Βαθμονόμηση αδρονικών πιδάκων στο πείραμα CMS

Προπτυχιακή Διπλωματική Εργασία του φοιτητή:

Νιφόρος Γεράσιμος

Επιβλέπων Καθηγητής: Κουσουρής Κωνσταντίνος

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΦΥΣΙΚΗΣ

Αθήνα, Ιούνιος 2023

Abstract

The subject of this thesis is the study of supervised machine learning regression algorithms and their applications on the field of elementary particle physics, the field that asks probably the most significant question on science, "What is matter made of?". In particular, we aim to explore the possibilities of using machine learning algorithms on calibrating hadron jets' transverse momentum measurements, like the ones produced and measured on the CMS (Compact Muon Solenoid) experiment at the Large Hadron Collider in Geneva. The algorithms studied on this thesis are Boosted Decision Trees, Artificial Neural Networks and Deep Neural Networks. The data used to both train and test the algorithms came from Monte Carlo simulations of proton-proton collisions. The analysis was made both theoretically and practically, first presenting the nature of the problem addressed, then presenting the theory behind the algorithms used, after that presenting the classical correction method, some first results and, as a conclusion, a final exploration of these results. The data used and the regression models were all created on C++ with the ROOT and TMVA software.

Περίληψη

Η παρούσα διπλωματική εργασία έχει σκοπό την μελέτη αλγορίθμων μηχανικής μάθησης παλινδρόμησης υπό επίβλεψη και τις εφαρμογές τους στον τομέα της φυσικής στοιχειωδών σωματιδίων, τον τομέα που ρωτά ίσως την πιο σημαντική ερώτηση για την επιστήμη, "Από τί είναι η ύλη φτιαγμένη;". Συγκεκριμένα, στοχεύουμε στην αναζήτηση της δυνατότητας χρήσης των αλγορίθμων μηχανικής μάθησης στη βαθμονόμηση μέτρησης της εγκάρσιας ορμής πιδάκων αδρονίων, όπως αυτοί παράγονται και μετρούνται στο πείραμα CMS (Compact Muon Solenoid), στον Μεγάλο Επιταχυντή Αδρονίων (LHC) στη Γενεύη. Οι αλγόριθμοι που μελετήθηκαν στην συγκεκριμένη εργασία είναι τα ενισχυμένα δέντρα απόφασης (BDTs), τα τεχνητά νευρωνικά δίκτυα (ANNs) και τα νευρωνικά δίκτυα βαθιάς εκμάθησης (DNNs). Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση αλλά και τον έλεγχο των αλγορίθμων προέρχονται από εξομοιώσεις Monte Carlo της σύγκρουσης πρωτονίου-πρωτονίου. Η ανάλυση έγινε από θεωρητική και πρακτική σκοπιά, πρώτα παρουσιάζοντας τη φύση του προβλήματος, μετά την θεωρία πίσω από τους αλγορίθμους που χρησιμοποιήθηκαν, ύστερα παρουσιάζεται ο κλασικός τρόπος διόρθωσης, κάποια πρώτα αποτελέσματα και, κλείνοντας, μία τελικά διερεύνηση των αποτελεσμάτων. Τα δεδομένα που αξιοποιήθηκαν καθώς και τα μοντέλα παλινδρόμησης έγιναν σε C++ , με τη χρήση των λογισμικών ROOT και TMVA.

Ευχαριστίες

Πρώτα και κύρια θα ήθελα να ευχαριστήσω τον επιβλέων καθηγητή Κωνσταντίνο Κουσουρή, καθώς η παρούσα διπλωματική εργασία δεν θα ήταν δυνατή χωρίς την αμέριστη βοήθεια, την καθοδήγηση, την στήριξη και την υπομονή του. Οφείλω και ένα μεγάλο ευχαριστώ στην ομάδα των διδακτορικών φοιτητών που ήταν πάντα εκεί να μου απαντήσουν κάποια απορία μου. Ένα τεράστιο ευχαριστώ πάει επίσης στους γονείς και την οικογένεια μου, που φρόντισαν από όταν ακόμα ήμουν μαθητής να έχω όσο το δυνατόν περισσότερους δρόμους μπροστά μου να επιλέξω, και για τη συνεχή στήριξη που μου προσφέρουν με κάθε τρόπο. Τέλος, ευχαριστώ από καρδιάς τους φίλους και τις φίλες μου, καθώς είναι οι δικοί μου "γίγαντες" από τους οποίους βλέπω πιο "μακριά".

Περιεχόμενα

Κεφάλαιο 1

Τα Στοιχειώδη Σωματίδια

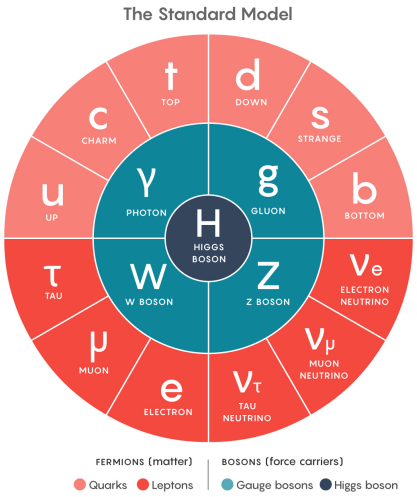
1.1 Εισαγωγή

Από την πληθώρα των σωματιδίων που έχουν παρατηρηθεί μέχρι σήμερα, μόνο μερικά έχουν βρεθεί ελεύθερα στη φύση. Τα περισσότερα από τα ελεύθερα αυτά σωματίδια παράχθηκαν και παρατηρήθηκαν στην κοσμική ακτινοβολία. Η κοσμική ακτινοβολία αποτέλεσε την σπουδαιότερη πηγή για τη μελέτη των διαφόρων σωματιδίων. Σήμερα, η ύπαρξη μεγάλων επιταχυντών μας δίνει τη δυνατότητα να παρατηρήσουμε μέσω των πειραμάτων σκέδασης μία πληθώρα από νέα σωματίδια, τα περισσότερα από τα οποία είναι πολύ ασταθή για να υπάρξουν ελεύθερα στη φύση. Τα νέα αυτά σωματίδια παρατηρούνται ως προϊόντα μιας αντίδρασης σωματίων. Σε ένα πείραμα σκέδασης έχουμε επιτάχυνση του ενός ή και των δύο αρχικών σωματιδίων τα οποία στη συνέχεια θα συγκρουστούν. Μετά τη σύγκρουση, τα προϊόντα της σκέδασης, η γωνία σκέδασης, η ενεργός διατομή και μερικά άλλα στοιχεία είναι ποσότητες που ανιχνεύονται και μετρούνται, και από τις οποίες μπορούν να εξαχθούν χρήσιμες πληροφορίες. Μία πρώτη ταξινόμηση των σωματιδίων, ανάλογα με το σπιν τους, είναι σε φερμιόνια και μποζόνια. Τα πρώτα έχουν σπιν ημιακέραιο, ενώ τα δεύτερα έχουν σπιν ακέραιο. Σε ένα δεύτερο διαχωρισμό, τα σωματίδια, ανάλογα με το είδος της αλληλεπίδρασης στην οποία συμμετέχουν χωρίζονται σε δύο μεγάλες κατηγορίες. Στην πρώτη κατηγορία κατατάσσονται εκείνα τα σωματίδια που μπορούν, εκτός των άλλων αλληλεπιδράσεων, να μετέχουν και σε ισχυρές αλληλεπιδράσεις. Στη δεύτερη κατηγορία ανήκουν τα σωματίδια που δεν συμμετέχουν καθόλου σε ισχυρές αλληλεπιδράσεις αλλά μόνο σε άλλου είδους αλληλεπιδράσεις. Τα σωματίδια της πρώτης κατηγορίας ονομάζονται αδρόνια, ενώ τα σωματίδια της δεύτερης κατηγορίας λεπτόνια. Τα αδρόνια με τη σειρά τους χωρίζονται σε δύο υποκατηγορίες, τα βαρυόνια, που είναι φερμιόνια και ταυτόχρονα τα πιο βαριά αδρόνια, και τα μεσόνια που είναι μποζόνια. Το βαρυόνιο είναι ένα σύνθετο υποατομικό σωματίδιο που αποτελείται από τρία κουαρκ/αντικουάρκ ενώ τα μεσόνια αποτελούνται από ένα κουάρκ και ένα αντικουάρκ.[1]

1.2 Το Καθιερωμένο Πρότυπο

Όλη η ύλη επομένως αποτελείται από 3 είδη στοιχειωδών σωματιδίων, λεπτόνια, κουάρκ και μεσολαβητές των δυνάμεων. Όπως υπάρχουν οι αριθμοί για το φορτίο (Q), τον λεπτονικό αριθμό (L_e) για τα λεπτόνια (και οι αντίθετοι αριθμοί για τα αντισωματίδια τους) κα, έτσι υπάρχουν και οι 6 "γεύσεις" των κουαρκ (up, down, strange, charm, top, bottom). Τέλος, υπάρχουν τα φωτόνια ως μεσολαβητές της ηλεκτρομαγνητικής δύναμης, τα μποζόνια W,Z για την ασθενή πυρηνική και τα γλουόνια για την ισχυρή πυρηνική (πιθανόν και το γκραβιτόνιο για την βαρυτική). Στο Καθιερωμένο Πρότυπο υπάρχουν 8 κουαρκ και αυτά με τη σειρά τους φέρουν την ιδιότητα του χρώματος (red,green, blue και τα "αντίστροφα" τους anti-red, anti-green και anti-

blue). Τα κουαρκ όμως δεν μπορούν να υπάρξουν μεμονωμένα στη φύση. Υπάρχουν μόνο σε "άχρωμες"-"λευκές" καταστάσεις μέσα σε αδρόνια, δηλαδή σε διπλέτες για τα μεσόνια (πχ. ως κόκκινο+αντικόκκινο=άσπρο) ή σε τριπλέτες για τα βαρυόνια (κόκκινο+πράσινο+μπλε=άσπρο). Ενώ όμως δεν μπορούμε να τα απομονώσουμε, μπορούμε να δούμε τα αποτελέσματα της ύπαρξης τους, ειδικά στους πίδακες που προκύπτουν από τη σκέδαση των πρωτονίων, θεωρώντας πλέον το κινούμενο πρωτόνιο σαν ένα κινούμενο σύνολο γλουονίων και κουαρκ.[7]



Σχήμα 1.1: Τα στοιχειώδη σωματίδια του Καθιερωμένου Προτύπου (πηγή εικόνας: Quanta Magazine)

Κβαντική Χρωμοδυναμική

Η δυσκολία που ανέδειξε η πολυπλοκότητα των αδρονίων, καθώς και οι φαινομενικά μη-σχετιζόμενες θεωρίες προσέγγισης τους, ανέδειξε μια ενιαία προσέγγιση για τις στοιχειώδεις αλληλεπιδράσεις. Αυτή η προσέγγιση βασίζεται στο γεγονός ότι, σύμφωνα με το Καθιερωμένο Πρότυπο, ο κόσμος γύρω μας αποτελείται κατά βάση από κουάρκ και λεπτόνια, τα οποία αλληλεπιδρούν μεταξύ τους ανταλλάζοντας είτε φωτόνια μέσω της ηλεκτρομαγνητικής αλληλεπίδρασης, είτε γλουόνια μέσω της ισχυρής αλληλεπίδρασης είτε μποζόνια μέσω της ασθενούς. Ενώ έχει ενοποιηθεί η θεωρία για τις ηλεκτρομαγνητικές και τις ασθενείς αλληλεπιδράσεις, η θεωρία για τις ισχυρές συνεχίζει να εμπλουτίζεται μέχρι και σήμερα. Αυτή η θεωρία, καθώς βασίζεται ως επί το πλείστον στη συμμετρία του χρώματος για τα κουαρκ, παίρνει την ονομασία Κβαντική Χρωμοδυναμική (Quantum ChromoDynamics – QCD). Η QCD είναι μια θεωρία πεδίου η οποία καταφέρνει με απλό τρόπο να περιγράψει μία δύναμη, που να εξαρτάται από την ιδιότητα του χρώματος, μεταξύ των κουάρκ. Η δύναμη αυτή παράγεται με την ανταλλαγή γλουονίων χρώματος, τα οποία είναι “δεμένα” μεταξύ τους αλλά και με τα κουάρκ που εμπλέκονται. Παρότι η συγκεκριμένη θεωρία σχετίζεται με την QED, έχει μια σημαντική διαφορά. Συγκεκριμένα η QED, ως θεωρία που περιγράφεται από αβελιανή ομάδα, δεν επιτρέπει στα φωτόνια να δημιουργήσουν ζεύγη μεταξύ τους. Από την άλλη, η QCD, ως ομάδα που περιγράφεται από μη αβελιανές ομάδες συμμετρίας (συγκεκριμένα την SU(3)) περιγράφει γλουόνια τα οποία μεταφέρουν “χρώμα” και επομένως μπορούν να δημιουργούν ζεύγη.

