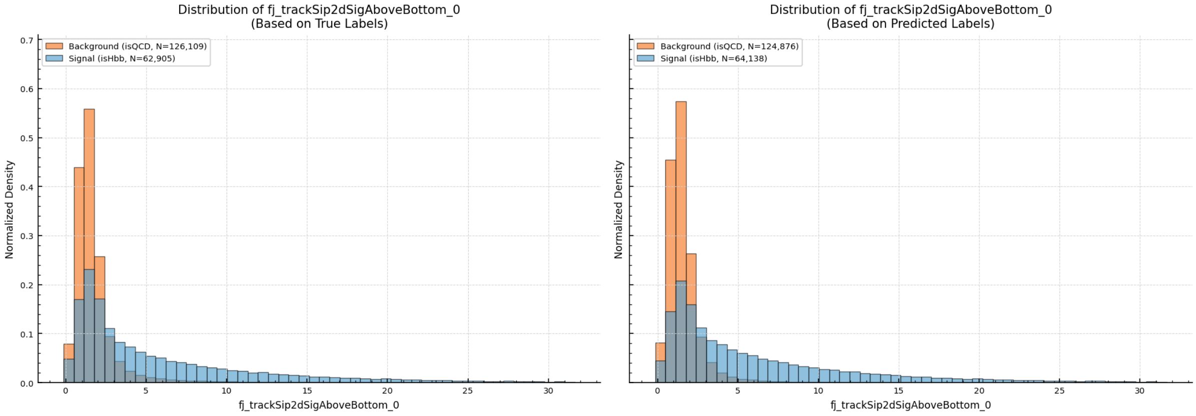

Figure 5.10: Distribution of the feature `fj_trackSip2dSigAboveBottom_0`. Left: Separated by true labels. Right: Separated by predicted labels using the optimal threshold (0.4800).

The high importance of these top three features highlights the synergy of substructure, vertexing, and tracking information for this task.

5.6.2 Least Important Features

Conversely, the three features found to have the lowest permutation importance are:

1. `fj_tau1_trackEtaRel_2` (Relative η of 3rd track w.r.t. subleading SV axis)
2. `fj_tau_vertexMass_1` (Mass of subleading SV)
3. `fj_tau1_trackEtaRel_1` (Relative η of 2nd track w.r.t. subleading SV axis)

These features, while included in the model, contribute minimally to the final discrimination performance in the context of the other 25 features. Potential reasons include focus on subleading/lower-rank objects, less discriminating power of the subleading vertex mass, or redundancy with more highly ranked features. Their inclusion does not significantly harm performance.

5.7 Quantifying the Impact of N-subjettiness

To further understand the relative contributions of different types of information to the tagger's performance, and to provide context for benchmark comparisons (discussed in Sec. 6), the impact of the single most important feature identified in Fig. 5.7—the N-subjettiness ratio `fj_tau21`—was explicitly evaluated. As discussed in Sec. 2.6, τ_{21} is designed to capture the two-prong substructure characteristic of boosted heavy particle decays like $H \to b\bar{b}$, distinguishing them from typically single-prong QCD jets.

An ablation study was performed by training and evaluating the identical optimized DNN architecture and procedure described in Sec. 4, but using only the other 27 input features (i.e., excluding `fj_tau21`). This assessment isolates the contribution of this key substructure variable.

The resulting performance on the independent test set yielded a ROC **AUC of approximately 0.92**. Comparing this to the final **AUC of 0.9441** achieved with the full set of 28 features reveals a significant performance gain (ΔAUC $\approx$ 0.024) attributable specifically to the inclusion of `fj_tau21`.

