Final Report
Italian Student Program 2023

New jet tagging techniques in Vector Boson Scattering (VBS) WV analysis in the semi-leptonic channel with the CMS experiment

Student:
RAFFAELE DELLI GATTI

Supervisors:
IRENE ZOI
JENNIFER NGADIUBA

October 16, 2023

Contents

Abstract

The first evidence for the electroweak vector boson scattering in the $l\nu q\bar{q}$ decay channel of two weak vector bosons WV (V=W, Z), produced in association with two parton jets, was reported in 2021 by the CMS experiment in proton-proton collisions at $\sqrt{s} = 13$ TeV, collected during 2016-2018, with an integrated luminosity of 138 fb^{-1}. The observed electroweak signal strength obtained corresponds to a signal significance of 4.4 standard deviations. The VBS production of vector boson pairs is a rare Standard Model process at the LHC since it is purely EW of order 6 of the neutral weak current coupling and it has a large background contamination. To increase the signal significance up to 5σ, required to claim the observation of the process, larger data sets are needed. The current Run III of the LHC started last year and it will provide more data for the analysis, at the same time it is possible to gain more sensitivity by taking advantage of the new jet tagging techniques recently developed, based on machine learning. In fact, the signal presents a V boson decaying hadronically, therefore its properties must be inferred from reconstructed jets in the final states, which need to be associated (tagged) to their emitting particles.

The present work aims at assessing the efficiency and sensitivity gain achieved with more performing boosted jet tagging techniques in the context of the VBS WV analysis, in particular with novel Deep Learning-based methods commissioned and applied throughout CMS, including the most performing ParticleNet tagger. Data are analyzed with the Coffea framework used in the CMS collaboration, which allows to easily read data in the new NANOAOD format. The report is organized as follows: Section 1 describes briefly the Large Hadron Collider at CERN and illustrates the Compact Muon Solenoid experiment and its sub-detectors, together with the coordinate system employed and the kinematic variables of interest. Section 2 introduces the physics process studied and provides an overview of jet tagging in high energy physics, together with the N-Subjettiness and ParticleNET taggers. In Section 3, the NANOAOD format and the columnar analysis with the Coffea framework are presented; moreover, the phase space is characterized by applying kinematic selections on the Monte Carlo sample, and the results of the efficiency and sensitivity studies are reported. The work ends with the conclusions.

1 The CMS experiment at the LHC

This Section provides an overview of the Large Hadron Collider (LHC), a device housed by the European Organization for Nuclear Research (CERN) that propels charged hadronic particles (protons or ions) along a circular path and collides them to investigate the creation of new particles. The Section also introduces the Compact Muon Solenoid (CMS) experiment at the LHC, offering insights into the structure of the sub-detectors that constitute the apparatus. More specifically, the CMS coordinate system and the kinematic variables used in the present work are described.

1.1 The Large Hadron Collider

A primary goal in a sub-nuclear physics experiment is to generate new particles via collisions [1]. The selection of a specific beam is dictated by the experiment's objective: the particles comprising the beam must be stable, and there must be a capacity to produce and accelerate them in large volumes. The minimal energy loss due to synchrotron radiation is a compelling reason for a hadronic collider, such as the Large Hadron Collider [2] at CERN. Established in 1954, CERN now includes 23 member states and collaborates with numerous scientists, engineers, and technicians worldwide. Despite its name, CERN's research interests extend beyond nuclear physics to primarily focus on particle physics. Several significant discoveries have been made through experiments at CERN, including the first observation of the Higgs boson in 2012, thanks to the CMS [3] (Compact Muon Solenoid) and ATLAS [4] (A Toroidal LHC Apparatus) experiments.

The LHC, an abbreviation for Large Hadron Collider, is a machine of considerable size (approximately 27 km in circumference) that accelerates and collides charged hadronic particles (protons or ions) arranged in two opposite beams to maximize energy in the center of mass [5]. The LHC, comprising two rings of superconducting magnets, is situated in a tunnel about 100 m deep on average. It is located on the France-Switzerland border and is hosted by CERN. The LHC consists of eight arcs and eight straight sections, four of which are the beam collision points where detectors are placed, while the other four are used for machine utilities, radio frequency, collimation, and beam interruption. The arcs contain the dipole bending magnets.

Before being injected into the LHC, the particles pass through an accelerator chain depicted in Figure 1. At present, the LHC collider hosts nine experiments, with ATLAS and CMS being the two largest. These are massive, general-purpose detectors designed to conduct precise tests on the Standard Model and explore new physics phenomena. It is crucial to have two separate detectors pursuing the same objectives to ensure independent measurements.

In the Large Hadron Collider, particles are propelled nearly at the speed of light under high vacuum conditions, to prevent collisions with gas molecules. Particle paths are curved using strong magnetic fields: the LHC employs magnetic dipoles for bending the beams and magnetic quadrupoles for focusing them; these magnets are superconductors maintained at a temperature near zero kelvin by a large cooling system containing superfluid helium-4. In addition to magnets, electromagnetic resonators acting as accelerating cavities are also used. The protons in the LHC circulate around the ring in groups known as bunches. This bunch structure is a result of the radio frequency acceleration scheme.

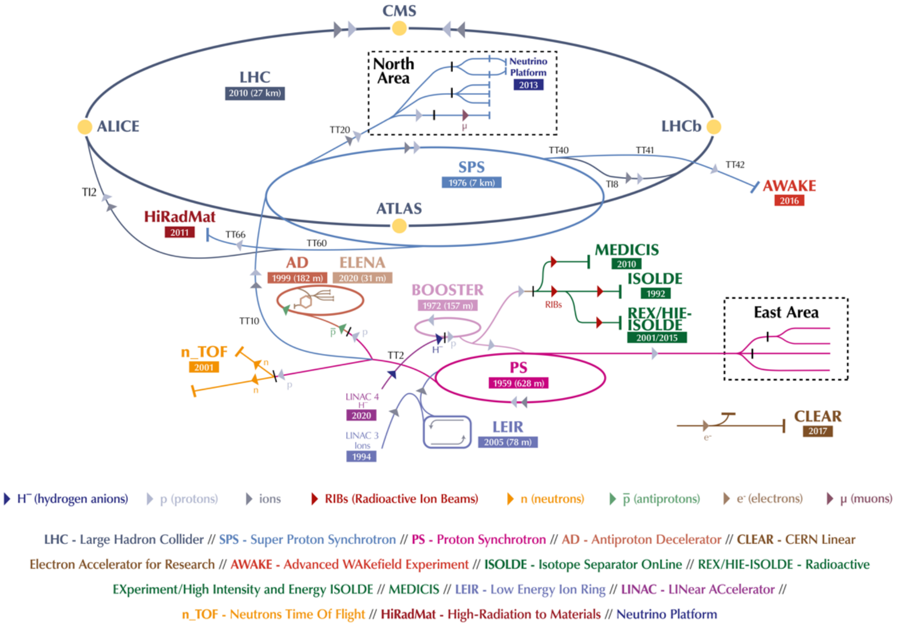

CMS
LHC
2010 (27 km)
North Area
Neutrino Platform
2013
ALICE
LHCb
TT20
TT40
TT41
TT42
SPS
1976 (7 km)
TI8
AWAKE
2016
TI2
ATLAS
HiRadMat
2011
TT66
TT60
AD
1999 (182 m)
ELENA
2020 (31 m)
BOOSTER
1972 (157 m)
MEDICIS
2010
ISOLDE
1992
RIBs
REX/HIE-ISOLDE
2001/2015
East Area
TT10
n_TOF
2001
TT2
PS
1959 (628 m)
LINAC 4
H⁻
2020
LINAC 3
Ions
1994
LEIR
2005 (78 m)
CLEAR
2017
H⁻ (hydrogen anions)
p (protons)
ions
RIBs (Radioactive Ion Beams)
n (neutrons)
p̄ (antiprotons)
e⁻ (electrons)
μ (muons)
LHC - Large Hadron Collider // SPS - Super Proton Synchrotron // PS - Proton Synchrotron // AD - Antiproton Decelerator // CLEAR - CERN Linear Electron Accelerator for Research // AWAKE - Advanced WAKefield Experiment // ISOLDE - Isotope Separator OnLine // REX/HIE-ISOLDE - Radioactive EXperiment/High Intensity and Energy ISOLDE // MEDICIS // LEIR - Low Energy Ion Ring // LINAC - LINear ACcelerator // n_TOF - Neutrons Time Of Flight // HiRadMat - High-Radiation to Materials // Neutrino Platform

Figure 1: The CERN accelerator complex.

The most crucial parameters for a collider are the total collision energy in the center-of-mass reference $\sqrt{s}$ and the number of significant collisions, measured by a quantity known as luminosity. A problem arises when a charged particle, accelerated along a curved path, emits electromagnetic radiation, known as synchrotron light, due to the influence of a magnetic field. The energy loss per revolution is inversely related to the radius of curvature of the path and the cube of the particle's mass, which is why a large hadron accelerator is more beneficial than one using electrons or positrons. The instantaneous luminosity L, measured in $cm^{-2}s^{-1}$, represents the number of potential collisions per unit area and time. The total luminosity $L = \int Ldt$ equates to the number of collisions per unit area that can occur over a specified duration and is measured in inverse barn b^{-1} (where 1 $b = 10^{-24}$ cm^2).
In order to investigate rare phenomena in LHC collisions, high beam energies and intensities are necessary. This means that the maximum number of particles must be concentrated into the smallest possible space at the interaction point to increase the likelihood of proton-proton collisions. The event generation rate is given by

$$R_{ev} = \frac{dN_{ev}}{dt} = \sigma_{ev} \cdot L, \tag{1}$$

where σ_{ev} is the cross section for the event being studied. High luminosity leads to a phenomenon known as pile-up, which is when a large number of interactions happen simultaneously at each collision. This results in many traces from simultaneous events overlapping with the interaction of interest, making it difficult to reconstruct individual events. Furthermore, the large number of interactions typically creates a background that is much larger than the signal, which consists of events with a very small cross section. While most interactions are elastic, the events of interest are those that are inelastic because they result in the formation of new particles; these inelastic productions of new particles occur with high transverse momentum due to momentum conservation.