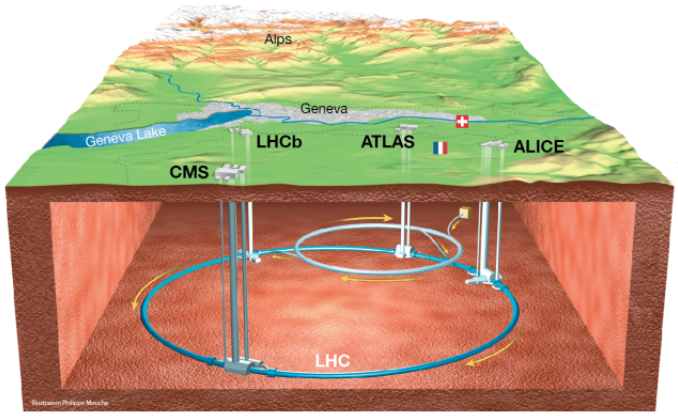
Η QCD περιγράφει με επιτυχία την αλληλεπίδραση παρτονίων (κουάρκ και γλουόνια) σε μικρές αποστάσεις ($r \ll R$) εκεί όπου η ισχύς της είναι χαμηλή και μπορεί να αξιοποιηθεί η θεωρία διαταραχών. Σε μεγάλες αποστάσεις ($r \sim R$) η αλληλεπίδραση γίνεται τόσο ισχυρή, ώστε φαίνεται να μην επιτρέπει στα κουάρκ ή στα γλουόνια να βρεθούν σε ελεύθερες καταστάσεις. Οι μόνες καταστάσεις που μπορούν να παρατηρηθούν είναι οι “άχρωμες” καταστάσεις, οι καταστάσεις εκείνες δηλαδή που ο συνδυασμός των χρωμάτων (ή των αντι-χρωμάτων) δημιουργούν

το λευκό χρώμα (κατά το μοντέλο RedGreenBlue). Αυτές οι καταστάσεις είναι τα αδρόνια που παρατηρούνται στα πειράματα.

Η δημιουργία των αδρονίων από τα "άχρωμα" κουαρκ παραμένει στον πυρήνα των προβλημάτων για την θεωρία των ισχυρών αλληλεπιδράσεων. Τα δεδομένα που συλλέγονται από τα πειράματα συνήθως αναλύονται αξιοποιώντας την θεωρία διαταραχών της QCD για μικρές αποστάσεις και φαινομενολογικά μοντέλα για αποστάσεις από r $\sim$ R και πάνω. Επομένως γίνεται εμφανής η ανάγκη μελέτης του μηχανισμού δημιουργίας σε υψηλές ενέργειες για την ολοκληρωμένη επεξήγηση της θεωρίας QCD αλλά και για τον έλεγχο της για μικρές αποστάσεις. Τα παρτόνια δεν βρίσκονται ελεύθερα, όμως σε υψηλές ενέργειες δημιουργούν πίδακες αδρονίων, οι οποίοι αντικατοπτρίζουν τις ιδιότητες των αρχικών παρτονίων πολύ πιο ολοκληρωμένα από μονάδες ελεύθερων αδρονίων, και ο μηχανισμός δημιουργίας τους αποτελεί το βασικό στοιχείο αυτής της εργασίας.

1.3 Ο Μεγάλος Επιταχυντής Αδρονίων (LHC) και ο ανιχνευτής CMS

Στις εγκαταστάσεις του Ευρωπαϊκού Κέντρου Πυρηνικών Ερευνών (CERN) βρίσκεται ο Μεγάλος Επιταχυντής Αδρονίων (Large Hadron Collider-LHC) ο οποίος αποτελεί τον μεγαλύτερο και ισχυρότερο επιταχυντή σωματιδίων στον κόσμο. Είναι κυκλικού σχήματος, με περίμετρο τα 27 χιλιόμετρα και βρίσκεται 50 έως και 175 μέτρα κάτω από τη γη. Αξιοποιώντας ισχυρό μαγνητικό πεδίο (8,3 Tesla) ο LHC επιταχύνει 2 δέσμες πρωτονίων σε αντίθετες κατευθύνσεις με ταχύτητες που πλησιάζουν την ταχύτητα του φωτός. Οι ηλεκτρομαγνήτες είναι κατασκευασμένοι από πηνία ειδικού ηλεκτρικού καλωδίου που λειτουργεί σε υπεραγώγιμη κατάσταση, διοχετεύοντας αποτελεσματικά την ηλεκτρική ενέργεια με ελάχιστη αντίσταση ή απώλεια ενέργειας. Αυτό απαιτεί ψύξη των μαγνητών στους 1,85 Kelvin, θερμοκρασία χαμηλότερη από την μέση θερμοκρασία στο διάστημα. Για τον λόγο αυτό μεγάλο μέρος του επιταχυντή συνδέεται με σύστημα διανομής υγρού ηλίου, το οποίο ψύχει τους μαγνήτες. Μια άλλη ειδική κατηγορία μαγνητών χρησιμοποιείται για να φέρει τα σωματίδια που αποτελούν την δέσμη πιο κοντά ώστε να επιτευχθεί η σύγκρουση με τα σωματίδια της αντίθετα κινούμενης δέσμης. Συγκρούσεις σωματιδίων γίνονται σε 4 προκαθορισμένα σημεία του LHC, με μέγιστη ενέργεια σύγκρουσης στο σύστημα κέντρου μάζας 14 TeV. Σε αυτά τα 4 σημεία έχουν τοποθετηθεί οι 4 ανιχνευτές ATLAS, CMS, LHCb και ALICE. Ο LHC ξεκίνησε την λειτουργία του τον Σεπτέμβρη του 2008 και από τότε έχει συμβάλει ανυπολόγιστα στην μελέτη της δομής, της συμπεριφοράς και των αλληλεπιδράσεων των στοιχειωδών σωματιδίων, καθώς και στην κατανόηση των θεμελίων που δομούν την ύλη και το σύμπαν.[11]



Σχήμα 1.2: Η τοποθεσία του LHC και των επιμέρους πειραμάτων (πηγή εικόνας: RTU)

Compact Muon Solenoid (CMS)

Ο Compact Muon Solenoid (CMS) είναι ένας ανιχνευτής γενικού ενδιαφέροντος, κυλινδρικού σχήματος, σχεδιασμένος να ανιχνεύει όλα τα γνωστά στοιχειώδη σωματίδια. Έχει μήκος 28,7 μέτρα, διάμετρο 15 μέτρα, ενώ ζυγίζει 14 χιλιάδες τόνους. Είναι κατασκευασμένος γύρω από ένα ισχυρό σωληνοειδή μαγνήτη μήκους 13 μέτρων, ο οποίος παράγει μαγνητικό πεδίο 3,8 Tesla παράλληλα με τον άξονα της δέσμης με σκοπό να καμπυλώνει τις τροχιές των σωματιδίων. Οι επιμέρους ανιχνευτικές διατάξεις τοποθετούνται ομοαξονικά, στη δέσμη ως κυλινρικές επιφάνειες. Ο ανιχνευτής κλείνει ερμητικά στα δύο του άκρα με δύο κάθετους δίσκους (endcaps), ώστε να είναι αποτελεσματικότερη η ανίχνευση όλων των σωματιδίων από την σύγκρουση των πρωτονίων στο κέντρο του.[11]

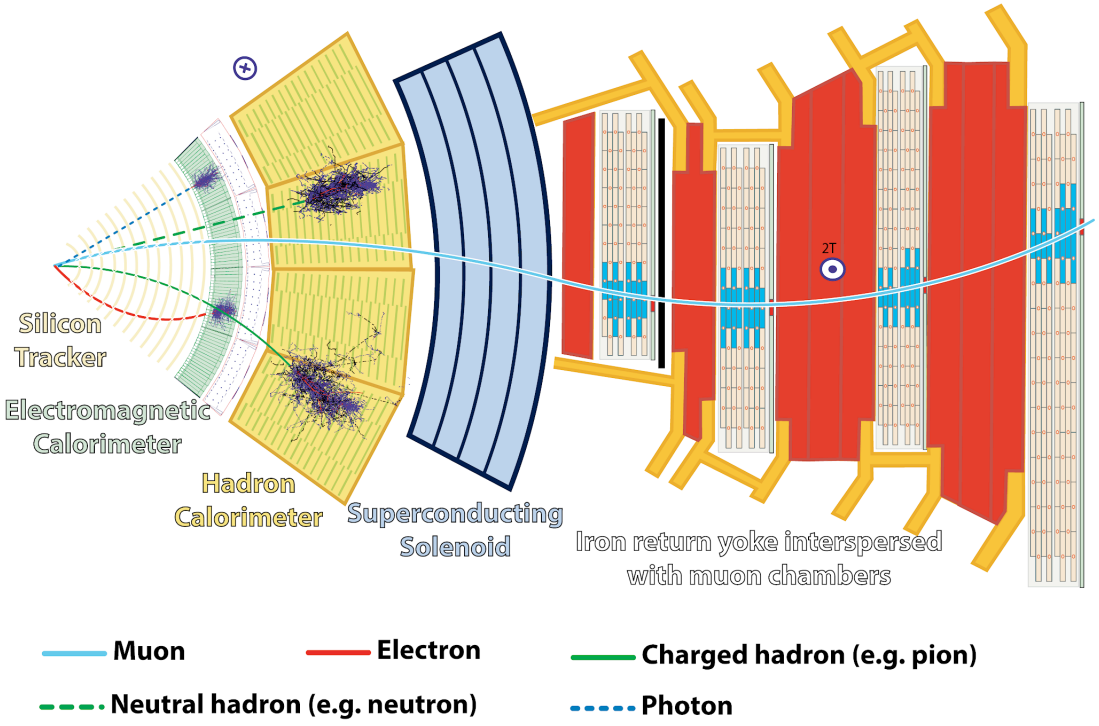
Τα μέρη του CMS που συμβάλουν στην ανίχνευση των προϊόντων κάθε κρούσης είναι:

• **TrackerDetector**: Σύστημα που ανιχνεύει τις τροχιές των σωματιδίων και δίνει πληροφορίες για το είδος και την ορμή των σωματιδίων που προέρχονται από την σύγκρουση των δεσμών.

• **ECalorimeter(ECAL)**: Σύστημα που ανιχνεύει που ανιχνεύει τα φωτόνια και τα ηλεκτρόνια.

• **HCalorimeter(HCAL)**: Αδρονικό θερμιδόμετρο το οποίο ανιχνεύει ουδέτερα αλλά και φορτισμένα αδρόνια. Σε αυτό το σημείο θα παρατηρηθεί μεγάλη αποτύπωση ενέργειας από ένα πίδακα αδρονίων.

• **MuonDetector**: Σύστημα υψηλής απόδοσης για την ανίχνευση των μιονίων.



Σχήμα 1.3: Τα μέρη του ανιχνευτή CMS και τα σημεία που ανιχνεύεται το κάθε είδος υποατομικού σωματιδίου (πηγή εικόνας: The TensorFlow Blog)

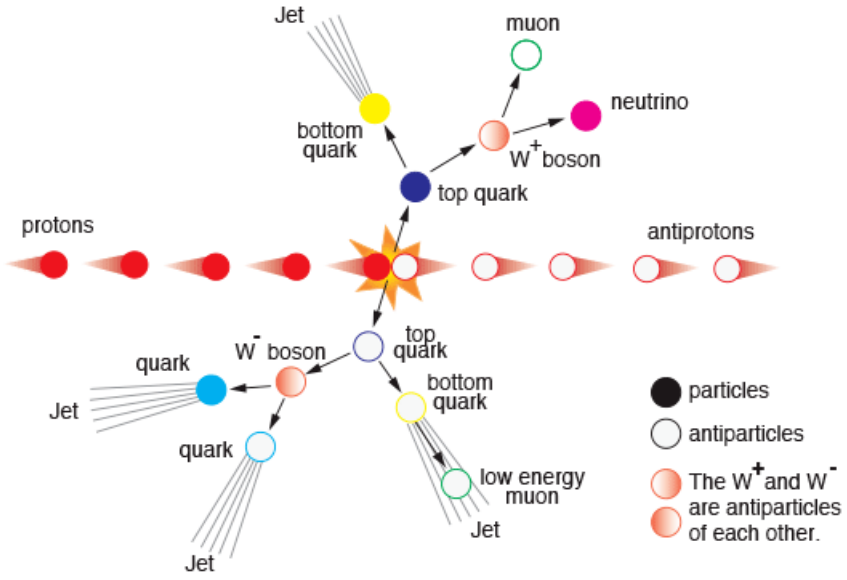
1.4 Πίδακες Αδρονίων

Οι πίδακες αδρονίων είναι ένα σύνολο από αδρόνια το οποίο έχει κάθετη ορμή ($p_\perp$) πολύ μικρότερη από την γραμική ορμή του ($p_\parallel$) ως προς τον άξονα διάδοσης του πίδακα. Η συγκεκριμένη ορολογία έχει επιλεχθεί από την αντίστοιχη για τους πίδακες (jets) υγρών ή αέριων ρευστών ($p_\perp \ll p_\parallel$). Η δημιουργία αυτών των πιδάκων κατά τη σκέδαση σωματιδίων στους επιταχυντές μπορεί να περιγραφεί με ένα απλό μοντέλο, όπως θα φανεί παρακάτω. Κατά τη σύγκρουση σωματιδίων, οι δημιουργούμενοι πίδακες αποτελούν φαινόμενο ενδιαφέροντος και έχει προκύψει η ανάγκη εύρεσης μοντέλων μελέτης τους. Η δυσκολία έγκειται στο ότι πολλές φορές οι πίδακες αδρονίων μπορούν να είναι τόσο πυκνοί και ισχυροί (μεγάλο πλήθος σωματιδίων με πολλές ενέργειες και μεγάλες ορμές), ώστε να καλύπτουν την ύπαρξη άλλων σωματιδίων ή γεγονότων στα πειράματα.