The substantial importance of τ_{21} stems directly from the distinct radiation patterns within the signal and background jets. Boosted $H \to b\bar{b}$ decays inherently produce two energetic partons (the b and $\bar{b}$ quarks), leading to a fat jet whose energy distribution is concentrated around two distinct axes or 'prongs'. In contrast, the dominant QCD background jets are typically initiated by a single quark or gluon, resulting in a more centrally concentrated, single-prong energy distribution within the fat jet. Mathematically, N-subjettiness (τ_N) quantifies how well the jet's constituents align with N candidate subjet axes [14]:

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} \min\{(\Delta R_{1,k})^\beta, (\Delta R_{2,k})^\beta, ..., (\Delta R_{N,k})^\beta\}$$

where the sum is over constituents k with transverse momentum $p_{T,k}$, $\Delta R_{J,k}$ is the angular distance to subjet axis J, $d_0 = \sum_k p_{T,k} R^\beta$ is a normalization factor with jet radius R, and β is an angular exponent (typically 1). The ratio used in this analysis, $\tau_{21} = \tau_2/\tau_1$, effectively quantifies the compatibility with a two-prong structure (low τ_{21}) versus a single-prong structure (high τ_{21}). While vertexing and impact parameter features (like `fj_tau_vertexEnergyRatio_0` and `fj_trackSip2dSigAboveBottom_0`) are crucial for identifying the $heavy - flavor$ nature of the jet constituents through their long lifetimes, τ_{21} provides complementary information about the underlying hard decay topology (two-body vs. one-body origin) based on the jet's energy flow. This synergy between substructure and lifetime information allows the DNN to achieve optimal discrimination.

5.8 Summary of Results

The DNN double b-tagger developed using 28 engineered features from CMS Open Data simulation achieved excellent performance in identifying boosted $H \to b\bar{b}$ signal jets against QCD multijet backgrounds. The model training converged successfully, and evaluation on the independent test set yielded an **AUC of 0.9441** and an **Average Precision of 0.9004**. A study excluding the top feature (`fj_tau21`) yielded an **AUC of approx. 0.92**, highlighting the significant contribution of substructure information. At an operating point optimized for the F1 score (threshold = 0.4800), the final 28-feature tagger achieves a signal efficiency of 83.3% with a background rejection factor of 11.1 (9.0% mistag rate) on the test set. Feature importance analysis confirmed that jet substructure (`fj_tau21`), vertex properties (`fj_tau_vertexEnergyRatio_0`), and track impact parameters (`fj_trackSip2dSigAboveBottom_0`) are critical for discrimination. The results demonstrate the effectiveness of applying optimized DNNs to high-level features for this challenging particle physics classification task using publicly available data.

Chapter 6

Discussion

6.1 Interpretation of Key Results

The tagger developed in this work achieved excellent performance on the independent test set, yielding an Area Under the ROC Curve (**AUC) of 0.9441** and an **Average Precision (AP) of 0.9004** using the full set of 28 input features. These headline metrics confirm that the DNN architecture effectively learned to distinguish the signal signature from background using the provided high-level features, demonstrating a strong capability for this challenging classification task.

Beyond these overall measures, the tagger demonstrates potent background rejection capabilities crucial for LHC analyses. For instance, at a 50% signal efficiency (ϵ_S), it achieves a QCD rejection factor ($1/\epsilon_B$) of approximately **74**, while at 70% signal efficiency, the rejection remains strong at roughly **26** (Table 5.2). This highlights its potential utility in suppressing backgrounds. Furthermore, at the specific operating point chosen by maximizing the F1-score on the validation set (threshold=0.4800), the tagger achieves a signal efficiency (Recall) of 83.3% with a corresponding precision of 81.9% on the test set (details in Table 5.1). This demonstrates a practical balance between identifying signal events and controlling false positives for this chosen working point. The stable training convergence observed (Fig. 5.2) further validates the chosen network design and training methodology.

6.2 Analysis of Performance Drivers, Feature Roles, Correlations, and Model Complexity

The DNN's strong performance stems from its ability to synthesize information from physically motivated features. As indicated by the feature importance (Fig. 5.7), a synergistic combination is vital. Key substructure information, like that from `fj_tau21`, acts as an initial filter for two-prong topologies. This is critically supported by vertexing features (e.g., `fj_tau_vertexEnergyRatio_0`) and lifetime information (e.g., `fj_trackSip2dSigAboveBottom_0`), which confirm the heavy-flavor origin of these prongs and reduce misidentification from lighter quark/gluon jets that might mimic a two-prong structure.