The LHC is not only CERN's largest experimental apparatus, but it is also currently the world's largest and most powerful particle accelerator. It began operations in 2008 and was recently upgraded, resuming data collection in 2022 with the new Run III. After the current Run III, the High Luminosity-Large Hadron Collider (HL-LHC) [6] project aims to enhance the performance of the LHC by increasing the luminosity beyond its design value.

1.2 The Compact Muon Solenoid

The Compact Muon Solenoid is a detector situated underground near Cessy, France, at the fifth interaction point along the Large Hadron Collider ring. Its purpose is to facilitate a variety of research endeavors, including studies of the Standard Model, Beyond the Standard Model (BSM), and measurements of Higgs boson properties. Weighing approximately $14,000$ tons, the CMS gets its name from its compact size, given the amount of detector material it houses. It measures 21 m in length and 15 m in height. The detector's design is a symmetrical cylinder, allowing for comprehensive coverage of the solid angle due to particles being produced in all directions during collisions.

The CMS apparatus features a superconducting solenoid with an internal diameter of 6 m, generating a 3.8 T magnetic field in the inner part of the detector; the magnetic flux is returned on the outside through a steel yoke, therefore the direction of the magnetic field in the two sections is opposite. For this reason, muons travel following a peculiar "s" shaped trajectory, reported also on the CMS logo. Within the solenoid volume, there are concentric sub-detectors (see Figure 2), which include a silicon tracker, a crystal electromagnetic calorimeter (ECAL), and a hadron calorimeter (HCAL). Each of these is made up of two sections: a central barrel and two end-caps. The hadron forward (HF) calorimeter extends the η coverage provided by the barrel and end-cap hadron calorimeters. The CMS allows precise detection of muons in gas-ionization detectors integrated into the steel return yoke outside the solenoid.

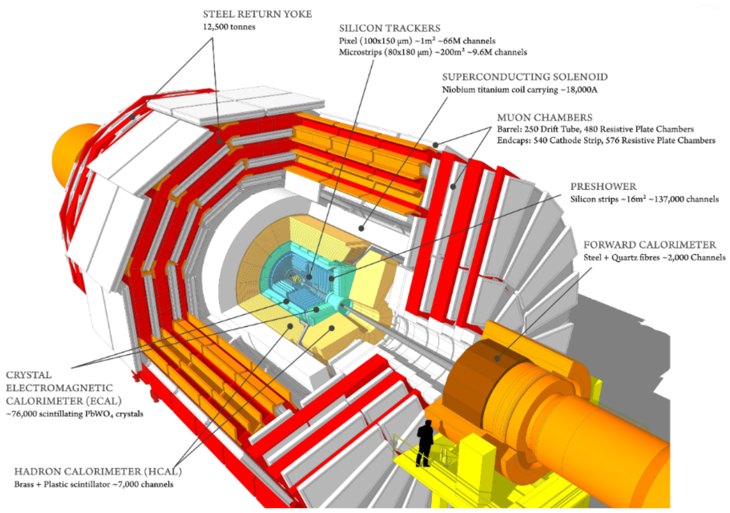

STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~1m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive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

Figure 2: A perspective view of the CMS detector structure.

The proton-proton collisions at the LHC result in a particle shower that travels across the CMS sub-detectors. Particles first enter the inner tracker from the beam interaction region, where the trajectories of charged particles are reconstructed from hits collected in the tracker layers. The magnetic field allows for the measurement of particles' charge and momentum. Electrons and photons are absorbed in the electromagnetic calorimeter cells, where corresponding electromagnetic showers are detected as clusters of energy. Hadrons may also initiate a shower in the ECAL, which is then fully absorbed in the hadron calorimeter. Muons are detected in the muon chambers located in the outermost part of the apparatus, while neutrinos remain undetected and their energy can only be inferred indirectly.

The particle-flow (PF) reconstruction approach [8] combines information from all sub-detectors to identify and reconstruct detected particles' properties. Given the fine spatial granularity of the CMS detector, a high magnetic field allowing good separation between neutral and charged hadrons, and excellent performance in identifying muon tracks, the PF algorithm can provide a global description of an event with high resolution and efficiency, as well as a reduced misidentification rate. Luminosity measurements are provided by various components including the HF calorimeter, silicon pixel tracker, drift tubes in the muon barrel detector [9], and Pixel Luminosity Telescope [10]. The latter is an online luminosity monitoring system rebuilt during an LHC shutdown and installed in 2021.

Protons in the Large Hadron Collider (LHC) collide at a frequency of 40 MHz, however, data storage capacity and computational resources for processing this data are limited. The CMS trigger system [11] selectively reads data in real-time, retaining only events that are relevant for further physics analysis. Following Run III, the LHC will transition to the High Luminosity phase, and CMS will incorporate new sub-detectors like the MIP (Minimum ionizing Particle) Timing Detector [12], enabling 4D reconstruction of final states by adding a time coordinate.

1.3 The CMS coordinate system

The origin of the Cartesian coordinate system is at the nominal collision point, with the x-axis pointing radially towards the LHC center, the y-axis pointing vertically upward, and the z-axis pointing along the beam direction. In the cylindrical coordinate system, the azimuthal angle ϕ is measured from the x-axis in the x-y plane, with r denoting the radial coordinate in this plane and θ representing the polar angle measured from the z-axis. Instead of using the θ angle, the pseudorapidity $\eta = -ln\ tan(\theta/2)$ is used. A sketch of the Cartesian and cylindrical systems is reported in Figure 3 [13]. Angular distances between particles are measured in the $\eta - \phi$ plane:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}, \tag{2}$$

where $\Delta\eta$ and $\Delta\phi$ are respectively the pseudorapidity and azimuthal angle difference between the directions of two particles. The momentum p_T and energy E_T transverse to the beam direction are calculated for a particle of mass m as

$$p_T = \sqrt{p_x^2 + p_y^2},\ E_T = \sqrt{m^2 + p_T^2}. \tag{3}$$

Since protons are only accelerated in the z-direction, transverse momentum must be zero before and after collision. The presence of missing energy in the transverse plane allows for the evaluation of momentum carried by neutrinos, which cannot be detected by CMS. The missing transverse energy $\not{E}_T$ is the magnitude of the missing transverse momentum $\vec{\not{p}}_T$, i.e. the projection on the plane orthogonal to the beams direction of the negative vector sum of the momenta of all final-state particles detected in an event [14]:

$$\not{E}_T = |\vec{\not{p}}_T| = |-\sum \vec{p}_T|. \tag{4}$$

For the momentum conservation in the transverse plane, $\vec{\not{p}}_T$ corresponds to the total transverse momentum of all undetected particles.

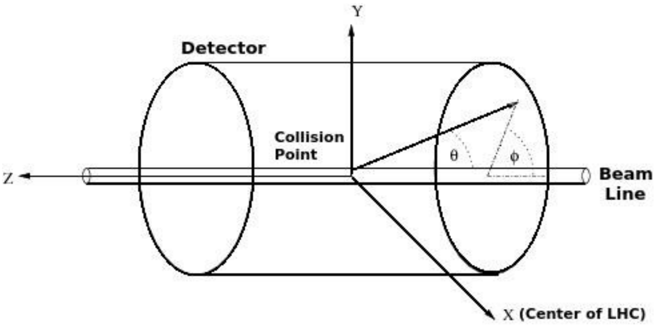


Figure 3: Coordinate system used by the CMS experiment.

2 The physics process

In this Section, the physics process studied in this report is described. In particular, the presence of a vector boson decaying hadronically requires to identify which particle gave origin to a jet in the final state by employing particular jet tagging techniques, introduced in this Section. More specifically, two jet taggers are faced: N-Subjettiness, a traditional tagger employed in the past years, and ParticleNET, based on new Deep Learning methods.

2.1 WV vector boson scattering in the semileptonic channel

The identification of the Higgs boson marked the completion of the observation of the particle constituents of the Standard Model of fundamental interactions; however, the exploration of its scalar and Yukawa sectors is just beginning. Vector boson scattering (VBS) is particularly significant because it prevents the violation of its unitarity arising from direct interaction between vector bosons through counterbalancing diagrams involving the Higgs boson. This cancellation of divergencies is a crucial feature of the SM and one of the main reasons to study VBS processes. Indeed, VBS measurements could offer additional insights into electroweak (EW) symmetry breaking and serve as a powerful tool to test effects beyond the SM.
The VBS production of vector boson pairs at the LHC is rare, as it is a purely EW process of order 6 of the neutral weak current coupling α_{EW}^6, and it has a large background contamination. Only recently the dataset collected at the LHC has become large enough to allow measurements in fully leptonic final states and in the $Z\gamma$ channel. It is compelling to study all VBS final states accessible at the LHC in addition to fully leptonic ones and those with photons. This work addresses the decay of one vector boson into quarks and another W boson into a lepton (electron or muon) and a neutrino. Figure 4 shows the Feynman diagram describing the purely electroweak WW vector boson scattering signal process contributing to this final state.

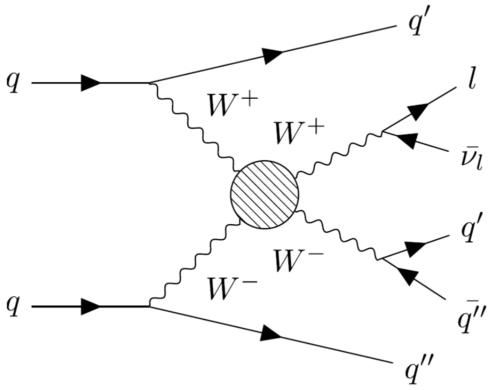


Figure 4: Feynman diagrams contributing to the semi-leptonic final state: purely electroweak WW vector boson scattering signal process contributions in the SM.