Αρχικά εξετάζεται ο τρόπος παραγωγής των πιδάκων. Είναι γνωστό ότι τα αδρόνια αποτελούνται από κουαρκ, τα οποία αλληλεπιδρούν μεταξύ τους ανταλλάζοντας γλουόνια. ΄Οταν η ισχύς της σύγκρουσης μεταξύ των σωματιδίων (στο σύστημα του Κέντρου Μάζας) είναι μεγάλη, τότε τα κουάρκ διασκεδάζονται προς διάφορες διευθύνσεις, κυρίως κάθετα στην αρχική διεύθυνση της δέσμης. Η ισχυρή αλληλεπίδραση, όμως, μεταξύ των κουαρκ δεν επιτρέπει την απομόνωση τους, όπως περιγράφηκε παραπάνω. Κατά την απομάκρυνση των κουάρκ αυξάνεται η ισχύς της μεταξύ τους αλληλεπίδρασης, η οποία στη συνέχεια γίνεται τόσο ισχυρή, ώστε δημιουργείται ένα νέο ζεύγος κουαρκ-αντικουαρκ, διατηρώντας έτσι το "λευκό" του συνολικού χρώματος που απαιτεί η QCD. Αυτή η διαδικασία επαναλαμβάνεται και έτσι δημιουργείται ένα πλήθος από κουαρκ-αντικουαρκ και γλουόνια (δηλαδή ένα σύνολο από παρτόνια) το οποίο, σε δεύτερη φάση, μπορεί και δημιουργεί με τη σειρά του ένα πλήθος από μεσόνια (ζεύγος κουάρκ-αντικουαρκ) και βαρυόνια (συνδυασμός τριών κουαρκ), τα οποία αποτελούν και τα αδρόνια του πίδακα που παρατηρούνται στον ανιχνευτή αδρονίων που βρίσκεται γύρω από την διεύθυνση της αρχικής δέσμης. Αυτή η διαδικασία παραγωγής αδρονίων από τα κουαρκ και τα γλουόνια ονομάζεται αδρανοποίηση (hadronization).

Από την διαδικασία δημιουργίας των πιδάκων, γίνεται εμφανές ότι η μεγαλύτερη συμβολή

στο μοντέλο παραγωγής τους προέρχεται από την πρώτη φάση δημιουργίας τους, κατά την σκέδαση των κουρκ στο εσωτερικό των συγκρουόμενων αδρονίων. Η σημασία ανάλυσης της συγκεκριμένης φάσης γίνεται ακόμα πιο ξεκάθαρη αναλογιζόμενοι ότι οι πίδακες διατηρούν τα φυσικά χαρακτηριστικά (όπως ορμή, κβαντικούς αριθμούς κλπ.) των αρχικών παρτονίων που τους δημιούργησαν και μάλιστα πιο ολοκληρωμένα από τα μεμονωμένα αδρόνια.[13]



Σχήμα 1.4: Αναπαράσταση της παραγωγής πιδάκων αδρονίων (πηγή εικόνας: electron6.phys.utk.edu)

Ταξινόμηση και Ανακατασκευή Πιδάκων

Σε κάθε προσπάθεια μελέτης των πιδάκων σωματιδίων που παρατηρούμε σε έναν ανιχνευτή μετά από κάθε γεγονός στον επιταχυντή, βρισκόμαστε αντιμέτωποι με 2 μεγάλες προκλήσεις. Η πρώτη είναι ο τρόπος με τον οποίο θα αντιστοιχηθούν οι τροχιές και οι ενέργειες που έχει καταγράψει ο ανιχνευτής σε πίδακα και στη συνέχεια θα ταξινομηθούν η τροχιά, ο τύπος σωματιδίου, η ενέργεια κλπ κάθε στοιχείου του πίδακα. Η δεύτερη πρόκληση είναι να επανακατασκευαστεί η πορεία εξέλιξης του πίδακα, μέχρι τη στιγμή της πρώτης σκέδασης των παρτονίων στις αρχικές δέσμες, ώστε να βγάλουμε χρήσιμα συμπεράσματα για την ενεργό διατομή των σκεδάσεων, τα σωματίδια που παράγονται κ.α.

Ταξινόμηση Πιδάκων

Σχηματικά, οι πίδακες έχουν κωνικό σχήμα, μέσα στους οποίους βρίσκουμε τις τροχιές των σωματιδίων που προέρχονται από την αδρανοποίηση των αρχικών παρτονίων της σκέδασης. Τα σωματίδια αυτά είναι συνήθως μεσόνια ή βαρυόνια, όπως πιόνια, καόνια, πρωτόνια, νετρόνια κλπ. Στο αδρονικό θερμιδόμετρο αυτά θα αποτυπωθούν σαν ένα τοπικό cluster ενέργειας. Τα jets (πίδακες) χωρίζονται σε δύο γενικές κατηγορίες:

- Prompt jets: Πίδακες οι οποίοι προέρχονται από την ίδια κύρια σύγκρουση δεσμών (primary vertex)
- Pileup jets: Πίδακες που προέρχονται από δευτερογενείς συγκρούσεις στο ίδιο crossing των δεσμών, είναι όμως μετατοπισμένες στον άξονα της δέσμης (secondary primary verteces)

Ως κύρια σύγκρουση (primary vertex PV) ορίζεται η κορυφή με το μεγαλύτερο άθροισμα του τετραγώνου των ορμών ($\sum p_T^2$) φορτισμένων τροχιών που συνδέονται με αυτή. Για να ικανοποιείται αυτή η συνθήκη απαιτείται επίσης στην κορυφή να περιέχονται τουλάχιστον τέσσερις τροχιές και η μέγιστη απόσταση από το ονομαστικό σημείο αλληλεπίδρασης να είναι μικρότερο από 24 cm κατά μήκος του άξονα z.

Το φαινόμενο των pileup jets προκαλείται καθώς, σε διατάξεις όπως του LHC, απαιτείται υψηλός ρυθμός συγκρούσεων, άρα και υψηλός ρυθμός συγκρούσεων δεσμών (bunches), για την λήψη μεγάλου αριθμού δεδομένων. Έτσι, γίνονται πολλές συγκρούσεις πρωτονίων ανά bunch crossing, με αποτέλεσμα, εκτός από την κύρια σύγκρουση, να παρατηρούνται πρόσθετες τροχιές σωματιδίων που παράγονται λόγω soft QCD διεργασιών.

Τα pileup jets, μαζί με τον ηλεκτρονικό θόρυβο υποβάθρου, μπορούν να επηρεάσουν την επεξεργασία και να αλλοιώσουν τα τελικά αποτελέσματα. Για τον λόγο αυτό έχουν αναπτυχθεί αλγόριθμοι, όπως ο Pile Up Per Particle Identification Algorithm (PUPPI), που συμβάλουν στις αναγκαίες διορθώσεις, μέχρι και στην ολική αφαίρεση, της επιροής των παραπάνω.

Ανακατασκευή Πιδάκων

Η διαδικασία ανακατασκευής των πιδάκων για την τελική τους ανάλυση ξεκινάει καταγράφοντας την ενέργεια που εναποτίθεται στα θερμιδόμετρα (ηλεκτρομαγνητικό και αδρονικό). Πρώτα αποκόπτονται ορισμένες καταγραφές που είναι κάτω από ένα όριο, διαφορετικό για κάθε σημείο του ανιχνευτή και για κάθε είδος θερμιδόμετρου, ώστε να καθαριστούν τα σήματα από τον θόρυβο. Στη συνέχεια η εναποτιθέμενη ενέργεια σε κάθε “χώρο” του ανιχνευτή συνδυάζεται σε “πύργους”. Πρόκειται ουσιαστικά για ένας είδος τρισδιάστατων ιστογραμμάτων όπου αποτελούνται από 1 κελί του αδρονικού θερμιδόμετρου και 3x3 κελιά του ηλεκτρομαγνητικού. Θα εφαρμοστεί ξανά μια αποκοπή δεδομένων κάτω από ένα όριο ενέργειας ($E_T = 0.5MeV$) ώστε να αποκλιστούν πιθανές ενέργειες από pileup jets και άλλα γεγονότα που μπορεί να συνέβησαν εντός των θερμιδόμετρων. Πριν την είσοδο των towers στους αλγορίθμους ανάλυσης των πιδάκων, γίνεται ακόμα ένας διαχωρισμός για τις ενέργειες κάτω των $10MeV$ καθώς jets με χαμηλές ενέργειες τείνουν να μην είναι καλά ορισμένα. Στην παρούσα εργασία θα παρουσιάσουμε 2 από τους βασικότερους αλγορίθμους ανάλυσης των jets.

Η λογική πίσω από κάθε αλγόριθμο ανακατασκευής ενός πίδακα είναι, ακολουθώντας μια αντίθετη πορεία από αυτή της εξέλιξης του jet, να προβλέψει το αρχικό σωματίδιο από το οποίο προέκυψε ένα ζεύγος σωματιδίων, αξιοποιώντας το γεγονός ότι οι περισσότερες διακλαδώσεις μια διαδικασίας QCD είναι συγγραμικές.

Ο αλγόριθμος k_t και $anti-k_t$

Πρόκειται για 2 αλγορίθμους της κατηγορίας αλγορίθμων Σειριακής Ομαδοποίησης (Sequencial Clustering Algorithms). Τα jets με αυτούς τους αλγορίθμους ομαδοποιούνται με βάση τις ορμές τους. Η γενική μορφή αυτού του είδους αλγορίθμου περιγράφεται ως εξής:

- Εισάγουμε την απόσταση μεταξύ των καταγραφών (σωματίδια, ψευτοπίδακες κλπ) i και j :

$$d_{ij} = min\{k_{ti}^{2p}, k_{tj}^{2p}\} \cdot \frac{\Delta_{ij}^2}{R^2}$$

- Όπως επίσης και την απόσταση μεταξύ του αντικειμένου i και της δέσμης B:

$$d_{iB} = k_{ti}^{2p}$$

Όπου $\Delta_{ij}^2 = (y_i - y_j)^2 + (\varphi_i - \varphi_j)^2$, k_{ti} είναι η αντίστοιχη εγκάρσια (ως προς τη δέσμη) ορμή (transverse momentum) και τα y_i, φ_i η ωκύτητα (rapidity) και το αζιμούθιο κάθε καταγραφής i. Το R είναι η παράμετρος της ακτίνας. Για $p = 1$ παίρνουμε τον αλγόριθμο k_t ενώ για $p = 0$ παίρνουμε την ειδική περίπτωση του αλγορίθμου Cambridge/Aachen. Ο $anti - k_t$ αλγόριθμος ορίζεται για $p = -1$.

Τα βήματα του αλγορίθμου είναι:

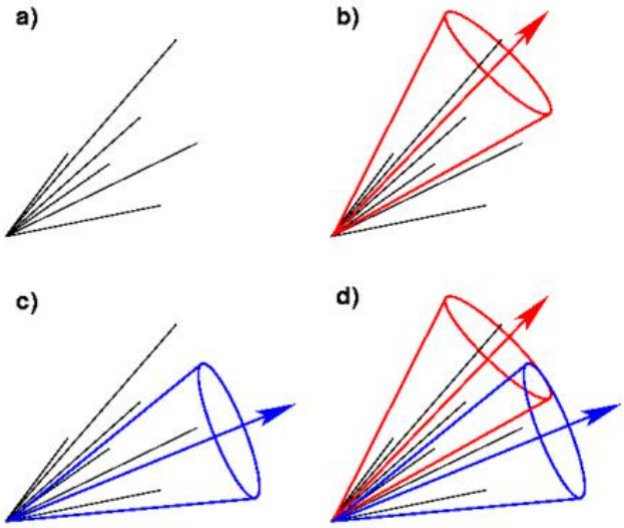
1. Βρίσκουμε την μικρότερη απόσταση από τις d_{ij}, d_{iB}

2. Αν είναι η d_{ij}, προσθέτουμε τις ορμές, ανανεώνουμε τις αποστάσεις και επαναλαμβάνουμε.

3. Αν είναι η d_{iB}, τότε ονομάζουμε το στοιχείο i πίδακα και το αφαιρούμε από την λίστα των καταγραφών-στοιχείων.