Beyond these top discriminants, other variables offer necessary detail for robust classification. For instance, secondary vertex data (like `fj_nSV` and `fj_tau_flightDistance2dSig_1`) aid in distinguishing genuine double b-decays from single b-jets that have additional vertices due to processes like gluon splitting ($g \to q\bar{q}$). Similarly, impact parameter measurements from multiple tracks (such as `fj_trackSipdSig_1`, `fj_trackSipdSig_2`, and `fj_trackSipdSig_3`) improve resilience against tracking variations or occasional high-IP tracks in background jets from other sources like strange hadron decays or material interactions. The DNN effectively learns to use this broader set of inputs to refine its classifications.

Feature correlations, as shown in Fig. 5.1, are anticipated, especially for variables describing the same physical object (like an SV). Unlike simpler models that might need feature pruning, DNNs can often leverage these correlations. The network probably learns complex, non-linear patterns; for example, unexpected relationships between SV energy and mass could offer discriminating power not available to linear models or simple cuts. Therefore, using the complete, correlated feature set was a strategic decision to maximize the information for the DNN, enable comprehensive benchmarking, and clearly evaluate the contribution of key features such as `fj_tau21` in a detailed context.

Consequently, the DNN's complexity appears well-justified and necessary. The task of identifying two simultaneous B-hadron decays within a dense, high-p_T fat jet—amidst substantial QCD background and detector resolution limitations—requires sensitivity to subtle, high-dimensional patterns. The achieved performance strongly indicates that simpler models would not be capable of capturing these nuances effectively using the 28 available features.

6.3 Significance and Context of Boosted $H \to b\bar{b}$ Tagging

The theoretical prediction of the Brout-Englert-Higgs mechanism earned Englert and Higgs the 2013 Nobel Prize in Physics. Following the experimental discovery in 2012, the focus shifted to precisely measuring the Higgs boson's properties to test the Standard Model (SM) and search for deviations indicating new physics. The fundamental importance and success of this ongoing experimental program were recently highlighted by the 2025 Breakthrough Prize in Fundamental Physics, awarded collectively to the ATLAS, CMS, ALICE, and LHCb collaborations. The citation explicitly recognized their ”detailed measurements of Higgs boson properties confirming the symmetry-breaking mechanism of mass generation” during LHC Run 2 [30].

The $H \to b\bar{b}$ decay channel, addressed in this thesis, is indispensable for this program because it:

- Is the dominant decay (∼58% BR), offering high statistics.
- Directly probes the Higgs coupling to the bottom quark (y_b), testing the SM mass mechanism.
- In the boosted regime, accesses key production modes (VH, ttH) and BSM searches, requiring advanced techniques like those used here (fat jets, substructure, double b-tagging).

Developing high-performance taggers for this channel directly aids these fundamental physics goals recognized by major scientific awards.

6.3.1 Benchmarking Performance

The `cernopendata/datascience/HiggsToBBMachineLearning` repository [31] provides a relevant public benchmark using the same Open Data record, reporting baseline performance around **AUC ≈ 0.90** with ∼27 features [32]. The final result here (**AUC = 0.9441** with 28 features) is significantly better. Notably, the optimized architecture and training developed here already achieved **AUC ≈ 0.92** on the same 27 features, outperforming the baseline even before adding `fj_tau21`. The final performance thus reflects benefits from both the effective network design and the inclusion of crucial substructure information.

This comparison shows competitive results are achievable with public data and optimized models on engineered features, relevant to the state-of-the-art efforts recognized by the recent Breakthrough Prize.

6.4 Critical Discussion of Limitations and Their Impact

A realistic assessment requires acknowledging study limitations and their potential impact:

- **Simulation Reliance:** The exclusive use of simulation means reported efficiencies and mistag rates likely differ from reality due to imperfect modeling of detector response, pileup, and hadronization. Impact: Without data calibration (deriving scale factors), the tagger cannot be reliably used for quantitative physics measurements, as biases would be uncontrolled.

- **Open Data Constraints:** Using a specific Open Data snapshot restricts the analysis to available variables and reconstruction versions, potentially limiting achievable performance compared to internal analyses with more data or features. Impact: The results demonstrate potential on public data but might not represent the absolute state-of-the-art achievable within the collaboration.

- **Engineered Feature Ceiling:** The tagger's knowledge is confined to the information captured by the 28 pre-processed features. Subtle correlations or patterns in raw constituent data are inaccessible. Impact: Performance might be inherently limited compared to end-to-end deep learning models (e.g., GNNs, PFNs) that learn representations directly from constituents, although these come with higher complexity.