The signal presents a single isolated lepton, a moderate level of missing transverse momentum (due to the undetected neutrino), and either three or four jets. A pair of jets is required to have a large invariant mass and large pseudorapidity separation, characteristic of VBS-like events, while the remaining jets result from a vector boson decay. If the boson possesses sufficient momentum in the laboratory frame, its decay products can be gathered in a single jet (*boosted category*), however, at lower momentum, the decay is resolved into two distinct jets (*resolved category*).
The CMS Collaboration has already studied the WV VBS process, where V stands for a W or Z boson, in final states where one boson decays leptonically and another decays hadronically, with data collected with the full Run II dataset, obtained from 2016 to 2018 (corresponding to an integrated luminosity of 138 fb^{-1}), reporting the first evidence for SM EW $l\nu\bar{q}q$ plus two jets production (2021) [15]. The observed significance for the signal was 4.4 standard deviations, with 5.1 expected, so it was sufficient to claim the evidence for the process but not to announce the observation. It is therefore needed, first of all, a larger dataset, for example exploiting the new data collected with the ongoing Run III; moreover, to gain more sensitivity, one could take advantage of the new jet tagging techniques based on novel Deep Learning (DL)-based methods. The purpose of the present report is to assess the sensitivity gain of the new taggers, in order to optimize the baseline analysis. In fact, the V boson decaying hadronically gives rise to one or two jets of colorless particles in the final state, it is therefore necessary to identify the particle that originated the jets as best as possible, to better analyze the process. The classification of the jets, according to the emitter particles, is known as *jet tagging* and will be described with a few more details in the next section. The CMS study of 2021 employs a traditional tagger called N-Subjettines, described in Section 2.3, while the present work aims to study a new tagger known as ParticleNET, based on a Dynamic Graph Convolutional Neural Network (see section 2.4).

2.2 Jet tagging

In the field of high-energy physics (HEP), a jet refers to a collimated sprays of subatomic long-lived particles that are generated from the hadronization process of a quark or gluon [16]. Due to the principle of quantum chromodynamics (QCD) confinement, particles with a color charge, such as quarks, cannot exist independently. When a particle brings color charge fragments, each fragment carries some of the color charge. To maintain confinement, these fragments generate other colored objects around them to form color-neutral entities. This collection of objects is termed a jet, as the fragments tend to move in the same direction, creating a narrow stream of particles. The kinematic properties of jets resemble that of the initial partons that produced them. These jets are detected and analyzed in particle detectors to infer the properties of the original quarks (Figure 5).

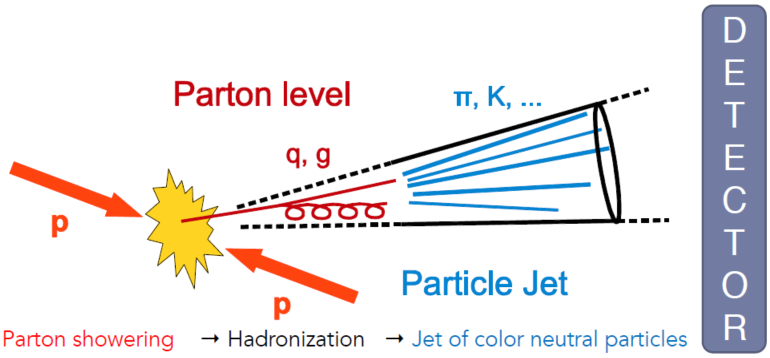


Figure 5: Schematic of a particle jet in a proton-proton collision in a hadronic collider [17].

A jet clustering algorithm is needed to collect the particles inside a shower of colorless particles [18]. Jets can be clustered using different inputs from the CMS detector. There are many algorithms for jet clusterization used by CMS, like the “kt algorithm”, the “Cambridge and Aachen algorithms” and the “anti-kt algorithm” [19], with the last one, abbreviated “AK”, being the most common. In short, one has to iteratively find the two particles i and j in the event which are closest in some distance measure:

$$d_{ij} = min\left(k_{ti}^{2p}, k_{tj}^{2p}\right)\frac{\Delta_{ij}^2}{R^2}, \quad \Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2, \tag{5}$$

and combine them if $d_{ij} < d_{iB}$, where $d_{iB} = k_{ti}^{2p}$ is the distance between particles i and the beam B, while k_{ti}, y_i and ϕ_i are respectively the transverse momentum, rapidity[1] and azimuth of particle i; if $d_{ij} > d_{iB}$ one stops and calls i a jet, removing it from the list of entities. The distances are recalculated and the procedure repeated until no entities are left. Besides the radius parameter R, a parameter p was added to govern the relative power of the energy versus geometrical (Δ_{ij}) scales; according to the power of the energy scale in the distance measure, one could have different algorithms:

- $p = 1 \rightarrow$ kt algorithm (KT);
- $p = 0 \rightarrow$ Cambridge Aechen algorithm (CA);
- $p = -1 \rightarrow$ anti-kt algorithm (AK).

The momentum power (-2) used by the anti-kt algorithm means that higher-momentum particles are clustered first. This leads to jets with a round shape that

[1] Given a beam along the z-direction, a particle with longitudinal momentum k_z, energy E and angle θ with respect to the beam direction has rapidity $y = \frac{1}{2}ln\frac{E+k_z}{E-k_z}$. Massless particles have $y = \eta$.

tend to be centered on the hardest particle. In CMS software this clustering is implemented using the fastjet package [2]. The AK algorithm leads to jets with shapes not influenced by soft perpendicular and strong parallel radiation; as a result, the cone has a stable radius. The size of the jet cone allows us to focus on a different object:

- larger cone: top jet;
- smaller cone: W/Z jet, H jet;
- much smaller: q/g jet, b-jet.

The internal structure of the jet constituents helps us to understand their origin (see Figure 6). The master formula for heavy objects allows us to compute the jet radius as:

$$\Delta R = \frac{2m}{k_t}; \tag{6}$$

the mass of QCD jets changes as a function of momentum, but the mass of heavy particle jets is relatively stable and therefore is possible to define a LO mass scale as the heavy object mass. We usually use two types of distance parameters, 0.4 and 0.8, and when the anti-kt algorithm is used we have AK4 and AK8 jets respectively.

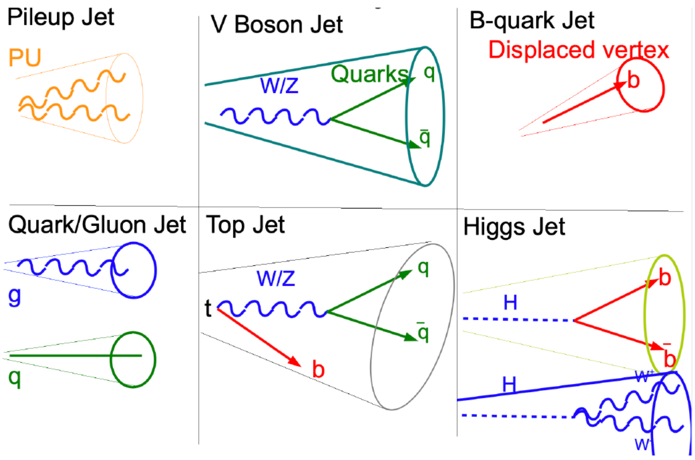


Figure 6: Table with different types of jets having various kinds of substructures.

The list of particle flow candidates includes particles that did not originate from the primary interaction point. The CMS detector at the LHC often encounters multiple collisions happening at the same time, a phenomenon known as *pileup*, during each “bunch crossing”. As a result, particles from these multiple collisions coexist within the detector. Several techniques are employed to eliminate their influence on

[2] https://www.fastjet.fr/

jets. One such method is Charged Hadron Subtraction (CHS), where all charged hadron candidates are linked to a track. If the track isn't linked to the primary vertex, that charged hadron can be excluded from the list. However, CHS is restricted to the area of the detector covered by the inner tracker. Another method is PileUp Per Particle Identification (PUPPI), which was available in Run II. In this method, CHS is applied first, and then all remaining particles are assigned weights based on their probability of being a result of pileup. This technique proves to be more stable and efficient in high pileup situations.

Infrared and collinear (IRC) safety is the property that if one modifies an event by a collinear splitting or the addition of a soft emission, the set of hard jets that are found in the event should remain unchanged: collinear safety is the invariance with random split ($\Delta R \to 0$), while infrared safety is the invariance with random particles ($E \to 0$). IRC safety is an important property of jet algorithms, which have historically been plagued by issues related to IRC safety[3], and a significant amount of the work on them has been directed towards understanding and eliminating these problems.

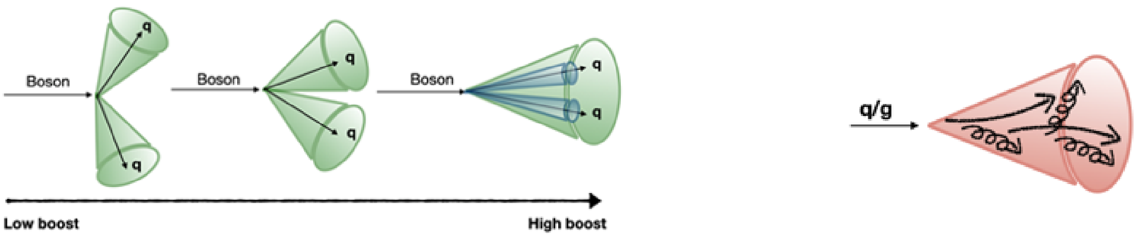
Jet energy corrections, instead, are adjustments made to the energy measurements of jets in particle physics experiments. In fact, the energy of the reconstructed jets does not correspond to the true particle-level energy, which is independent of detector response. The jet energy corrections relate these two values. The process involves several steps: firstly, pileup correction is applied, and then jets are corrected to particle level jets (gen-jet) as a function of p_T and η.

Jets can be contaminated by unenergetic wide-angle radiation which is not associated with the underlying hard substructure, jet *grooming* is essentially a post-processing treatment of jets to remove this wide-angle radiation [20]. This improves the measurement of jet properties. In essence, jet grooming helps to isolate theoretically controlled jet observables and explore possible modifications to the hard substructure of jets. The *soft drop* (SD) algorithm is chosen for the latest analyses for recursively removing soft wide-angle radiation from a jet.