4. Επαναλαμβάνουμε από το βήμα 1 έως ότου να μην μείνει κανένα στοιχείο-καταγραφή.

Η παραπάνω οικογένεια αλγορίθμων χρησιμοποιείται ευρέως στα πειράματα LEP και HERA. Η απόδοση τους, ειδικά αυτή του αλγορίθμου $anti - k_t$, προτιμάται στα περισσότερα πειράματα καθώς έχει γραμμική ανταπόκριση στα ελαφριά σωματίδια αλλά είναι κι όλας πιο αποδοτική και ασφαλής στις διακυμάνσεις του ανιχνευτή.



Σχήμα 1.5: Αναπαράσταση της λειτουργίας ενός Sequential Clustering αλγορίθμου

Οι Cone Type αλγόριθμοι

Αυτού του είδους οι αλγόριθμοι, αξιοποιούν περισσότερο γεωμετρικά χαρακτηριστικά στην ανίχνευση των ενεργειών πάνω στους ανιχνευτές. Υπάρχουν δύο είδη τέτοιων αλγορίθμων, ο Iterative Cone Algorithm και ο SISCone Algorithm. Τα βήματα του πρώτου είναι:

1. Βρίσκουμε το πιο ενεργητικό σωματίδιο σε ένα γεγονός και αυτό γίνεται το SEED

2. Ορίζουμε έναν κώνο ακτίνας R γύρω από το SEED και προσθέτοντας όλες τις ορμές εντός του κώνου έχουμε το TRIAL JET

3. Συγκρίνουμε τον άξονα του TRIAL JET με αυτόν του SEED

4. Αν είναι όμοιοι, εντός μια δεδομένης ακρίβειας, τότε TRIAL JET γίνεται STABLE CONE που γίνεται το σταθερό JET

5. Διαφορετικά, το TRIAL JET γίνεται το νέο SEED και επαναλαμβάνεται η διαδικασία μέχρι να υπάρξει σύγκλιση.

6. Η διαδικασία επαναλαμβάνεται έως ότου δεν υπάρχουν SEEDS πάνω από ένα όριο. Για το πείραμα στο CMS αυτό το όριο είναι το 1 GeV.

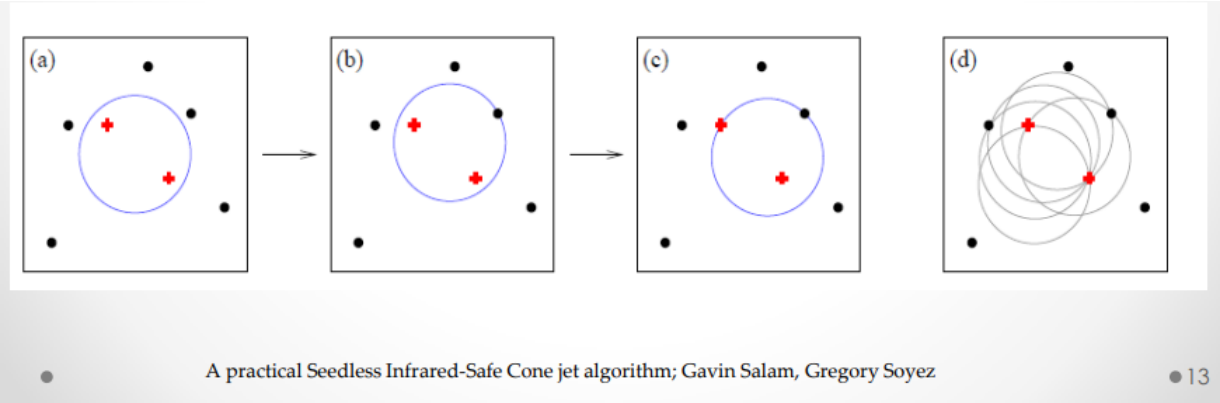
Ο αλγόριθμος SISCone αποτελεί μια βελτιωμένη έκδοση του παραπάνω αλγορίθμου, ενώ αποτελεί και τον βασικό cone-type αλγόριθμο στο CMS. Η διαδικασία που ακολουθεί μοιάζει με τον παραπάνω αλγόριθμο, με τις εξής διαφορές:

1. Αρχικά μετακινούμε ένα κυκλικό όριο σε τυχαίες κατευθύνσεις μέχρι να να βρεθεί ένα σωματίδιο στην περιφέρεια του.

2. Ύστερα περιστρέφουμε τον κύκλο, με άξονα περιστροφής το σωματίδιο, μέχρι να βρεθεί και δεύτερο σωματίδιο στην περιφέρεια του.

3. Ορίζουμε όλους τους κύκλους που ορίζονται από δύο σωματίδια στην περιφέρεια τους σαν STABLE CONEs.

Τα βασικά πλεονεκτήματα αυτού του αλγορίθμου είναι ότι μπορεί χωρίς την χρήση κάποιου SEED και σε ικανοποιητικό υπολογιστικό χρόνο, να βρεί πιθανώς όλους τους σταθερούς κώνους για την ανάλυση του πίδακα. Επίσης ο συγκεκριμένος αλγόριθμος είναι πιο ανθεκτικός στις διακυμάνσεις εντος του ανιχνευτή.[12]



Σχήμα 1.6: Αναπαράσταση της λειτουργίας του SISCone αλγορίθμου

Κεφάλαιο 2

Η Μηχανική Μάθηση

2.1 Περιεχόμενο, είδη και σκοπός

Η μηχανική μάθηση είναι ένας υποκλάδος της Τεχνητής Νοημοσύνης (Artificial Intelligence-AI), ο οποίος έχει σαν σκοπό υπολογιστικά συστήματα να εκπαιδεύονται ώστε να κάνουν προβλέψεις ή να παίρνουν αποφάσεις χωρίς να είναι προγραμματισμένα για συγκεκριμένα προβλήματα με αυστηρά ορισμένες παραμέτρους. Για τη μηχανική μάθηση είναι απαραίτητη η ανάπτυξη αλγορίθμων και μοντέλων που θα μπορούν να αναλύσουν μεγάλο όγκο δεδομένων ώστε να αναγνωρίσουν σε αυτόν ακολουθίες και πρότυπα, να βγάλουν συμπεράσματα και να πάρουν ακριβείς αποφάσεις ή να προβούν σε ασφαλείς προβλέψεις. Η βασική ιδέα πίσω από τη μηχανική μάθηση είναι να επιτραπεί στα υπολογιστικά συστήματα να μάθουν εμπειρικά ή μέσω παραδειγμάτων και να βελτιώνουν κάθε φορά την απόδοση τους. Αντί να δίνονται συγκεκριμένες οδηγίες στο σύστημα για το πώς να εκτελεί συγκεκριμένες εργασίες, οι αλγόριθμοι μηχανικής μάθησης μαθαίνουν από τα δεδομένα αναγνωρίζοντας ακολουθίες και συσχετισμούς μεταξύ τους. Αυτοί οι αλγόριθμοι αξιοποιούν τα συμπεράσματα από την παραπάνω διαδικασία για να κάνουν προβλέψεις, να ταξινομήσουν νές δεδομένα ή για να εκτελέσουν άλλες εργασίες. Υπάρχουν διάφορα είδη αλγορίθμων μηχανικής μάθησης:

- Μάθηση υπό επίβλεψη (Supervised Learning): Σε αυτό τον τρόπο εκπαίδευσης το επιθυμητό αποτέλεσμα είναι γνωστό, επομένως στον αλγόριθμο εισάγονται τα δεδομένα και αυτός με τη σειρά του μαθαίνει να τα αντιστοιχεί κατάλληλα στο αποτέλεσμα-στόχο, βρίσκοντας κατάλληλες ακολουθίες και σχέσεις μεταξύ των δεδομένων εκπαίδευσης (training set). Στη συνέχεια, έχοντας ολοκληρώσει την παραπάνω διαδικασία, μπορεί να κάνει προβλέψεις για νέα, άγνωστα σε αυτόν δεδομένα. Αυτή είναι και η γενική μέθοδος που θα χρησιμοποιηθεί στην παρούσα εργασία.

- Μάθηση χωρίς επίβλεψη (Usupervised Learning): Εδώ ο αλγόριθμος δεν γνωρίζει τα στοιχεία του επιθυμητού αποτελέσματος-στόχου, προσπαθεί να βρεί όμως ακολουθίες και δομές στα δεδομένα που του εισάγουμε μέσω διάφορων μεθόδων όπως ομαδοποίηση όμοιων δεδομένων ή μείωση της διάστασης τους.

- Ενισχυμένη μάθηση (Reinforced Learning): Αυτού του είδους η μέθοδος χρησιμοποιεί "ποινές" και "επιβραβεύσεις" κατά την αλληλεπίδραση του αλγορίθμου με κάποιο περιβάλλον ώστε να οδηγήσει τελικά στην βελτίωση του και στην διαμόρφωση ενός βέλτιστου αλγορίθμου.

- Βαθιά μάθηση (Deep Learning): Η Deep Learning μέθοδος είναι ουσιαστικά υποκατηγορία των μεθόδων μηχανικής μάθησης και επικεντρώνεται στη χρήση τεχνητών νευρωνικών δικτύων (Artificial Neural Networks), λαμβάνοντας έμπνευση από τον τρόπο δομής και λειτουργίας του ανθρώπινου εγκεφάλου. Τα μοντέλα που προκύπτουν από αυτή τη μέθοδο

είναι ικανά να μαθαίνουν πολύπλοκες δομημένες αναπαραστάσεις δεδομένων εφαρμόζοντας πολλαπλά στρώματα από συνδεδεμένους κόμβους (interconnected nodes).

Ένας άλλος τρόπος διαχωρισμού των μεθόδων μηχανικής μάθησης είναι από το είδος του επιθυμητού αποτελέσματος και διακρίνονται οι εξής κατηγορίες:

- Ταξινόμηση (Classification): Οι αλγόριθμοι ταξινόμησης χρησιμοποιούν ήδη κατηγοριοποιημένα δεδομένα και στη συνέχεια προβλέπουν την κατηγοριοποίηση (ταξινόμηση) των νέων, άγνωστων.

- Παλινδρόμηση (Regression): Οι συγκεκριμένοι αλγόριθμοι προβλέπουν την συμπεριφορά από συνεχείς αριθμητικές τιμές βασιζόμενες στη συμπεριφορά των εισαγόμενων δεδομένων. Εκπαιδεύονται μέσω της σχέσης των εισαγόμενων δεδομένων και των αντίστοιχων εξαγόμενων τιμών.

- Συσταδοποίηση (Clustering): Αυτού του είδους αλγόριθμοι στοχεύουν στην ομαδοποίηση παρόμοιων δεδομένων, βασιζόμενοι όμως σε πιθανά κοινά χαρακτηριστικά μεταξύ τους που μπορεί να αναγνωριστούν και όχι σε κάποιο γνωστό, τελικό σύνολο, γι'αυτό και η συγκεκριμένη μέθοδος αποτελεί τυπική εργασία μάθησης χωρίς επίβλεψη.

- Εκτίμηση Πυκνότητας (Density Estimator): Αυτή η μέθοδος βρίσκει την κατανομή των δεδομένων εισόδου σε κάποιο χώρο.

- Μείωση Διάστασης (Dimensionality Reduction): Βασικός τρόπος λειτουργίας της συγκεκριμένης μεθόδου είναι η μείωση του αριθμού των παραμέτρων ή των μεταβλητών από τα δεδομένα εισόδου χωρίς να χαθεί σημαντική πληροφορία. Βοηθούν την απλοποίηση της παρουσίασης των δεδομένων, την καλύτερη χρήση της υπολογιστικής δύναμης και την μείωση πιθανού θορύβου.

Οι μέθοδοι της Παλινδρόμησης και της Ταξινόμησης είναι οι δύο κύριες και πιο ευρέως χρησιμοποιούμενες μέθοδοι μηχανικής μάθησης. Με την ταξινόμηση κατηγοριοποιούμε αντικείμενα σε διακριτά σύνολα, όπως για παράδειγμα συγκεκριμένα χαρακτηριστικά σε εικόνες (αναγνώριση προσώπων σε φωτογραφίες/βίντεο) ή αναγνώριση γραφικού χαρακτήρα για ψηφιοποίηση χειρογράφων. Με την παλινδρόμηση προβλέπουμε την συμπεριφορά συνεχών μεταβλητών, όπως η τιμή ενός ακινήτου, η κατανάλωση ενέργειας ή η ενέργεια-ορμή ενός σωματιδίου. Γίνεται προφανές πως στην συγκεκριμένη εργασία θα χρησιμοποιήσουμε την μέθοδο της παλινδρόμησης, καθώς θέλουμε να διερευνηθεί η σχέση μεταξύ ανεξάρτητων μεταβλητών (ορμή, αζιμούθια γωνία κλπ) και της εξαρτημένης μεταβλητής της εγκάρσιας ορμής του αδρονικού πίδακα.