- **Simplified Background Model:** Training primarily against QCD neglects other potential backgrounds (W/Z+jets, $t\bar{t}$) that could pass selections in a real analysis. Impact: The tagger's discrimination against these non-QCD backgrounds is unknown and could be significantly worse, potentially leading to underestimation of total background or requiring dedicated vetoes.

- **Absence of Systematics:** Lacking systematic uncertainty evaluation prevents a complete assessment. Impact: Uncertainties arising from sources like Jet Energy Scale/Resolution (JES/JER), flavor tagging efficiencies, pileup modeling, and theoretical cross-sections would add significant error bars to any physics result derived using this tagger, potentially dominating the statistical uncertainty.

6.5 Concrete Proposals for Future Work with Rationale

Building on this work requires addressing its limitations and exploring extensions:

- **Collision Data Validation:** This is paramount for usability. We can define signal-depleted control regions (e.g., mass sidebands, anti-tagged regions) in Open Data collision samples to measure mistag rates. Use $t\bar{t}$ samples, if feasible, via tag-and-probe to estimate b-tagging efficiency, ultimately deriving simulation-to-data scale factors.

- **Systematic Uncertainty Assessment:** Initially,we can focus on dominant experimental uncertainties by propagating variations in JES/JER (if possible with Open Data tools) and incorporating data-derived scale factor uncertainties for b/c/light-jet tagging including theoretical uncertainties on signal/background cross-sections.

- **Further Training Optimization:** Potentially marginal gains are possible.Since early stopping wasn't triggered at 100 epochs, we can conduct longer runs (e.g., 200 epochs) or employ more sophisticated hyperparameter optimization (e.g., Bayesian optimization extending the initial Keras Tuner search) to probe for further improvements.

- **Lower-Level Feature Exploration:** We can identify Open Data formats with particle-flow constituents, implement suitable architectures (e.g., ParticleNet, PFN), manage the increased computational load, and directly compare performance to quantify the gains from constituent-level learning vs. high-level features.

- **Multi-Class Extension:** We can identify and process relevant Open Data MC samples (e.g., $Z \to b\bar{b}, t\bar{t}$) and train a multi-class network to specifically distinguish $H \to b\bar{b}$ from individual background categories, potentially improving overall purity.

In conclusion, this thesis presented a high-performance DNN $H \to b\bar{b}$ tagger using CMS Open Data, achieving competitive results relevant to the ongoing, highly recognized scientific effort to characterize the Higgs boson. While acknowledging limitations requiring further work (especially data validation and systematics), this study provides a strong foundation and highlights the potential of applying optimized deep learning techniques to public LHC datasets.

Chapter 7

Conclusion

This thesis presented the development, training, and comprehensive evaluation of a Deep Neural Network (DNN) aimed at the challenging task of identifying boosted jets originating from Higgs boson decays to bottom quarks ($H \to b\bar{b}$) and discriminating them from a large background of Quantum Chromodynamics (QCD) multijet events. The study utilized publicly available Monte Carlo simulation samples from the CMS experiment corresponding to Run 2 conditions, accessed via the CERN Open Data Portal (Record 12102), demonstrating the potential of these resources for advanced physics analysis and machine learning development. The focus was on leveraging a specific set of 28 high-level, engineered features derived from reconstructed fat jets (AK8) capturing tracking, vertexing, and jet substructure information.

7.1 Summary of Key Findings

The primary objective—developing and evaluating an effective DNN-based double b-tagger for boosted $H \to b\bar{b}$ using engineered features on Open Data was successfully achieved. The key findings derived from the evaluation on an independent test set are:

- **High Discrimination Performance:** The final tagger, utilizing 28 input features, demonstrated excellent overall discrimination power, achieving a **ROC AUC of 0.9441** and an **Average Precision (PR AUC) of 0.9004**.
- **Effective Working Point:** At an operating point optimized to balance precision and recall (maximizing F1-score on validation), the tagger yields a practical **signal efficiency (Recall) of 83.3%** with a corresponding **precision of 81.9%**. It also provides strong background rejection across various efficiency targets (e.g., rejection factor ∼74 at 50% efficiency).