A possible origin for the partons that lead to jets is that they come from the hadronic decay of a heavy particle, for example a top quark, a Higgs boson, or some other yet-to-be-discovered resonance. Jets may also originate radiatively, for example from the emission of a gluon off some other parton in the event. Jet tagging is a classification problem in HEP experiments that aims to identify jets from particle collisions and tag them to their emitter particle. It is essentially a task that aims to distinguish jets arising from particles of interest, such as the Higgs boson, the W/Z bosons or the bottom/top quarks, from other less interesting types of jets. Different particles give rise to jets with varying attributes. For instance, gluon-initiated jets typically have

[3] In fixed-order perturbative QCD calculations, soft emissions and collinear splittings are associated with divergent tree-level matrix elements. Normally the two sources of divergence should cancel, but for IRC unsafe jet algorithms the cancellation is broken and leads to infinite cross sections. Experimental detectors provide some regularization because of their finite resolution and non-zero momentum thresholds, but the extent to which this happens depends on the experiment.

a wider energy distribution compared to those initiated by quarks. Heavy particles with high momentum, such as top quarks and W, Z, and Higgs bosons, can decay hadronically, resulting in jets with unique multiprong structures. As a result, the source particle's identity can be deduced from the properties of the reconstructed jet. This information about particle identity offers valuable insights into the collision events being studied and can significantly aid in distinguishing events stemming from different physics processes. This, in turn, enhances the sensitivity of searches for new particles and measurements of standard model processes. Though most uses essentially identify a jet as coming from a single parton, this association is ambiguous. For example, when a highly boosted W or Z boson decays to two partons, those partons may be so collimated by the Lorentz boost that they will lead to a single jet with substructure (two distinct hard subjets), as shown in Figure 7a, having an invariant mass of 80 GeV; QCD radiative corrections also give substructure to jets, for example a Boosted QCD jet of 80 GeV, as in figure 7b, originates from a hard parton and gains mass through large angle soft splittings.



(a) Boosted V boson jet.

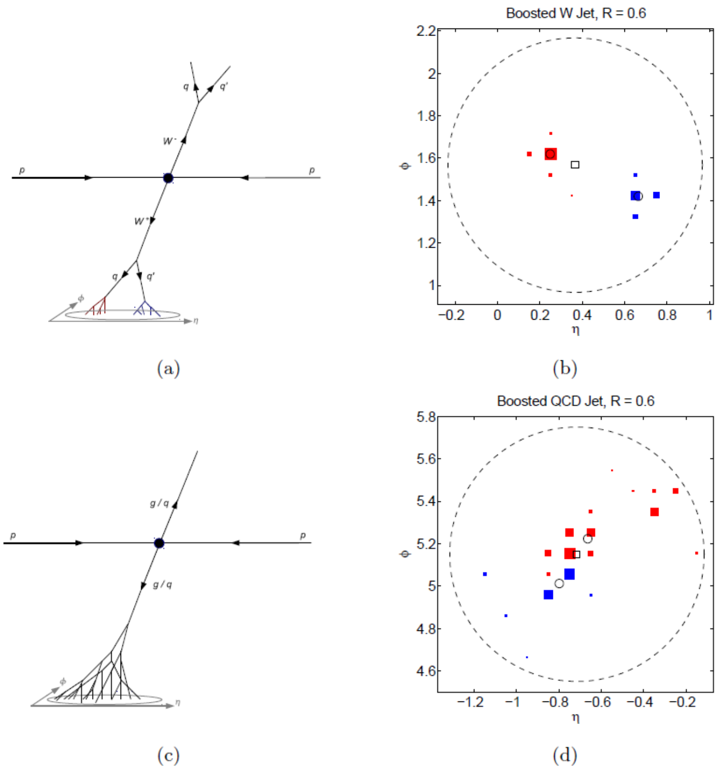
(b) Boosted QCD jet.

Figure 7: Representation of the difference pattern between jets originating from a V boson (7a) and a hard parton (7b).

2.3 N-Subjettiness

N-subjettiness [21] is a method for detecting boosted objects that decay into hadrons, such as electroweak bosons and top quarks. When used with a jet invariant mass cut, N-subjettiness serves as an effective variable for tagging these boosted objects and filtering out the background noise of QCD jets with a large invariant mass. When the boost factor is large, the decay and fragmentation of a boosted object result in a concentrated burst of hadrons; this burst would be identified as a single jet by a standard jet algorithm. Consequently, conventional reconstruction techniques for electroweak bosons and top quarks are rendered ineffective due to the overwhelming background noise from regular QCD jets. One potential solution is to concentrate on channels where the boosted object decays into leptons. However, such methods may discard a significant portion of the original signal, making them potentially suboptimal for detecting new heavy resonances.
Boosted hadronic objects exhibit a distinct energy distribution compared to QCD

jets with a similar invariant mass. To illustrate, let us consider the example of a boosted W boson, as depicted in Figure 8. The jets are grouped using the anti-kt jet algorithm with $R = 0.6$, and the dashed line represents the approximate boundary of the jet. The cells are color-coded based on how the exclusive kt algorithm segregates the cells into two potential subjets. The open square symbolizes the overall direction of the jet, while the open circles represent the directions of the two subjets. The discriminating variable τ_2/τ_1 is used to measure how closely the energy distribution of the jet aligns with the open circles as opposed to the open square.



(a)

(b)

(c)

(d)

Figure 8: Left: Diagram of the fully hadronic decay processes in (a) W^+W^- and (c) dijet QCD events. Right: event displays for (b) W jets and (d) QCD jets with invariant mass near m_W. The jets are clustered with the anti-kt jet algorithm, with $R = 0.6$.

A tagging method employed in the past years for boosted objects is "N-subjettiness", based on a jet shape denoted by τ_N:

$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} min(\Delta R_{1,k}, \Delta R_{2,k}, ..., \Delta R_{N,k}), \tag{7}$$

that basically tells how likely a jet is composed of N subjets. Here, k runs over the constituent particles in a given jet, $p_{T,k}$ are their transverse momenta, and $\Delta R_{J,k}$

is the distance in the rapidity-azimuth plane between a candidate subjet J and a constituent particle k. The normalization factor is $d_0 = \sum_k p_{T,k} R_0$, where R_0 is the characteristic jet radius used in the original jet clustering algorithm. Jets that have $\tau_N \approx 0$ contain all their radiation in line with the proposed subjet directions, indicating they have N (or fewer) subjets. On the other hand, jets with $\tau_N >> 0$ have a significant portion of their energy spread away from the proposed subjet directions, suggesting they have at least $N+1$ subjets. Specifically, τ_2/τ_1 is an effective variable for distinguishing two-prong objects like boosted W, Z, and Higgs bosons. As shown in Figure 9, W jets exhibit smaller τ_2/τ_1 values compared to QCD jets. Conversely, τ_3/τ_2 is useful for identifying three-prong objects like boosted top quarks. Similar to other jet shape methods, τ_N can be computed for each jet. A flexible one-dimensional cut on a function $f(\tau_1, ..., \tau_N)$ can then be used to determine the efficiency/rejection curve.

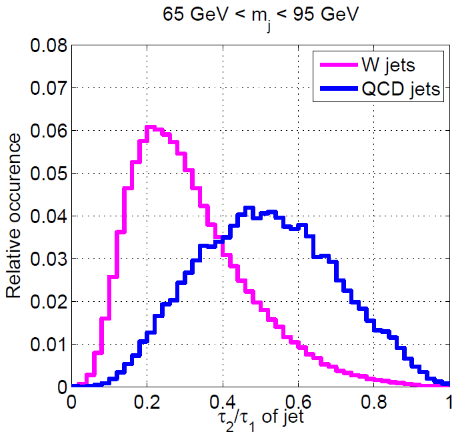


Figure 9: Distribution of τ_2/τ_1 for boosted W and QCD jets, the discriminating variable τ_2/τ_1 gives a good separation between W jets and QCD jets. The mass windows is 65 GeV $< m_{jet} <$ 95 GeV for W jets.

2.4 ParticleNET

Based on the notion of point cloud, ParticleNet [22] is a customized neural network architecture using Dynamic Graph Convolutional Neural Network (DGCNN) for jet tagging problems that consider a jet as an unordered, permutation-invariant set of its constituent particles, effectively a "particle cloud".

Machine learning (ML) has recently revitalized the field of jet tagging. Jets are perceived as images, sequences, trees, graphs, or sets of particles, and ML techniques, particularly deep neural networks (DNNs), are employed to automatically construct new jet tagging algorithms from simulated samples or even real data. This has led to novel insights and enhancements in jet tagging. In this study, we deal with a new deep-learning method for jet tagging that utilizes an innovative representation of jets. Rather than arranging a jet's constituent particles into an ordered structure (like a sequence or a tree), it considers a jet as an unordered collection of particles. This is similar to the point cloud representation of three-dimensional (3D) shapes in computer vision, where each shape is depicted by a set of points in space, with the points themselves being unordered. Hence, a jet can be thought of as a particle cloud. The ParticleNET neural network architecture is used to process the input particle-flow candidates and secondary vertices in a permutation-invariant way. It was tested on two jet tagging benchmarks and was found to significantly outperform all existing methods. In particular, Figure 10 [23] shows a comparison of its efficiency with the DeepAK8 algorithm, a deep one-dimensional convolutional neural network to process particle-flow candidates and secondary vertices associated with the jet for identifying hadronic decays of highly Lorentz-boosted objects and classifying different decay modes based on AK8 jets. On the axis, signal (SE) and background efficiency (BE) are defined as follows:

$$SE = \frac{selected\ sig\ ev}{sig\ ev}, \quad BE = \frac{selected\ bkg\ ev}{bkg\ ev}. \tag{8}$$

ParticleNET presents significant performance improvement. One feature of these taggers is the correlation with the jet mass, so the jet mass shape of the background becomes similar to that of the signal after selection with the tagger ("Mass sculpting"): a mass-independent tagger is needed if one uses the mass variable to separate signal and background tagging signal jets with an unknown mass. Various methods were explored in CMS to reduce the tagger's correlation with the jet mass: the Designing Decorrelated Tagger (DDT) transforms the tagger response as a function of the jet variables, instead DeepAK8-MD has outputs decorrelated with the jet mass and ParticleNET-MD achieves mass independence by exploiting a dedicated signal sample for training.