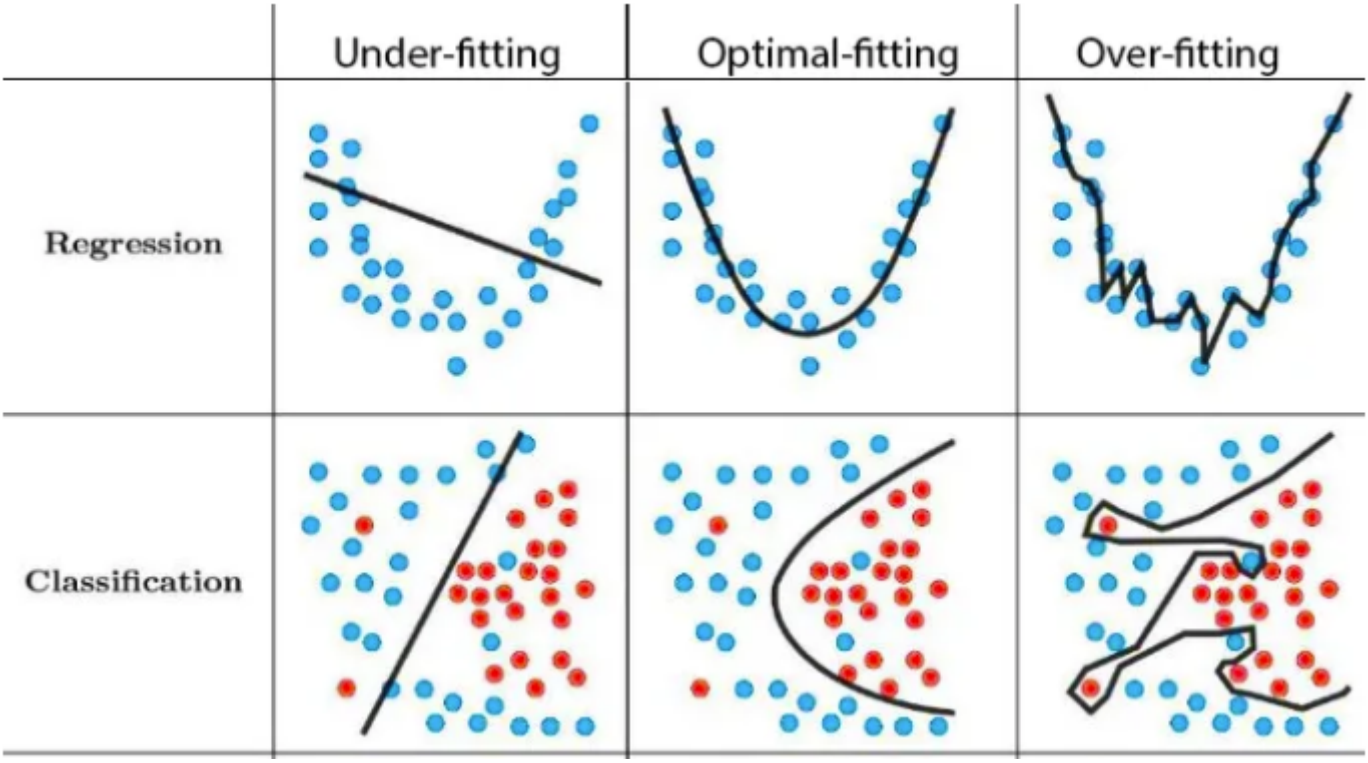
Επίσης, θα χρησιμοποιήσουμε μέθοδο εκμάθησης με επίβλεψη, χρησιμοποιώντας σαν τελικά δεδομένα-στόχο για την τετραορμή των πιδάκων αυτά που μας έδωσε η εξομοίωση της σύγκρουσης $t - \bar{t}$ μέσω Monte Carlo προσομοιώσεων. Καθώς όμως η εκμάθηση θα είναι με επίβλεψη, θα δοθεί ιδιαίτερη προσοχή ώστε τα δεδομένα για την εκμάθηση να επιλεχθούν σωστά σε σχέση με τον συνολικό όγκο δεδομένων, καθώς διαφορετικά ενδέχεται το μοντέλο πρόβλεψης που θα προκύψει να είναι υπερβολικά προσαρμοσμένο στα δεδομένα που το εκπαιδεύσαμε αλλά που δεν αντιπροσωπεύουν νέα δεδομένα, επομένως θα προκύπτουν ανακριβείς προβλέψεις μόλις αναπτυχθεί το μοντέλο και γενικευτεί. Το τελευταίο αποτελεί το φαινόμενο του Overfitting-Overtraining και θα σχολιαστεί στο επόμενο κεφάλαιο.[4, 3, 2]

2.2 Υπερεκπαίδευση (Overtraining)

Η υπερεκπαίδευση ενός συνόλου δεδομένων προκύπτει σε περίπτωση που το μοντέλο το οποίο προκύπτει από τον αλγόριθμο της μηχανικής εκμάθησης είναι τόσο σύνθετο ώστε να φτάνει να έχει την ικανότητα να προσομοιώνει τόσο μικρές διαφοροποιήσεις στα δεδομένα όσο

αυτές που προκύπτουν από τον θόρυβο στο δείγμα του συνόλου. Ουσιαστικά κατά την υπερεκπαίδευση στοιχεία του υποβάθρου έχουν ληφθεί σαν στοιχεία της υποκείμενης δομής του συνόλου εκπαίδευσης.[3] Υπερεκπαίδευση μπορεί να έχουμε όμως ακόμα κι όταν δεν υπάρχει θόρυβος στα δεδομένα, στην γενική περίπτωση όπου το μοντέλο ουσιαστικά δεν "μαθαίνει" από τα χαρακτηριστικά του δείγματος αλλά τα "απομνημονεύει". Για παράδειγμα, εάν ο αριθμός των παραμέτρων είναι ίδιος ή μεγαλύτερος από τον αριθμό των παρατηρήσεων, τότε ένα μοντέλο μπορεί να προβλέψει τέλεια τα δεδομένα εκπαίδευσης απλώς απομνημονεύοντας τα δεδομένα στο σύνολό τους. Σε κάθε περίπτωση, το βασικό πρόβλημα που θα προκύψει είναι ότι το μοντέλο δεν θα μπορέσει να δώσει μια ικανοποιητική εκτίμηση/πρόβλεψη όταν εισαχθούν νέα δεδομένα, ενώ εμφανίζει επίσης ψευδή υψηλή απόδοση στις δοκιμές (test set) μετά την εκπαίδευση του.

Η πιθανότητα εμφάνισης του φαινομένου της υπερεκπαίδευσης (overtraining/overfit) εξαρτάται από τον αριθμό των παραμέτρων και των δεδομένων, το επίπεδο θορύβου ή σήματος υποβάθρου, αλλά και από την δομή του μοντέλου μηχανικής μάθησης που εφαρμόζεται σε κάθε περίπτωση (πχ. Boosted Decision Tree, Neural Network κλπ). Γενικότερα πρέπει να λάβουμε υπ' όψιν πως κάθε μοντέλο, όσο καλά και να έχει εκπαιδευτεί, θα αποδίδει λιγότερο καλά σε ένα νέο σύνολο δεδομένων από ό,τι στο σύνολο εκπαίδευσης.

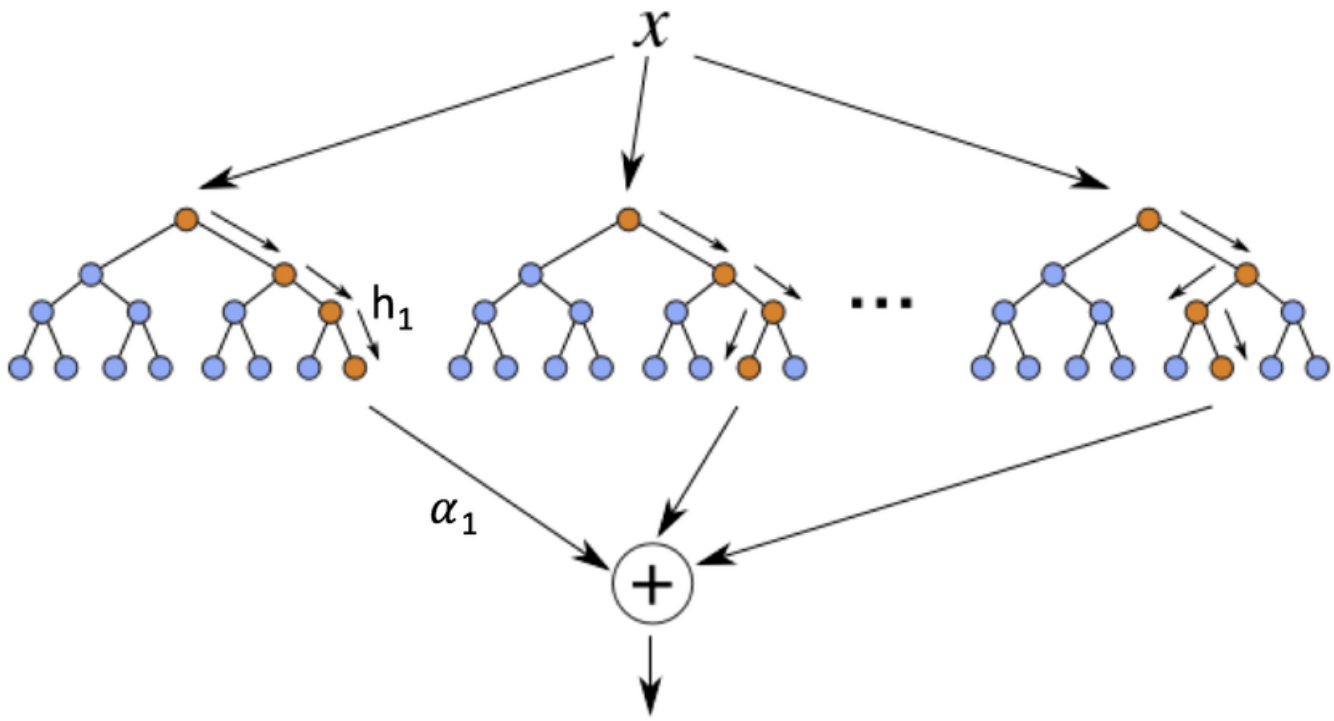


Σχήμα 2.1: (πηγή εικόνας: towardsdatascience.com)

Μία πρακτική αντιμετώπισης του φαινομένου της υπερεκπαίδευσης είναι η κανονικοποίηση (regularization), η οποία προσθέτει κάποιο ποινή (penalty) στην συνάρτηση υπολογισμού του σφάλματος απόκλισης σε κάθε βήμα. Μία δεύτερη πρακτική, η οποία θα χρησιμοποιηθεί και στην συγκεκριμένη εργασία, είναι να χωρίζουμε το σύνολο των δεδομένων σε δύο μέρη, ένα μέρος από το σύνολο για την εκπαίδευση του μοντέλου (training set) και το άλλο μέρος για να δοκιμαστεί η ικανότητα γενίκευσης του μοντέλου που προέκυψε (testing set).

2.3 Ενισχυμένα Δέντρα Απόφασης (Boosted Decision Trees)

Τα (ενισχυμένα) δέντρα απόφασης (BDTs) είναι από τις πιο διαδεδομένες και ταυτόχρονα πρακτικές μεθόδους μηχανικής μάθησης. Η μέθοδος που ακολουθούν φτιάχνει μία δυαδική δομή δέντρου. Ξεκινώντας από τη "ρίζα"(root) και θέτοντας επαναλαμβανόμενες αποφάσεις "ναι/όχι"(αριστερά/δεξιά) σε μία μεμονωμένη μεταβλητή φτιάχνονται τα "κλαδιά"(nodes) του δέντρου, τα οποία καταλήγουν σε μία πιθανή τιμή για την μεταβλητή, τα "φύλλα"(leaves) του δέντρου. Η διαδικασία επαναλαμβάνεται μέχρις ότου ικανοποιηθεί κάποιο κριτήριο διακοπής. Ο χώρος φάσης χωρίζεται με αυτόν τον τρόπο σε πολλές περιοχές και τελικά τα γεγονότα ταξινομούνται ως σήμα ή υπόβαθρο επάνω στον τελευταίο κόμβο του δέντρου (φύλλο).[2] Συγκεκριμένα στα δέντρα παλινδρόμησης (regression BDTs), στόχος είναι να προβλεφθεί η τιμή μίας μόνο μεταβλητής από ένα N-διάστατο διάνυσμα $\boldsymbol{x} = (x_1, ..., x_N)^T$ μεταβλητών εισόδου. Τα δεδομένα για την εκπαίδευση αποτελούνται από διανύσματα $\{x_1, ..., x_K\}$ με τις αντίστοιχες ετικέτες $\{t_1, ..., t_K\}$. Αν δίνεται ο διαχωρισμός του χώρου των φάσεων και ελαχιστοποιηθεί η μέση τιμή του τετραγώνου του σφάλματος, τότε η βέλτιστη τιμή για συγκεκριμένη μεταβλητή σε οποιαδήποτε περιοχή δίνεται από τη μέση τιμή των τιμών t_k εκείνων των σημείων που ανήκουν σε αυτή την περιοχή. Στο τέλος της διαδικασίας κάθε φύλλο αντιπροσωπεύει μια συγκεκριμένη τιμή αυτης της μεταβλητής στόχου.