- **Superiority over Baseline and Feature Impact:** The optimized DNN architecture and training procedure significantly outperformed baseline examples associated with the dataset. The model achieved an **AUC of approximately 0.92** using only 27 features (excluding N-subjettiness), already surpassing the benchmark **AUC of ∼0.90**. The inclusion of the `fj_tau21` substructure variable further boosted performance to the **final AUC of 0.9441**, quantifying the importance of both model optimization and jet substructure information.

- **Identification of Key Discriminants:** Feature importance analysis confirmed that the tagger's success relies on synergistically combining information from jet **substructure** (`fj_tau21`), **vertexing** (e.g., `fj_tau_vertexEnergyRatio_0`), and **B-hadron lifetime** (e.g., `fj_trackSip2dSigAboveBottom_0`), aligning well with the underlying physics principles.

These results collectively demonstrate the efficacy of the developed approach for this specific, challenging tagging task within the constraints of using high-level features and public data.

7.2 Concluding Remarks and Outlook

This thesis successfully developed and validated a high-performance DNN tagger for boosted $H \to b\bar{b}$ identification, contributing to the ongoing efforts in Higgs boson characterization—a field recognized by the highest scientific honors like the 2013 Nobel Prize and the 2025 Breakthrough Prize. **The Python code developed for the data processing, model implementation, training, evaluation, and figure generation presented in this work is publicly available on GitHub [33] at: `https://github.com/TheDevNair/Thesis_project`.** The work underscores the value of CMS Open Data for enabling meaningful research and demonstrates that optimized deep learning models applied to well-engineered features can achieve highly competitive performance relevant to LHC Run 2 standards.

While the tagger shows excellent potential on simulation, its practical application requires addressing the inherent limitations, most critically the **need for validation on collision data** and a **thorough assessment of systematic uncertainties**. These steps, outlined in the future work proposals (Sec. 6.5), are essential to calibrate the tagger and understand the full uncertainty on its performance before it can be reliably used in a physics analysis aiming to measure SM properties or search for new phenomena.

In conclusion, this work provides a robust implementation and a strong performance benchmark for boosted $H \to b\bar{b}$ tagging using deep learning on high-level features within the accessible framework of CMS Open Data, laying a solid foundation for subsequent studies involving real data and physics analysis integration.

Bibliography

[1] R. L. Workman et al. “Review of Particle Physics”. In: *Prog. Theor. Exp. Phys.* 2022 (2022), p. 083C01. DOI: 10.1093/ptep/ptac097.

[2] P. W. Higgs. “Broken Symmetries and the Masses of Gauge Bosons”. In: *Phys. Rev. Lett.* 13 (1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.

[3] G. Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Phys. Lett. B* 716 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

[4] S. Chatrchyan et al. “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”. In: *Phys. Lett. B* 716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

[5] LHCHiggsCrossSectionWorkingGroupCollaboration et al. *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector.* Tech. rep. CERN-2017-002-M. CERN, 2016. DOI: 10.23731/CYRM-2017-002. arXiv: 1610.07922 [hep-ph].

[6] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. “The Anti-k(t) jet clustering algorithm”. In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].

[7] Andrew J. Larkoski et al. “Jet Substructure as a New Paradigm”. In: *Ann. Rev. Nucl. Part. Sci.* 70 (2020), pp. 189–215. DOI: 10.1146/annurev-nucl-101918-023702. arXiv: 1907.01131 [hep-ph].

[8] K. Albertsson et al. “Machine learning in high energy physics community white paper”. In: *J. Phys. Conf. Ser.* 1525 (2020), p. 012003. DOI: 10.1088/1742-6596/1525/1/012003. arXiv: 1807.02876 [physics.comp-ph].

[9] CERN. *CERN Open Data Portal.* http://opendata.cern.ch/. Accessed: May 9, 2025. 2025.

[10] S. Chatrchyan et al. “Particle-flow event reconstruction in CMS and performance for jets, taus, and MET”. In: *JINST* 12.10 (2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003.