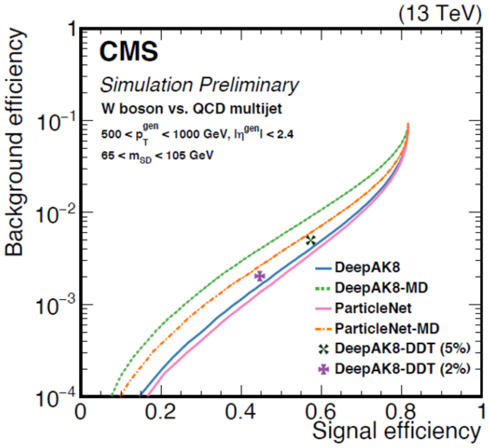


Figure 10: Performance comparison between two particle identification algorithms, particleNET and DeepAK8, for identifying hadronically decaying W bosons.

3 Data analysis

The present Section contains the analysis of the WV vector boson scattering in the semi-leptonic channel with the purpose of assessing the sensitivity gain of the ParticleNET tagger, described in the previous section, with respect to the traditional N-Subjettiness. Monte Carlo simulation of signal and background, stored in the NANOAOD format, are analyzed by means of the Coffea Framework. The kinematic selections applied to spot a fiducial region for the analysis are reported, together with the distributions of the kinematic variables of interest. Finally, efficiency and sensitivity studies are conducted to estimate the gain offered by the novel tagger in the context of the current analysis.

3.1 The NANOAOD data format

With LHC entering a regime where the collision energy will not be dramatically increased, much larger datasets will need to be processed in physics analyses. Therefore, further data reduction is needed to ensure the reach of the experiment's scientific goals. The CMS collaboration has developed and tested a new event data format, called NANOAOD [24], to meet the requirements of a significant portion of physics analyses with an event size of approximately 1 kB. This format is over 20 times smaller than the MINIAOD format and only includes high-level information typically used in the final stages of analysis. The standard data analysis process in CMS involves several stages of data processing and reduction. While collisions occur at a rate of 40 MHz, the final distributions included in published papers usually only include a few observables from a very reduced set of significant events. Data reduction begins in the detector hardware itself, with zero suppression algorithms and trigger systems. The former reduces the event content, while the latter reduces the number of events to be processed. Additional data reduction steps are then applied in the reconstruction and analysis chain. Different types of analysis are expected to require different levels of data reduction: calibration processes depend on a high level of detail, while searches and precision measurements require a large number of events with less detail on low-level detector information. NanoAOD [25] format consists of an Ntuple like format, readable with bare root and containing the per-event information that is needed in most generic analyses. A NanoAOD file contains a main TTree (the ROOT columnar format) named Events, which in turn contains only scalar branches or simple array branches. The current content of NANOAOD consists of all physics objects, including jets, electrons, photons, muons, tau leptons, trigger information, missing transverse energy, generator information, event weights and cleaning flags, primary and secondary vertices, isolated tracks, fatjets and their substructures, Soft Activity information (trackjets) and more. The size used by each object collection is shown in Figure 11.

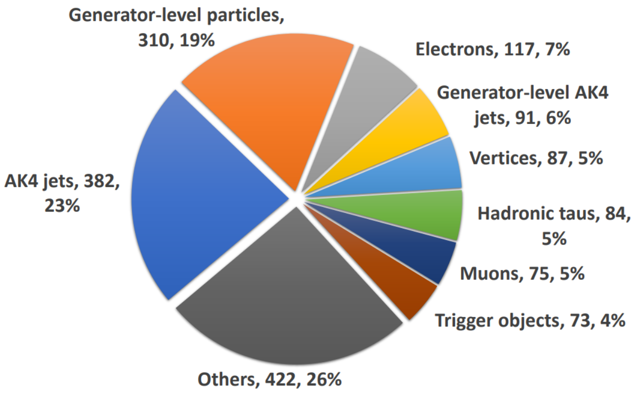


Figure 11: NANOAOD relative size of each physics object collection on a simulated sample.

3.2 The Coffea framework

Coffea[4] (Columnar Object Framework For Effective Analysis) is a preliminary package designed to consolidate the common requirements of a HEP experiment analysis using the scientific Python ecosystem. It leverages uproot and awkward-array to offer an array-based syntax for manipulating HEP event data in an efficient and numpythonic manner. It includes sub-packages that implement histogramming, plotting, and look-up table functionalities that are essential for conveying scientific insight, applying transformations to data, and correcting for discrepancies in Monte Carlo simulations compared to data. Coffea is a collaborative project within the HEP community and is currently in its prototype phase. In High Energy Physics, the conventional method for data analysis is the event loop, where an explicit loop is written to traverse each event and every field within an event that needs to be cut. However, this approach is quite cumbersome compared to the columnar method, as shown in Figure 12, which ideally doesn't involve any explicit loops. In the columnar method, data fields are treated as arrays and analyzed using operations similar to those in numpy. But numpy falls short in HEP as our data is non-rectangular (for instance, an event can have varying numbers of muons). This issue is addressed by awkward arrays[5], which allow us to access data in a columnar manner. However, accessing data is just the first step in analysis.

Coffea provides a variety of features that improve our analysis capabilities. *NanoEvents* is a feature that allows us to assign a schema to our awkward array. These schemas help us to better structure our file and apply physics methods to our data. There are schemas available for some standard file formats, notably NanoAOD, and there is a *BaseSchema* that functions similarly to uproot. *Processors* are Coffea's solution for packaging an analysis in a way that is independent of deployment. Once

[4] https://coffeateam.github.io/coffea/

[5] https://awkward-array.org/doc/main/getting-started/index.html

an analysis is prepared with Coffea, it can be incorporated into a processor and executed using any of several executors to divide it into chunks and distribute it across multiple workers, therefore allowing parallel processing and distributed computing resources. In the next Section, the Monte Carlo simulations analyzed with Coffea are reported.

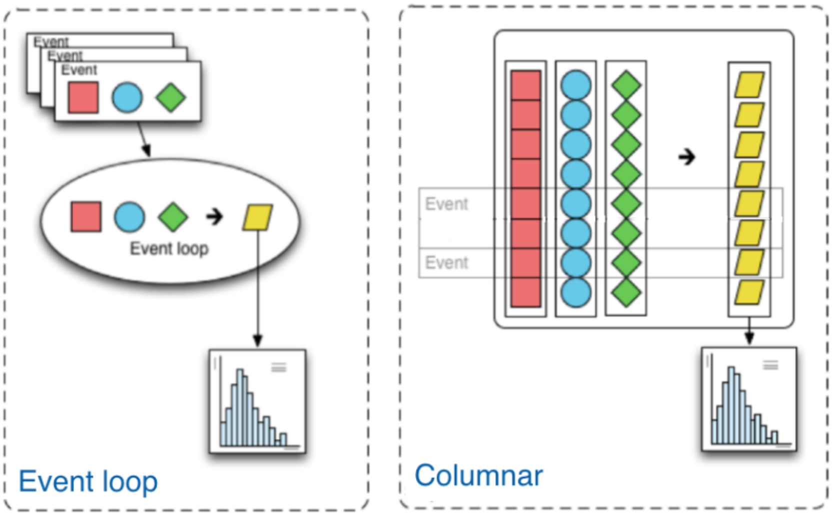


Figure 12: Comparison between the event loop approach in analyzing HEP data (left) and the columnar analysis with the Coffea framework (right).

3.3 Signal and background simulations

The signal sample for this work was simulated with madgraph[6] at the leading order (LO) accuracy, requiring two vector bosons and two quarks in the final state, interfaced with PYTHIA8[7] for parton shower, hadronization, and the simulation of the underlying event; the CP5 MC tune is employed [26]. The NNPDF3.1 NNLO PDF [27] set for parton distribution functions is used. The simulated dataset, for proton-proton collisions at a center of mass energy of 13 TeV, was obtained during the production campaign of Summer (2018) - Run II, with a total integrated luminosity of the data collected of $59.7fb^{-1}$. The sample is availaible in the NANOAODSIMv9 data format. The complete list of the SM signal samples can be found in Table 1.

The main sources of background contamination originate from the production of a single W boson accompanied by jets (W+jets), from $t\bar{t}$ pairs, where one of the W bosons produced by the top quark decays hadronically, and from single top, tW and tZ. Figure 13 shows the Feynman diagrams of W+jets (13a) and $t\bar{t}$ (13b) background. Other contributions to the background are:

[6] http://madgraph.phys.ucl.ac.be/

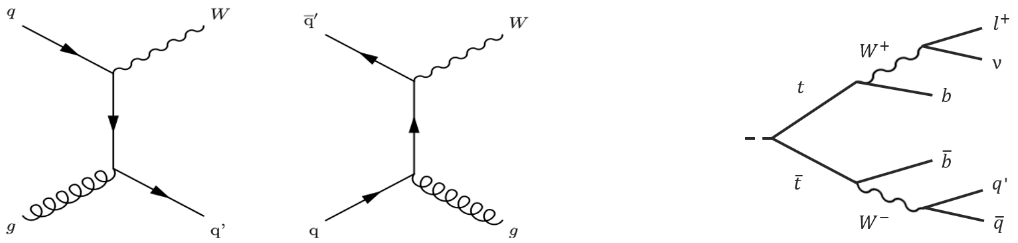
[7] https://pythia.org/

- QCD-WV: nonresonant QCD-associated diboson production;
- non-prompt: data-driven estimation with fakable object technique;
- VBF-V: single V boson EW production in the vector boson fusion channel;
- Drell-Yan lepton pair production;
- ggWW, VVV, Vγ: very small contribution.

In this work, for simplicity, we analyze only the W+jets background, whose complete list of samples can be found in Table 2. W+jets backgrounds are modeled with HT[8] binned generated samples at LO in order to increase the number of MC events describing the high jets multiplicity phase space; LO samples are used to cover the low HT phase space. The stitching of HT bins cross-sections is checked in Figure 14 in an inclusive region defined only by the NanoAOD data format requirements,i.e. one charged lepton with $p_T > 15$ GeV.