Σχήμα 2.2: Αναπαράσταση ενός "δάσους" από δέντρα απόφασης (πηγή εικόνας: Research Gate)

Η απόδοση των δέντρων απόφασης βελτιώνεται δραματικά μέσω της διαδικασίας της ενίσχυσης (boosting) ενός δέντρου. Η διαδικασία αυτή επεκτείνει την διαδικασία που περιγράφηκε παραπάνω από ένα δέντρο σε πολλά δέντρα μαζί (δάσος). Συγκεκριμένα, ο πρώτος ταξινομητής είναι ένα δέντρο χαμηλής απόδοσης, ορίζοντας ένα στατιστικό βάρος σε κάθε τιμή από τα δεδομένα του συνόλου εκπαίδευσης. Αρχικά όλα τα βάρη θα είναι ίσα. Πριν ξεκινήσει οποιαδήποτε διαδικασία εκμάθησης το επόμενο δέντρο απόφασης, θα αυξήσει την τιμή του βάρους για τα δεδομένα που δεν έχουν κατηγοριοποιηθεί σωστά ενώ θα μειώσει την αντίστοιχη τιμή για τα σωστά ταξινομημένα δεδομένα. Αυτό θα επαναλαμβάνεται από κάθε δέντρο πριν την εκτέλεση του. Επομένως κάθε επόμενος ταξινομητής αναγκάζεται να δώσει έμφαση σε δεδομένα που τα-

ξινομούνται εσφαλμένα. Η τελική πρόβλεψη λαμβάνεται συνδυάζοντας τα αποτελέσματα από κάθε δέντρο. Η συμβολή κάθε δέντρου ορίζεται από την απόδοση του κατά την διαδικασία εκπαίδευσης του, δηλαδή τα πιο ακριβή δέντρα επηρεάζουν περισσότερο το τελικό αποτέλεσμα από ότι αυτά που δεν απέδωσαν τόσο καλά. Η ενίσχυση σταθεροποιεί την απόκριση των δέντρων απόφασης σε σχέση με τις διακυμάνσεις στο δείγμα εκπαίδευσης και είναι σε θέση να μειώσει τα φαινόμενα της υπερεκπαίδευσης.

Εκπαίδευση των BDTs

Για να εκπαιδεύσουμε ή να κατασκευάσουμε ένα δέντρο απόφασης πρέπει ουσιαστικά να καθορίσουμε τα κριτήρια διαχωρισμού κάθε κόμβου. Ξεκινώντας από τη "ρίζα" (πρώτος κόμβος) χωρίζουμε, βάσει κάποιου αρχικού κριτηρίου, το συνολικό δείγμα εκπαίδευσης. Η ίδια διαδικασία θα επαναληφθεί, χωρίζοντας κάθε φορά κάποιο από τα υποσύνολα σε άλλα 2 υποσύνολα, κατασκευάζοντας έτσι ολόκληρο το δέντρο. Σε κάθε κόμβο (node) επιλέγεται κατάλληλο κριτήριο και επιλέγεται η βέλτιστη μεταβλητή διαχωρισμού ώστε να διασφαλίζεται η ομοιογένεια των δεδομένων ως προς τη μεταβλητή-στόχο. Καθώς έχει βρεθεί εμπειρικά ότι δεν μπορεί να υπάρξει ένας γενικός κανόνας για το μετά από πόσους κόμβους το σφάλμα θα πέσει κάτω από μια συγκεκριμένη τιμή, η διαδικασία επέκτασης των κόμβων του δέντρου σταματάει μόλις το δέντρο φτάσει στο επιθυμητό όριο που έχει καθορίσει αρχικά ο χρήστης σχετικά με τις παραμέτρους Μέγιστο Βάθος (Max Depth) και Αριθμός Δέντρων (Number of trees-NTrees).

Παρακάτω παρουσιάζονται μερικά κριτήρια διαχωρισμού για την αξιολόγηση μιας μεταβλητής, τα οποία μπορούν να εφαρμοστούν για κοπή σε έναν κόμβο. Από δοκιμές δεν έχουν προκύψει αξιόλογες διαφορές στην μεταξύ τους απόδοση. Ο δείκτης καθαρότητας p είναι ίσος με $0,5$ όταν έχουμε πλήρως μπερδεμένα δείγματα, ενώ μηδέν όταν το δείγμα αποτελείται από μόνο μία κλάση.

- Gini Index: $Q = p(1-p)$
- Cross Entropy: $Q = p \cdot lnp$
- Average squared error (Μέσο Τετραγωνικό Σφάλμα): $Q = \frac{1}{N}\sum\{t-y\}^2$

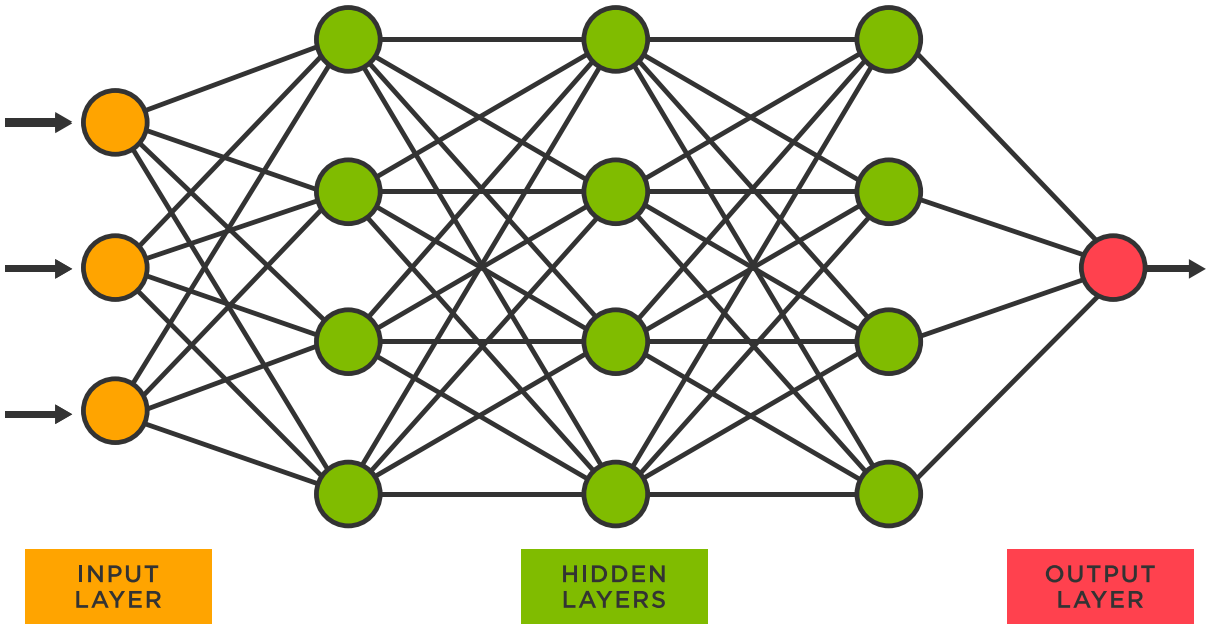
όπου t είναι η τιμή του στόχου της παλινδρόμησης για κάθε γεγονός και y η μέση τιμή από όλα τα γεγονότα στον κόμβο.

Η ιδιότητα των δέντρων απόφασης να χρειάζονται παρέμβαση σε λίγες από τις παραμέτρους τους για να πετύχουν σχετικά ακριβή αποτελέσματα τα έχει κάνει να θεωρούνται οι καλύτεροι "out of the box" ταξινομητές. Η παραπάνω απλότητα οφείλεται ότι σε κάθε κόμβο διάσπασης (node) χρειάζεται μόνο μία μονοδιάστατη βελτιστοποίηση κοπής του συνόλου. Επίσης οι μεταβλητές που δεν έχουν καλή διακριτική ικανότητα δεν επηρεάζουν τόσο το τελικό αποτέλεσμα. Παρόλα αυτά, τα BDTs τείνουν συχνά να εμφανίζουν το φαινόμενο της υπερεκπαίδευσης, ενώ εμφανίζουν μεγάλη ευαισθησία σε δεδομένα θορύβου και είναι πολύ απαιτητικά σε υπολογιστικούς πόρους. [2]

2.4 Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks)

Ο όρος "νευρωνικό δίκτυο" έχει τις ρίζες του στην προσπάθεια εύρεσης μαθηματικής διατύπωσης της πληροφορίας που επεξεργάζονται βιολογικά συστήματα. Στην πραγματικότητα "νευρωνικά δίκτυα" αποκαλούνται ένα σύνολο από συναρτήσεις που εξάγουν μοντέλα από σύνολα δεδομένων. Η αναπαράσταση τους γίνεται συμβολίζοντας με κουτιά την τοποθεσία όπου βρίσκονται τα δεδομένα εισόδου και με κύκλους τους νευρώνες. Κάθε νευρώνας είναι υπεύθυνος για την εκτέλεση μιας συνάρτησης, κατά την οποία θα δεχθεί έναν αριθμό από τιμές εισόδου

και θα τις αντιστοιχίσει σε κάποιες τιμές εξόδου. Τα βέλη συμβολίζουν ότι η έξοδος από έναν νευρώνα αποτελεί την είσοδο για έναν άλλον.[3] Εφαρμόζοντας ένα εξωτερικό σήμα στους νευρώνες εισόδου το δίκτυο τίθεται σε μια καθορισμένη κατάσταση που μπορεί να μετρηθεί από την απόκριση των νευρώνων εξόδου. Μπορούμε λοιπόν να φανταστούμε ένα νευρωνικό δίκτυο ως μία αντιστοίχιση από έναν χώρο μεταβλητών εισόδου $\{x_1, ..., x_n\}$ σε έναν χώρο μεταβλητών εξόδου $\{y_1, ..., y_n\}$. Η αντιστοίχιση μπορεί να είναι μη γραμμική εάν τουλάχιστον ένας νευρώνας έχει μη γραμμική ανταπόκριση στη συμβολή του.



Σχήμα 2.3: Αναπαράσταση της διαδικασίας ροής και επεξεργασίας των δεδομένων από ένα Τεχνητό Νευρωνικό Δίκτυο (πηγή εικόνας: TIBCO Software)

Από τα παραπάνω προκύπτει ότι ένα δίκτυο με n νευρώνες μπορεί να διαθέτει n^2 συνδέσεις, κάνοντας το αρκετά πολύπλοκο. Στην συγκεκριμένη εργασία θα χρησιμοποιηθεί ένα είδος νευρωνικού δικτύου που ονομάζεται Multilayer Perseptron και έχει την ιδιότητα να μειώνει κατά πολύ την πολυπλοκότητα του δικτύου. Αυτό επιτυγχάνεται οργανώνοντας τους νευρώνες σε στρώματα (layers) και επιτρέποντας μόνο άμμεσες συνδέσεις από το ένα στρώμα νευρώνων στο άλλο. Τα δεδομένα εισόδου αποτελούν το στρώμα εισόδου του νευρωνικού δικτύου (input layer), ενώ το τελευταίο στρώμα θα είναι το στρώμα εξόδου (output layer), με όλα τα ενδιάμεσα στρώματα να είναι κρυφά στρώματα (hidden layers).