[11] CMS Collaboration. “Double Higgs production at CMS”. In: *Proceedings of The 41st International Conference on High Energy physics (ICHEP2022)*. Vol. ICHEP2022. PoS 507. Figure 3 (left panel) adapted for fig:boosted$_{j}et_{c}lustering$.. Sissa Medialab, Dec. 2022. DOI: 10.22323/1.414.0507. URL: https://pos.sissa.it/414/507/.

[12] G. Aad et al. “Impact of the Insertable B-layer on b-tagging Performance for ATLAS”. In: *Nucl. Instrum. Meth. A* 768 (2014), pp. 156–160. DOI: 10.1016/j.nima.2014.09.034. arXiv: 1405.3191 [physics.ins-det].

[13] Andrew J. Larkoski et al. “Soft Drop”. In: *JHEP* 05 (2014), p. 146. DOI: 10.1007/JHEP05(2014)146. arXiv: 1402.2657 [hep-ph].

[14] Jesse Thaler and Ken Van Tilburg. “Identifying Boosted Objects with N-subjettiness”. In: *JHEP* 03 (2011), p. 015. DOI: 10.1007/JHEP03(2011)015. arXiv: 1011.2268 [hep-ph].

[15] S. Chatrchyan et al. “The CMS experiment at the CERN LHC”. In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[16] CMS Collaboration. *Sample with jet, track and secondary vertex properties for Hbb tagging ML studies HiggsToBBNTuple_HiggsToBB_QCD_RunII_13TeV_MC*. 2021. DOI: 10.7483/OPENDATA.CMS.A0N1.Y1B5. URL: https://doi.org/10.7483/OPENDATA.CMS.A0N1.Y1B5.

[17] 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”. In: *JHEP* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[18] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Z. Skands. “A Brief Introduction to PYTHIA 8.1”. In: *Comput. Phys. Commun.* 178 (2008), pp. 852–867. DOI: 10.1016/j.cpc.2008.01.036. arXiv: 0710.3820 [hep-ph].

[19] Apache Software Foundation. *Apache Parquet*. https://parquet.apache.org/. Accessed: May 9, 2025. 2025.

[20] Jim Pivarski et al. “Uproot: Scikit-HEP project Python package”. In: *Journal of Open Source Software* 5.49 (2020). This cites Uproot 3. Check for newer citations or Zenodo DOIs if using Uproot 4/5., p. 2220. DOI: 10.21105/joss.02220.

[21] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. *Searching for Activation Functions*. 2017. arXiv: 1710.05941 [cs.NE].

[22] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG].

[23] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1929–1958. ISSN: 1532-4435. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[24] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90. arXiv: 1512.03385 [cs.CV].

[25] Tom O’Malley et al. *KerasTuner*. https://github.com/keras-team/keras-tuner. Accessed: May 9, 2025. 2019.

[26] François Chollet et al. *Keras*. https://keras.io. Accessed: May 9, 2025. 2015.

[27] Martín Abadi et al. “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems”. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI ’16)*. 2016. arXiv: 1603.04467 [cs.DC]. URL: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.

[28] Tsung-Yi Lin et al. “Focal Loss for Dense Object Detection”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.324. arXiv: 1708.02002 [cs.CV].

[29] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. Conference version of the arXiv preprint. 2015. arXiv: 1412.6980 [cs.LG].

[30] Breakthrough Prize Foundation. *Breakthrough Prize Announces 2025 Laureates In Life Sciences, Fundamental Physics, And Mathematics*. https://breakthroughprize.org/News/88. Accessed: May 9, 2025. Verify the official announcement URL. Sept. 2024.

[31] cernopendata-datascience. *HiggsToBBMachineLearning: Machine learning on CMS Open Data for $H \to bb$ identification*. https://github.com/cernopendata-datascience/HiggsToBBMachineLearning. Accessed: May 9, 2025. Check repository for specific release versions if needed. 2020.

[32] cernopendata-datascience. *HiggsToBBMachineLearning Documentation / README*. https://github.com/cernopendata-datascience/HiggsToBBMachineLearning#readme. Accessed: May 9, 2025. Performance details are within the repository README or linked documentation. 2020.

[33] Dev Nair. *Code for Master Thesis: Development of a Deep Neural Network Double b-tagger for Boosted Topologies using CMS Open Data*. https://github.com/TheDevNair/Thesis_project. Accessed: May 9, 2025. 2025.