Signal process dataset name	Cross section (pb)
/WPHADWMLEPjj_EWK_LO	0.9107
/WPLEPWMHADjj_EWK_LO	0.9114
/WPLEPZHADjj_EWK_LO	0.1825
/WMLEPZHADjj_EWK_LO	0.1000
/WPLEPWPHADjj_EWK_LO	0.0879
/WMLEPWMHADjj_EWK_LO	0.0326
TuneCP5_13TeV-madgraph-pythia8/UL2018-NANOAODSIMv9	

Table 1: Complete list of EW standard model VBS samples.



(a) W+jets background.

(b) Top background.

Figure 13: Feynman diagrams of the main sources of background.

[8] Variables characterizing the visible energy in the transverse plane, defined as the scalar sum of the transverse energy E_T of jets: $HT = \sum_{i=1}^{N_{jet}} E_T$, where N_{jet} is the number of jets [28].

Background process dataset name	Cross section (pb)
/WJetsToLNu_HT-70To100	1292.0
/WJetsToLNu_HT-100To200	1395.0
/WJetsToLNu_HT-200To400	407.9
/WJetsToLNu_HT-400To600	57.48
/WJetsToLNu_HT-600To800	18.77
/WJetsToLNu_HT-800To1200	5.366
/WJetsToLNu_HT-1200To2500	1.074
/WJetsToLNu_HT-2500ToInf	0.008001
TuneCP5_13TeV-madgraphMLM-pythia8/RunIISummer20UL18NanoAODv9	

Table 2: Complete list of EW standard model VBS samples.

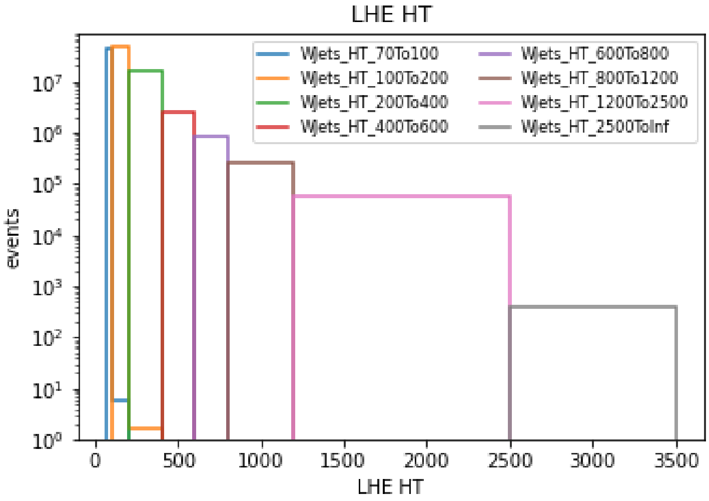


Figure 14: Cross section of W+jets HT binned samples.

3.4 Event selection

Events are selected by applying the same cuts of the official analysis for the electron signal region in the boosted category. The final electron candidates are required to have $p_T > 35$ GeV and a pseudorapidity of $|\eta| < 2.5$. For each event, hadronic jets are clustered from reconstructed particles using the infrared- and collinear-safe anti-kt algorithm (see Section 2.2) with a distance parameter of 0.4 (0.8), respectively AK4 and AK8 jets. Reconstructed jets overlapping with isolated leptons are discarded: $\Delta R(j,l) = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} > 0.4$ (0.8) for AK4 (AK8) jets. In an event, AK4 and AK8 jets are considered in the analysis if they have a $p_T > 30$ GeV and $|\eta| < 4.7$ or $p_T > 200$ GeV and $|\eta| < 2.4$, respectively. AK4 jets overlapping with AK8 jets with $\Delta R(j_{AK4}, j_{AK8}) < 0.8$ are removed from the event.

The analysis targets the VBS production of pairs of vector bosons, WV, in association with two jets originating from the scattered incoming partons; in the channel chosen the W boson decays leptonically and the second boson decays hadronically. Candidate events are required to contain exactly one tightly identified and isolated electron associated with the W boson leptonic decay. Events containing a second loosely identified lepton with $p_T > 10$ GeV are vetoed. Finally, we require moderate missing energy in the transverse plane: PuppyMET > 30 GeV in the event (the PUPPI algorithm is applied to reduce the pileup dependence of the E_T^{miss} observable). An event is assigned to a boosted category if it contains only one AK8 jet that passes the selection criteria as a hadronically decaying vector boson V_{had}, together with at least two AK4 jets. The two AK4 jets with the largest invariant mass are identified as the VBS jets. The fraction of VBS events in the sample is enhanced requiring a large invariant mass $m_{jj}^{VBS} > 500$ GeV and large pseudorapidity interval $\Delta\eta_{jj}^{VBS} = |\eta_{j1}^{VBS} - \eta_{j2}^{VBS}| > 2.5$ for the VBS jets. The leading VBS jet is required to have $p_T > 50$ GeV and the transverse mass of the leptonically decaying W is required to be $m_T^W < 185$ GeV, defined as

$$m_T^W = \sqrt{2p_T(l)E_T^{miss}\left[1 - cos\left(\Delta\phi(p_T(\vec{l}), \vec{p}_T^{miss})\right)\right]}, \tag{9}$$

where $p_T(l)$ is the p_T of the lepton and $\Delta\phi(p_T(\vec{l}), \vec{p}_T^{miss})$ is the azimuthal distance between the lepton and the $\vec{p}_T^{miss}$. The signal region consists of events where no b-jet candidates are found according to the loose working point (WP) of the DeepCSVtagger [29], and the hadronically decaying vector boson invariant mass m_V is between $70 - 115$ GeV for the boosted category, which is consistent with an on-shell W or Z decaying hadronically (soft drop mass of the AK8 jet is used for removing soft, wide-angle radiation from the large radius jet, in order to improve the modeling of the jet mass observable). Finally, in the official analysis the AK8 jets are identified as hadronic decays of Lorentz-boosted W/Z bosons by requiring $\tau_2/\tau_1 < 0.45$ and a groomed AK8 jet mass between 40 and 250 GeV. In this work, the last conditions are released because our goal is to study the efficiency and sensitivity of N-subjettines and ParticleNET and compare them. Figure 15 depicts the final state of the EW VBS in the semi-leptonic channel for the boosted category, while in Table 3 the selections applied to obtain the signal region are summarized.

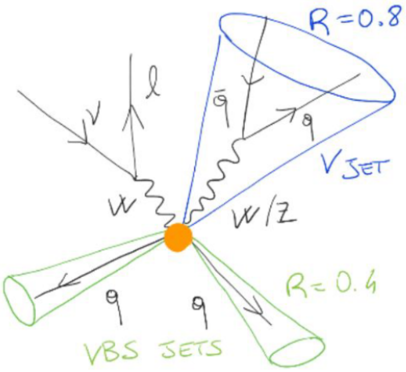


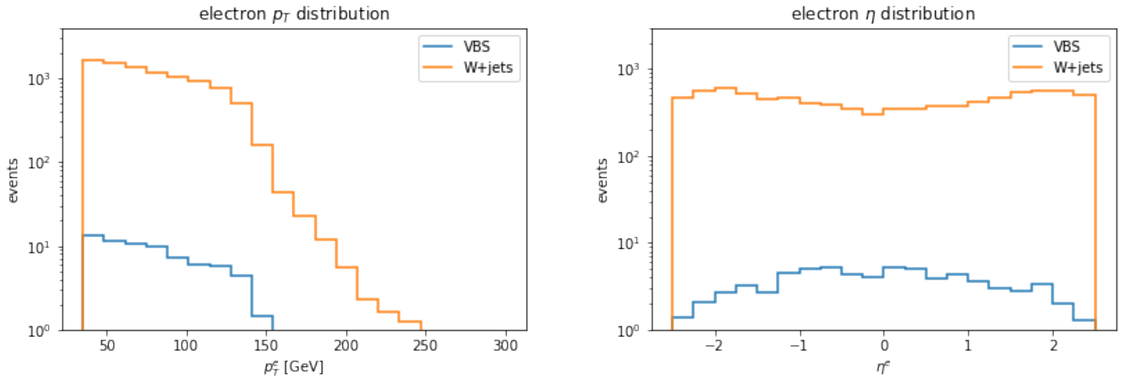
Figure 15: Schematic of the WV vector boson scattering in the semi-leptonic channel for the boosted category.

Event selection	
One isolated tight electron	One AK8 jet
$p_T^e > 35$ GeV, $\lvert\eta^e\rvert < 2.5$	$p_T^{AK8} > 200$ GeV, $\lvert\eta^{AK8}\rvert < 2.4$
PuppyMET > 30 GeV	At least two AK4 jets
No second loosely identified lepton	$p_T^{AK4} > 30$ GeV, $\lvert\eta^{AK4}\rvert < 4.7$,
with $p_T > 10$ GeV	$p_T^{lead} > 50$ GeV, $p_T^{trail} > 30$ GeV
b veto with Loose DeepCSV WP	$m_{jj}^{VBS} > 500$ GeV, $\Delta\eta_{jj}^{VBS} > 2.5$
$m_T^W < 185$ GeV	$\Delta R(j_{AK4}, l) > 0.4$, $\Delta R(j_{AK8}, l) > 0.8$
70 GeV $< m_V <$ 115 GeV	$\Delta R(j_{AK4}, j_{AK8}) > 0.8$

Table 3: Selection criteria defining the fiducial region.