Η διαδικασία επεξεργασίας των δεδομένων από κάθε νευρώνα αφορά την συνάρτηση απόκρισης του, δηλαδή τον συνδυασμό μίας συνάρτησης σύναψης και μίας συνάρτησης ενεργοποίησης. Η πρώτη λειτουργεί ως εξής. Ο νευρώνας δέχεται n στοιχεία εισόδου $[x_1, ..., x_n]$ από n διαφορετικές εισερχόμενες συνδέσεις, έχοντας κάθε μία ένα διαφορετικό βάρος $[w_1, ..., w_n]$. Το άθροισμα συναρτήσει του κάθε βάρους γίνεται πολλαπλασιάζοντας την είσοδο με τα βάρη και προσθέτοντας μεταξύ τους τα γινόμενα.

$$z = \sum_{i=1}^{n} x_i \times w_i$$

Το επόμενο στάδιο επεξεργασίας των δεδομένων από τον νευρώνα είναι η προώθηση της παραπάνω υπολογισμένης τιμής z μέσα από μια συνάρτηση ενεργοποίησης. Οι πιο συνηθισμένες συναρτήσεις ενεργοποίησης είναι :

- Γραμμική

$$\alpha(x) = x$$

- Σιγμοειδής

$$\alpha(x) = \frac{1}{1+e^{-kx}}$$

- Υπερβολική Εφαπτομένη

$$\alpha(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$$

- Ακτινική

$$\alpha(x) = e^{-\frac{x^2}{2}}$$

Εκπαίδευση Νευρωνικού Δικτύου (NN)-Η Οπισθοδρόμηση (Back-propagation)

Ο πιο συνηθισμένος αλγόριθμος για τον υπολογισμό των βαρών για την βελτιστοποίηση της απόδοση ενός νευρωνικού δικτύου είναι η οπισθοδρόμηση (Back-Propagation). Ανήκει στην κατηορία των μεθόδων μάθησης υπό επίβλεψη, όπου, όπως αναφέρθηκε προηγουμένως, είναι γνωστή η επιθυμητή έξοδος για κάθε συμβάν εισόδου. Ο αλγόριθμος της οπισθοδρόμησης χρησιμοποιείται για να εκπαιδεύσει νευρωνικά δίκτυα με κρυμμένα στρώματα νευρώνων (Multilayer Perceptrons). Αρχικά ο αλγόριθμος ορίζει τυχαία βάρη σε κάθε σύνδεση στο νευρωνικό δίκτυο. Ύστερα η μέθοδος συνεχίζει κανονικά από αριστερά προς τα δεξιά, με τα δεδομένα εισόδου να περνούν μέσα από τους νευρώνες και να καταλήγουν στους νευρώνες εξόδου (forward pass). Στη συνέχεια ο αλγόριθμος ξεκινάει από τους νευρώνες εξόδου και πηγαίνει προς τα πίσω, μέχρι να φτάσει τους νευρώνες εισόδου (backward pass). Κατά τη διάρκεια αυτού του βήματος, ο αλγόριθμος υπολογίζει ένα σφάλμα από κάθε νευρώνα σε σχέση με τον νευρώνα εξόδου, το οποιο σφάλμα αξιοποιείται για να μεταβάλλει τα βάρη των νευρώνων εξόδου. Το σφάλμα κάθε νευρώνα εξόδου μοιράζεται προς τα πίσω στους κρυφούς νευρώνες που είναι συνδεδεμένοι με αυτόν (backpropagation), συναρτήσει πάντα του βάρους της σύνδεσης κάθε κρυφού νευρώνα με τον νευρώνα εξόδου. Μόλις ολοκληρωθεί αυτή η προς τα πίσω διανομή για έναν κρυφό νευρώνα, αθροίζεται η συνολική διαφοροποίηση για τον συγκεκριμένο (κρυφό) νευρώνα και αυτό το άθροισμα χρησιμοποιείται για να φτιάξει τα νέα βελτιωμένα βάρη του. Η διαδικασία επαναλαμβάνεται προς τα πίσω για όλους τους νευρώνες μέχρι να ανανεωθούν όλα τα βάρη. Επειδή η μέθοδος της οπισθοδρόμησης απαιτεί η συνάρτηση ενεργοποίησης κάθε νευρώνα να μπορεί να παραγοντοποιηθεί, δεν μπορεί να δουλέψει με κάποια γραμμική συνάρτηση, γι’αυτό και επιλέγεται μία σύνάρτηση ενεργοποίησης του είδους "Υπερβολική Εφαπτομένη" ή "Σιγμοειδής".

Παραπάνω είδαμε πως για την ενημέρωση του κάθε βάρους χρησιμοποιείται το άθροισμα των σφαλμάτων όλων των δεδομένων εκπαίδευσης, δηλαδή έχουμε μαζική εκμάθηση (Bulk-Learning). Στην συγκεκριμένη εργασία όμως θα χρησιμοποιηθεί μία εναλλακτική μέθοδος, η διαδικτυακή μάθηση (Online Learning), κατά την οποία η ενημέρωση για το κάθε βάρος γίνεται σε κάθε ένα γεγονός ξεχωριστά. Επιλέχθηκε η παραπάνω μέθοδος καθώς, ενώ έχει την απαίτηση να χρησιμοποιείται δείγμα εκπαίδευσης με αρκετά καλή τυχαιότητα, προσφέρει πολλά πλεονεκτήματα στη συνολική απόδοση του αλγορίθμου.

Εκπαίδευση Νευρωνικού Δικτύου (NN)-Η μέθοδος BFGS

Η μέθοδος BFGS (Broyden-Fletcher-Goldfarb-Shannon) διαφέρει από την μέθοδο της οπισθοδρόμησης στη συνάρτηση σφάλματος ως προς την προσαρμογή των βαρών, καθώς εδώ χρησιμοποιείται η δεύτερη παράγωγος. Βασικό στοιχείο της μεθόδου είναι ο επαναλαμβανόμενος υπολογισμός ενός αντίστροφου Hessian πίνακα:

$$H^{-1(k)} = \frac{D \cdot D^T \cdot (1 + Y^T \cdot H^{-1(k-1)} \cdot Y)}{Y^T \cdot D} - D \cdot Y^T \cdot H + H \cdot Y \cdot D^T + H^{-1(k-1)}$$

Για τα παραπάνω, το Y είναι το διάνυσμα των σφαλμάτων της κλίσης και το D το διάνυσμα μεταβολών του βάρους, τα οποία υπολογίζονται ως εξής:

$$D_i^{(k)} = w_i^k - w_i^{(k-1)}$$
$$Y_i^{(k)} = g_i^k - g_i^{(k-1)}$$

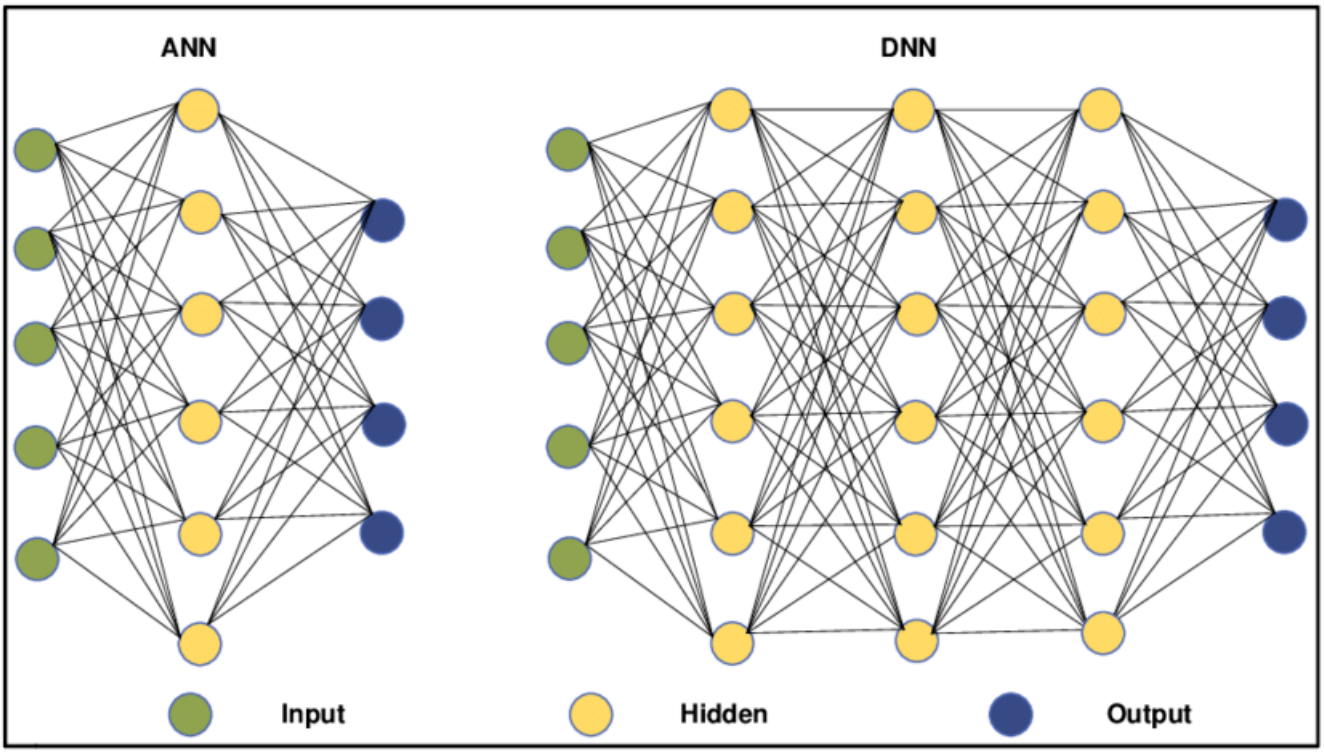
όπου $w_i,\ g_i$ το βάρος και η κλίση (gradient) αντίστοιχα της κάθε σύναψης, ενώ k ο αριθμός του κάθε βήματος. Το κάθε νέο διάνυσμα μεταβολών του βάρους υπολογίζεται από τη σχέση:

$$D^{(k)} = -H^{-1(k)} \cdot Y(k)$$

Μετά τον αρχικό υπολογισμό-προσέγγιση του ανάστροφου Hessian πίνακα, υπολογίζεται ένα διάνυσμα κατεύθυνσης, πολλαπλασιάζοντας τον αρχικό Hessian matrix με την κλίση προσέγγισης του ελάχιστου σημείου μίας παραβολής. Ύστερα εφαρμόζεται μία αναζήτηση γραμμής (line search) για να βρεθεί ένα κατάλληλο βήμα για την παραπάνω κατεύθυνση. Υπολογίζεται το Y_i και ο νέος Hessian πίνακας σύμφωνα με την πρώτη σχέση. Στη συνέχεια, από την τελευταία σχέση για το $D^{(k)}$ υπολογίζεται η νέα προσέγγιση για τον ανάστροφο Hessian πίνακα του επόμενου βήματος $H^{(-1)}$.[3, 2]

2.4.1 Deep Neural Networks

Όπως εξηγήσαμε παραπάνω, ένα τεχνητό νευρωνικό δίκτυο αποτελείται από το στρώμα εισόδου, το στρώμα εξόδου και ένα ή παραπάνω κρυφά στρώματα νευρώνων. Η βασική διαφορά των κλασικών Neural Networks με τα Deep Neural Networks (DNNs) έγκειται στο πλήθος των κρυφών στρωμάτων (hidden layers). Σε αντίθεση με τα κλασικά νευρωνικά δίκτυα, ένα DNN μπορεί να έχει δεκάδες ή εκατοντάδες από κρυφά στρώματα, τα οποία του επιτρέπουν να εκπαιδεύεται πολύ αποδοτικότερα, προσθέτοντας βέβαια και επιλέον υπολογιστικό κόστος.



Σχήμα 2.4: Σχηματική αναπαράσταση των τεχνητών (ή απλών) νευρωνικών δικτύων (ANN) και των Deep Neural Networks (πηγή εικόνας: Research Gate)

Αυτή η βασική διαφορά προσφέρει στα Deep NNs αρκετά αυξημένη απόδοση στην διαδικασία μηχανικής μάθησης. Δεν πρέπει όμως να θεωρηθεί ότι το συγκεκριμένο είδος αλγορίθμων μπορεί να αντικαταστήσει τα απλά νευρωνικά δίκτυα. Όπως κανένα είδος αλγορίθμου μηχανικής μάθησης δεν είναι κατάλληλο για όλες τις περιπτώσεις, έτσι και εδώ έχει παρατηρηθεί ότι πολλές φορές υπάρχουν ορισμένα είδη προβλημάτων όπου τα DNNs δεν αποδίδουν πάντα καλύτερα από τα απλά Neural Networks στο τελικό μοντέλο πρόβλεψης, συνήθως, λόγω εμφάνισης overtraining.