3.5 Kinematic distributions

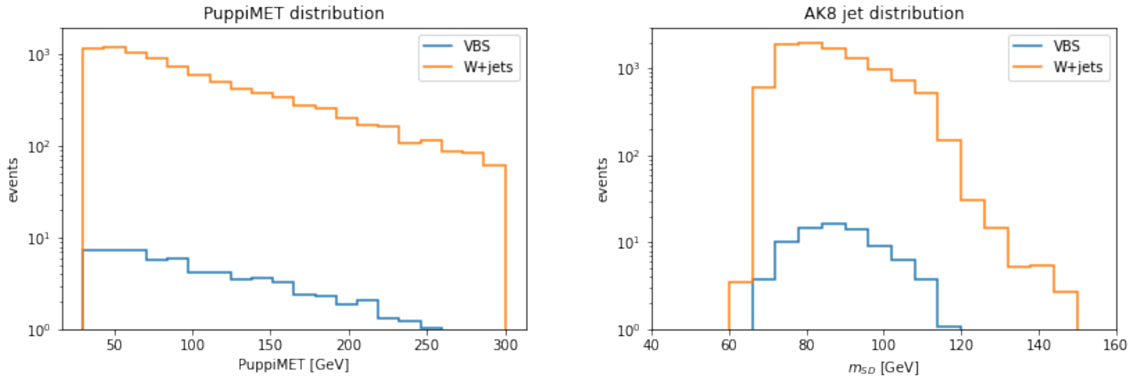
The distributions of the kinematic variables of interest for the $WV \rightarrow l\nu\bar{q}q$ process (in blue) and the W+jets background (in orange), in the electron signal region for the boosted category, are produced with Coffea. The distributions of the transverse momentum and the pseudorapidity of the electron are shown in Figure 16. Figure 17, instead, shows the PuppyMET and the $AK8$ jet m_{SD} distributions, in particular in Figure 17b one can see that the mass of the $AK8$ jet peaks around 80/90 GeV, consistently with an on-shell W or Z decaying hadronically. In Figure 18, the distributions of the transverse momentum for the vector boson scattering jets are reported: the p_T distribution for the leading VBS jet is in Figure 18a, while the one for the trailing VBS jet is in Figure 18b. Finally, VBS jets invariant mass and $\Delta\eta$ distributions are shown in Figure 19, in particular one can see from the $\Delta\eta$ distribution in Figure 19b that the VBS jets present large pseudorapidity separation (the typical signature of VBS-like events).



(a) 35 GeV $< p_T^e <$ 300 GeV.

(b) $-2.5 < \eta^e < 2.5$.

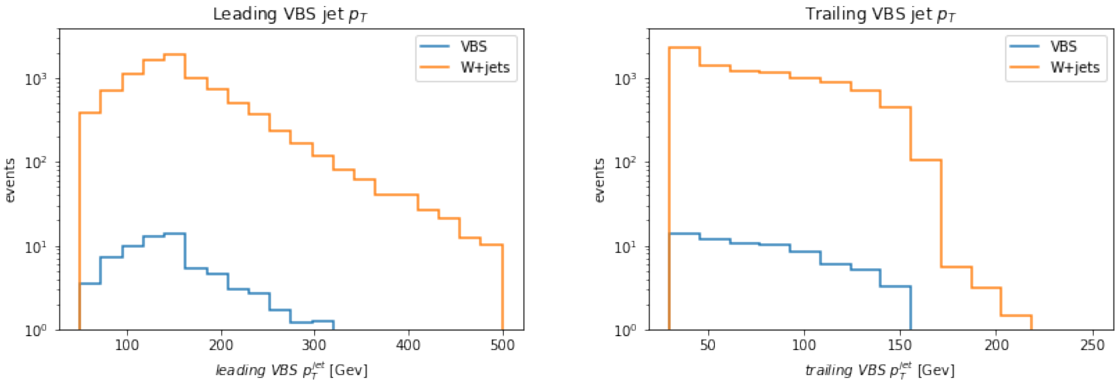
Figure 16: Electron transverse momentum and pseudorapidity distributions for the $WV \rightarrow l\nu\bar{q}q$ process (in blue) and the W+jets background (in orange), in the electron signal region for the boosted category.



(a) 30 GeV < PuppiMET < 300 GeV.

(b) 50 GeV $< m_{SD} <$ 150 GeV.

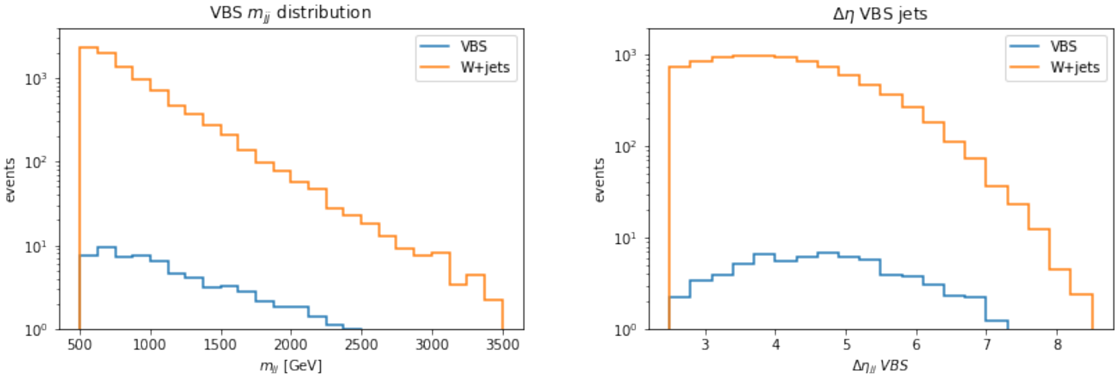
Figure 17: PuppyMET and $AK8$ jet m_{SD} distributions for the $WV \rightarrow l\nu\bar{q}q$ process (in blue) and the W+jets background (in orange), in the electron signal region for the boosted category.



(a) 50 GeV $< p_T^{lead} <$ 500 GeV.

(b) 30 GeV $< p_T^{trail} <$ 250 GeV.

Figure 18: VBS jets leading and trailing p_T distributions for the $WV \rightarrow l\nu\bar{q}q$ process (in blue) and the W+jets background (in orange), in the electron signal region for the boosted category.



(a) 500 GeV $< m_{jj}^{VBS} <$ 3500 GeV.

(b) $2.5 < \Delta\eta_{jj}^{VBS} < 8.5$.

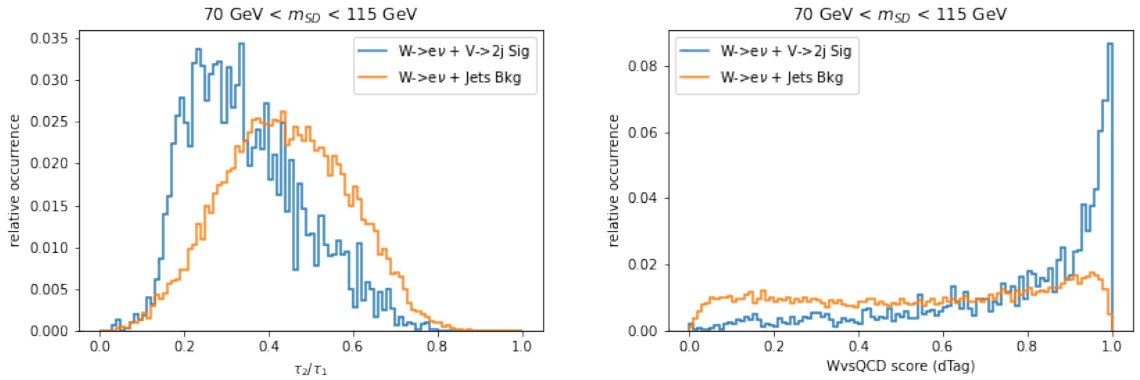
Figure 19: VBS jets invariant mass and $\Delta\eta$ distributions for the $WV \rightarrow l\nu\bar{q}q$ process (in blue) and the W+jets background (in orange), in the electron signal region for the boosted category.

3.6 Efficiency and sensitivity studies

In this last section, we investigate the signal efficiency for individual V jets and the background efficiency for QCD jets, by employing different taggers; sensitivity studies are also performed. To do that, first of all one needs the distributions of the scores of the different taggers, both for the signal and the background. Figure 20a shows the distribution of τ_2/τ_1 for the signal (blue) and the background (orange), the discriminating variable τ_2/τ_1 gives a good separation between boosted V jets and QCD jets. The working point of the official analysis is obtained by setting the cut $\tau_2/\tau_1 < 0.45$, corresponding to a signal efficiency of 78% and a background efficiency of 53%. In Figure 20b, instead, is reported the W vs QCD score of the dTag for both the signal (blue) and the background (orange). Figure 21 shows the W vs QCD score obtained with ParticleNET for both the signal (blue) and the background (orange), in particular Figure 21b is the mass decorrelated tagger (see Section 2.4), where a mixed discriminant, defined as

$$[p(X \to cc) + p(X \to qq)]/[p(X \to cc) + p(X \to qq) + p(QCD)] \tag{10}$$

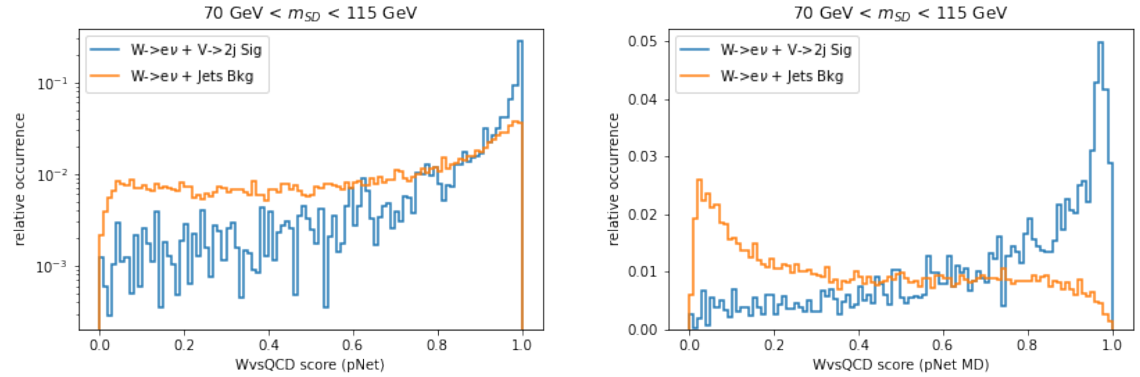
is used for W boson identification.



(a) N-Subjettiness.

(b) dTag.

Figure 20: The τ_2/τ_1 of N-Subjettiness (20a), W vs QCD score obtained with dTag (20b).



(a) ParticleNET.

(b) ParticleNET mass decorrelated.

Figure 21: W vs QCD score distributions obtained with ParticleNET.

The efficiency of the different taggers, as defined in Equation 8, is studied by varying the working point; while for N-Subjettiness the WP is defined by applying the cut $\tau_2/\tau_1 < X$, with the purpose of obtaining high signal efficiency and low background rejection, for the other taggers the WP is defined by the following requirement: score $> X$. Figure 22 shows the background efficiency versus the signal efficiency of the different taggers; the working point of the official analysis with N-Subjettiness is highlighted with a dot. The background efficiency reduction obtained with ParticleNET respect to N-Subjettiness is $\sim$ 20% for a signal efficiency of 78%, in the context of the VBS WV analysis. Figure 23 shows the sensitivity $S/\sqrt{B}$ [30] versus the signal efficiency of the different taggers; the working point of the official analysis with N-Subjettiness is highlighted with a dot. The sensitivity gain of ParticleNET with respect to N-Subjettiness is $\sim$ 11% for a signal efficiency of 78%, in the context of the VBS WV analysis.

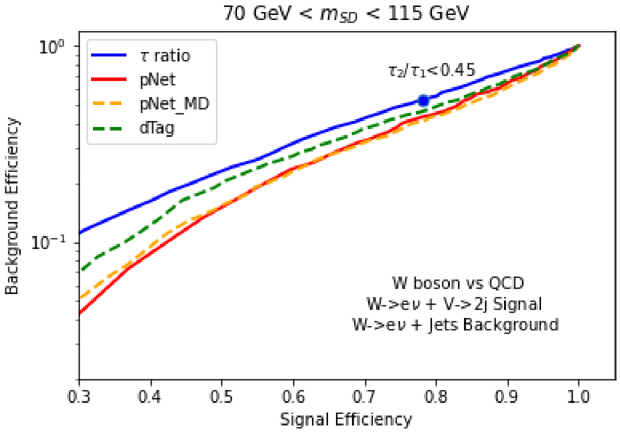


Figure 22: Efficiency of the algorithms for identifying hadronically decaying W bosons.

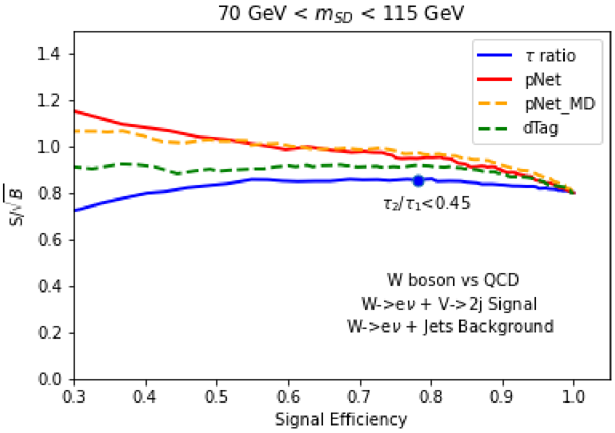


Figure 23: Sensitivity of the algorithms for identifying hadronically decaying W bosons.

Conclusions

The WV(V=Z, W) vector boson scattering process in the semi-leptonic channel was studied at the CERN Large Hadron Collider with the Compact Muon Solenoid experiment, using a Monte Carlo sample for proton-proton collisions at 13 TeV, with an equivalent total luminosity of 59.7 fb^{-1} (2018). Such a final state is characterized by a single isolated tight electron, together with moderate missing energy in the transverse plane, in association with the W boson leptonic decay. In particular, the present analysis aimed to study the events assigned to the boosted category, characterized by only one AK8 jet that passed the selection criteria as a hadronically decaying vector boson V_{had}, together with at least two AK4 jets. The background source for the event analyzed in this work was the W+jets background contribution. The Coffea Framework was employed to analyze the MC simulations, available in the new NANOAOD data format (version 9).

The latest developments in ML-based identification algorithms of highly Lorentz boosted heavy particles in CMS were explored, in particular the purpose was to study the efficiency and the sensitivity of a novel jet tagger, called ParticleNET, and to compare them with the traditional tagger N-Subjettiness. Compared to N-Subjettiness, ParticleNET showed a background efficiency reduction $\sim$ 20% and a sensitivity gain $\sim$ 11%, for a signal efficiency of 78% (corresponding to the working point of the official analysis), in the context of the VBS WV analysis. Thus, ParticleNET presented significant performance improvement. The present studies can be further expanded by:

- studying the efficiency and sensitivity also in the muon signal region;
- exploring also the Z vs QCD score of the jet taggers;
- including also the top background;
- exploiting full Run II statistics.

The analysis can be found for reference at the GitHub repo[9].

[9] `https://github.com/rdelliga/Coffea_Swan.git`

References

[1] M. Thomson, “Modern particle physics”, Cambridge University Press (2013).

[2] Evans L., Byant P., “LHC machine. The CERN Large Hadron Collider: Accelerator and Experiments.”, Journal of Instrumentation, Volume 3 (2008). DOI: https://dx.doi.org/10.1088/1748-0221/3/08/S08001

[3] CMS Collaboration, “The CMS experiment at the CERN LHC”, JINST 3 S08004 (2008). DOI: https://dx.doi.org/10.1088/1748-0221/3/08/S08004

[4] ATLAS Collaboration, “The ATLAS Experiment at the CERN Large Hadron Collider.” Bulletin of the American Physical Society 75 (2008). DOI: https://dx.doi.org/10.1088/1748-0221/3/08/S08003

[5] Lopes, Ana and Perrey, Melissa Loyse, “FAQ-LHC The guide”, CERN-Brochure, (2017). https://cds.cern.ch/record/2809109

[6] Brüning, Oliver Sim and Lucio Rossi. “The High-Luminosity Large Hadron Collider.” Nature Reviews Physics 1 (2019): 241-243.

[7] CMS Collaboration, “Particle-flow reconstruction and global event description with the CMS detector.” (2017). DOI: https://doi.org/10.48550/arXiv.1706.04965

[8] CMS Collaboration, “Particle-flow reconstruction and global event description with the CMS detector.” (2017). DOI: https://doi.org/10.48550/arXiv.1706.04965

[9] CMS Collaboration, “Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS.”, The European physical journal. C, Particles and fields 81 9 (2021): 800. DOI: https://doi.org/10.48550/arXiv.2104.01927

[10] CMS BRIL Collaboration, “The Pixel Luminosity Telescope: A detector for luminosity measurement at CMS using silicon pixel sensors.” (2022). DOI: https://doi.org/10.48550/arXiv.2206.08870

[11] Tosi, Mia. “The CMS trigger in Run 2.” (2017). https://cds.cern.ch/record/2290106

[12] CMS Collaboration, “Precision Timing with the CMS MTD Barrel Timing Layer for HL-LHC.” (2022). DOI: https://doi.org/10.22323/1.380.0116

[13] Perez, Genessis, “Unitarization Models For Vector Boson Scattering at the LHC.” 10.5445/IR/1000082199 (2018).

[14] CMS Collaboration, “Performance of the CMS missing transverse momentum reconstruction in pp data at $\sqrt{s} = 8$ TeV.” Journal of Instrumentation 10 (2014): P02006 - P02006. DOI: https://doi.org/10.48550/arXiv.1411.0511

[15] CMS Collaboration, “Evidence for WW/WZ vector boson scattering in the decay channel $\ell\nu$qq produced in association with two jets in proton-proton collisions at $\sqrt{s} = 13$ TeV.” (2021). DOI: https://doi.org/10.48550/arXiv.2112.05259

[16] https://cms-opendata-guide.web.cern.ch/analysis/selection/objects/jets/

[17] https://twiki.cern.ch/twiki/bin/viewauth/CMS/SWGuideCMSDataAnalysisSchoolLPC2023JetExercise

[18] Salam, Gavin P.. “Towards jetography.” The European Physical Journal C 67 (2009): 637-686. DOI: https://doi.org/10.48550/arXiv.0906.1833

[19] Cacciari, Matteo et al. “The anti-k t jet clustering algorithm.” (2008). https://doi.org/10.48550/arXiv.0802.1189

[20] Larkoski, Andrew J. “Improving the understanding of jet grooming in perturbation theory.” Journal of High Energy Physics 2020 (2020). DOI: https://doi.org/10.48550/arXiv.2006.14680

[21] Thaler, Jesse and Ken Van Tilburg, “Identifying boosted objects with N-subjettiness.” Journal of High Energy Physics 2011 (2010): 1-28. DOI: https://doi.org/10.48550/arXiv.1011.2268

[22] Qu, Huilin and Loukas Gouskos, “ParticleNet: Jet Tagging via Particle Clouds.” (2019). DOI: https://doi.org/10.48550/arXiv.1902.08570

[23] CMS Collaboration, “Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques.” (2020). https://cds.cern.ch/record/2707946

[24] CMS collaboration “NANOAOD: a new compact event data format in CMS.” EPJ Web Conf 245 (2020). https://inspirehep.net/files/07fb1b522cd4c9166b9a4d0167ef02ec

[25] https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookNanoAOD#The_Events_TTree

[26] CMS Collaboration,“Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements”, Eur. Phys. J. C 80 (2020) no.1, 4. DOI: https://doi.org/10.48550/arXiv.1903.12179

[27] Ball, Richard D., et al. “Parton Distributions from High-Precision Collider Data.” The European Physical Journal C, vol. 77, no. 10, Oct. 2017. DOI: https://doi.org/10.1140/epjc/s10052-017-5199-5

[28] Cms Collboration, “Search for Supersymmetry at the LHC in Events with Jets and Missing Transverse Energy.” American Physical Society, vol. 107, no. 22, Nov. 2011. DOI: https://doi.org/10.48550/arXiv.1109.2352

[29] CMS Collaboration, “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV.” J. Instrum. 13 (2018) P05011. DOI: https://doi.org/10.1088/1748-0221/13/05/P05011

[30] Punzi, Giovanni. “Sensitivity of Searches for New Signals and Its Optimization.” (2003). DOI: https://doi.org/10.48550/arXiv.physics/0308063