Κεφάλαιο 3

Δεδομένα και Επεξεργασία

Τα δεδομένα τα οποία επεξεργαστήκαμε προέρχονται από εξομοίωση διάσπασης ζεύγους top-antitop quarks. Η συγκεκριμένη διάσπαση επιλέχθηκε καθώς προσφέρει μεγάλο αριθμό hadron jets ώστε να γίνει η μελέτη τους. Συγκεκριμένα, η παραπάνω διάσπαση εξελίσσεται με τους παρακάτω τρόπους:

1. $t\bar{t} \to W^+bW^-\bar{b} \to q\bar{q}'bq''\bar{q}'''\bar{b}$ (45.7%)
2. $t\bar{t} \to W^+bW^-\bar{b} \to q\bar{q}'bl^-\bar{\nu}_l\bar{b} + l^+\nu_l bq''\bar{q}'''\bar{b}$ (43.8%)
3. $t\bar{t} \to W^+bW^-\bar{b} \to l^+\nu_l bl'^-\bar{\nu}_{l'}\bar{b}$ (10.5%)

Επειδή όλα τα quark που εμφανίζονται στις τελικές καταστάσεις θα εξελιχθούν σε hadron jets, είναι εμφανή τα πλεονεκτήματα επιλογής της συγκεκριμένης διάσπασης. Τα παραπάνω αναγραφόμενα ποσοστά για την κάθε διάσπαση δεν κάνουν διαχωρισμό ως προς το είδος του λεπτονίου (l) που συμβάλει κάθε φορά, το οποίο μπορεί να είναι $e,\ \mu$ ή τ λεπτόνιο. Στα αποτελέσματα της εξομοίωσης που επεξεργαστήκαμε υπάρχει διαχωρισμός ως προς το είδος του λεπτονίου που ανιχνεύεται σε κάθε γεγονός. Είναι σημαντικό επίσης να αναφέρουμε ότι εκτός από τα jets που προέρχονται από την διάσπαση $t\bar{t}$, λόγω έξτρα QCD ακτινοβολίας (quarks και γλουόνια) από σωματίδια χρώματος σε κάθε γεγονός, είναι πολύ πιθανόν να παρατηρηθούν και έξτρα jets. Τέλος, ο αριθμός των jet που θα παρατηρηθεί σε κάθε γεγονός εξαρτάται από την κινηματική της κάθε διάσπασης καθώς επίσης και από αλγόριθμο προσδιορισμού του jet που χρησιμοποιείται.[5]

3.1 Οι μεταβλητές για την κινηματική μελέτη στους ανιχνευτές

Οι συγκρουόμενες ακτίνες στον ανιχνευτή έχουν ένα φάσμα από ορμές κατά μήκος τους άξονά τους, οι οποίες καθορίζονται από την συνάρτηση κατανομής των παρτονίων. Το κέντρο μάζας της σκέδασης παρτονίου-παρτονίου ενισχύεται (boosted) ως προς το αντίστοιχο των δύο εισερχόμενων παρτονίων. Μας είναι χρήσιμο λοιπόν να κατηγοριοποιήσουμε την τελική κατάσταση με όρους μεταβλητές που μετασχηματίζονται απλά κάτω από διαμήκη boosts. Για αυτό τον σκοπό εισάγουμε τους όρους ωκύτητα (rapidity) y, εγκάρσια ορμή P_t και αζιμουθιακή γωνία $\varphi\,(-\pi \leqslant \varphi \leqslant \pi)$. Με την χρήση των παραπάνω μεταβλητών, η τετραορμή ενός σωματιδίου με μάζα m μπορεί να γραφτεί

$$\begin{aligned} p^\mu &= (E, p_x, p_y, p_z) \\ &= (m_T cosh, p_T sin\varphi, p_T cos\varphi, m_T sinhy) \end{aligned}$$

όπου η εγκάρσια μάζα ορίζεται ως

$$m_T = \sqrt{p_T^2 + m^2}$$

Η ωκύτητα y ορίζεται ως

$$y = \frac{1}{2} ln(\frac{E+p_z}{E-p_z}).$$

Πρακτικά, προτιμάται η χρήση της μεταβλητής της ψευδωκύτητας (pseudorapidity) η, η οποία ορίζεται

$$\eta = -lntan(\frac{\theta}{2})$$

αφού η γωνία θ, η γωνία δηλαδή μεταξύ της ορμής του σωματιδίου **p** και του θετικού ημιάξονα της δέσμης, μπορεί να μετρηθεί απευθείας από τον ανιχνευτή.

Οι τιμές που παίρνει το η σύμφωνα με την μαθηματική περιγραφή του είναι από $-\infty$ έως $+\infty$, στην πραγματικότητα όμως, εντός τους ανιχνευτή είναι από -5 έως $+5$.

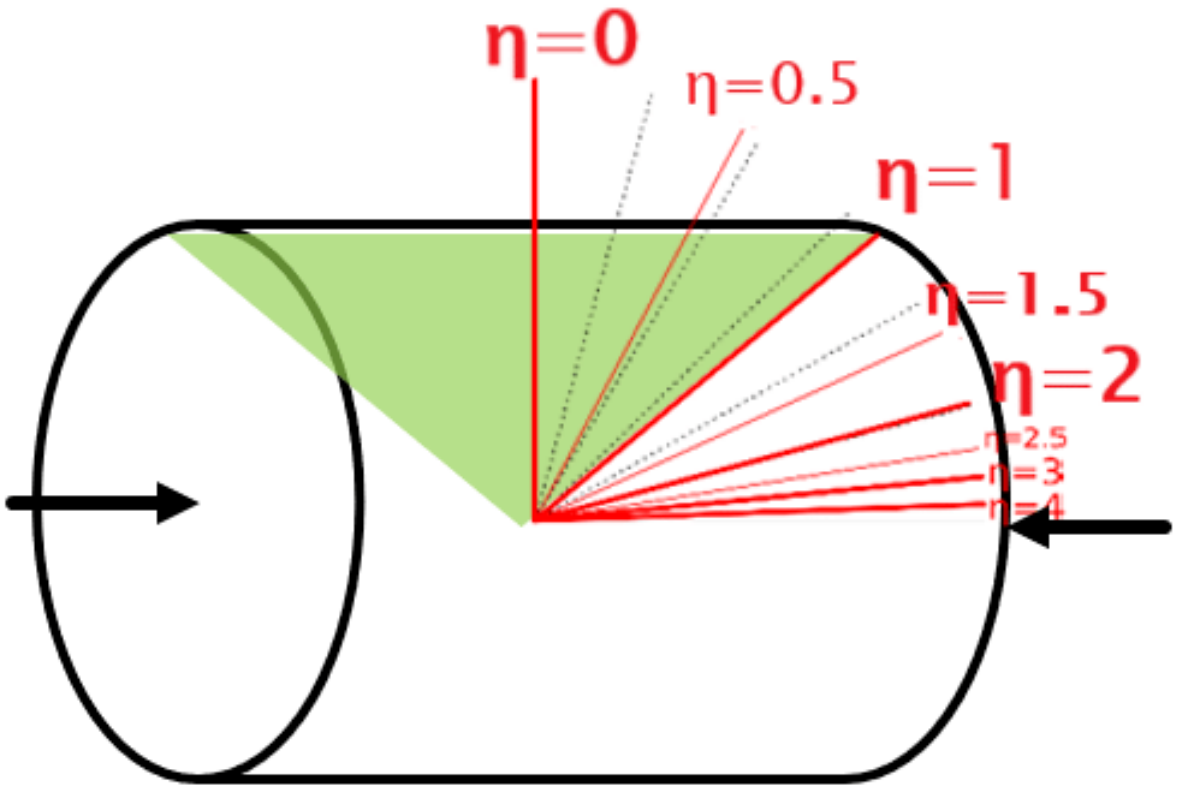
Η ψευδωκύτητα η μπορεί να προσεγγίζει την ωκύτητα y όταν η μάζα του σωματιδίου τείνει στο 0 ($m \to 0$) ή, αντίστοιχα, όταν η ταχύτητα του τείνει στην ταχύτητα του φωτός. Τότε έχουμε την προσέγγιση

$$m << p \Rightarrow E \approx p \Rightarrow \eta \approx y.$$

Είναι επίσης σύνηθες να μετράται η εγκάρσια ενέργεια

$$E_T = Esin\theta$$

αντί της εγκάρσιας ορμής p_T καθώς αυτό είναι το μέγεθος που μετράται στα αδρονικά καλορίμετρα.[14]

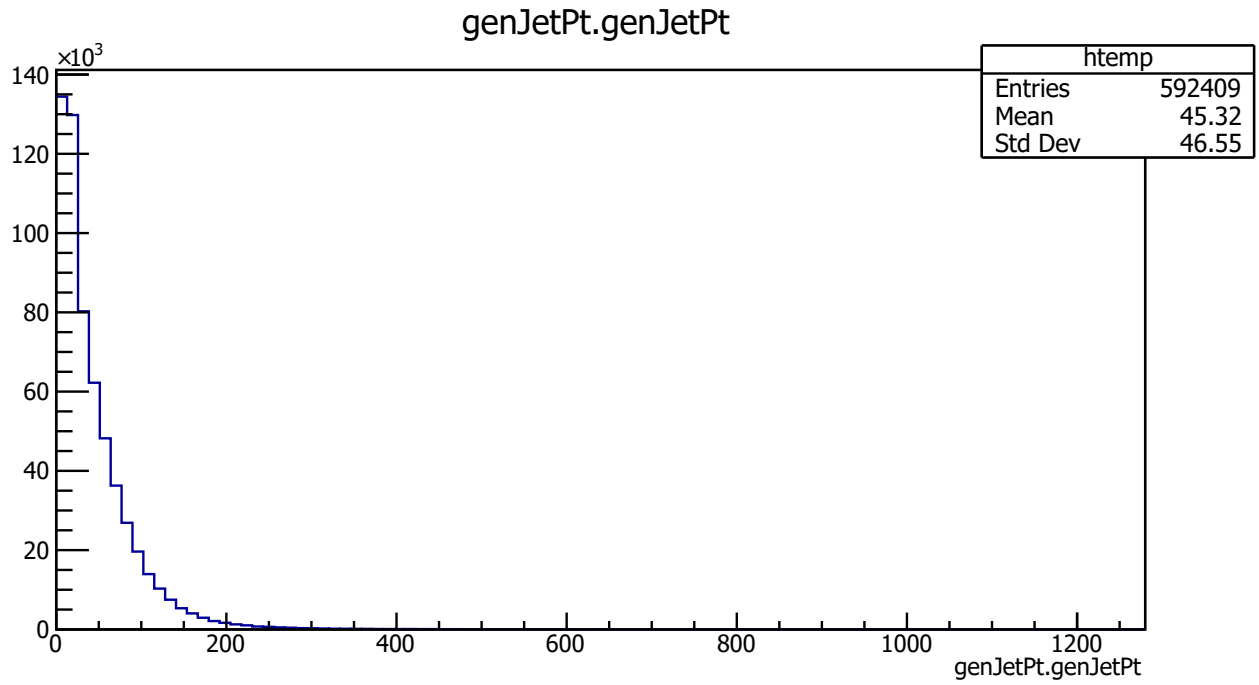


Σχήμα 3.1: Αναπαράσταση της εξέλιξης των τιμών της ψευδωκύτητας (η) σε έναν κύλινδρο (πηγή εικόνας: Research Gate)

Παρακάτω παρουσιάζονται οι κατανομές των μεταβλητών του πίδακα που θα αξιοποιήσουμε για την εκπαίδευση των αλγορίθμων μηχανικής μάθησης, καθώς επίσης, εν συντομία, κάποια φυσικά χαρακτηριστικά τους και η φυσική τους σημασία.

Η μεταβλητή genJetPt

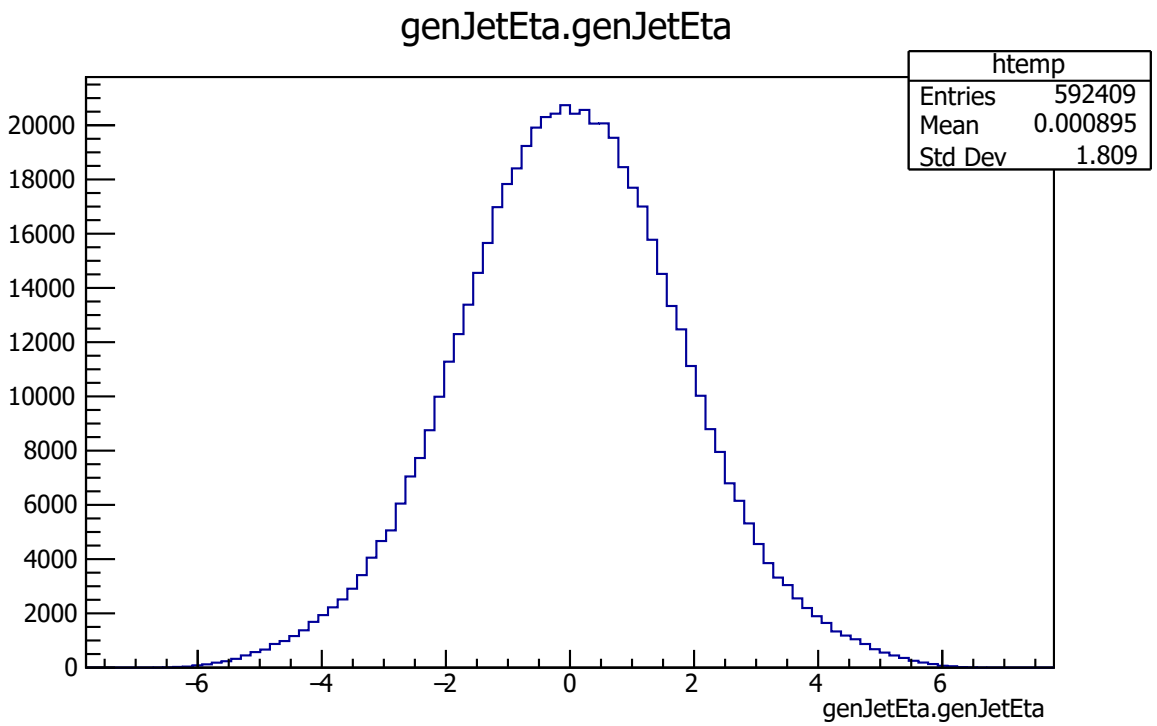
Η μεταβλητή αυτή περιγράφει την κάθετη ορμή ($p_{Transverse}$) του jet, όπως αυτό διαμορφώνεται από τα σωματίδια που το αποτελούν, πρίν συναντήσει τον ανιχνευτή και μετά το επίπεδο παρτονίων και την διαδικασία hadronization.



Σχήμα 3.2: Απεικόνιση της μεταβλητής genJetPt με το λογισμικό ROOT

Η μεταβλητή genJetEta

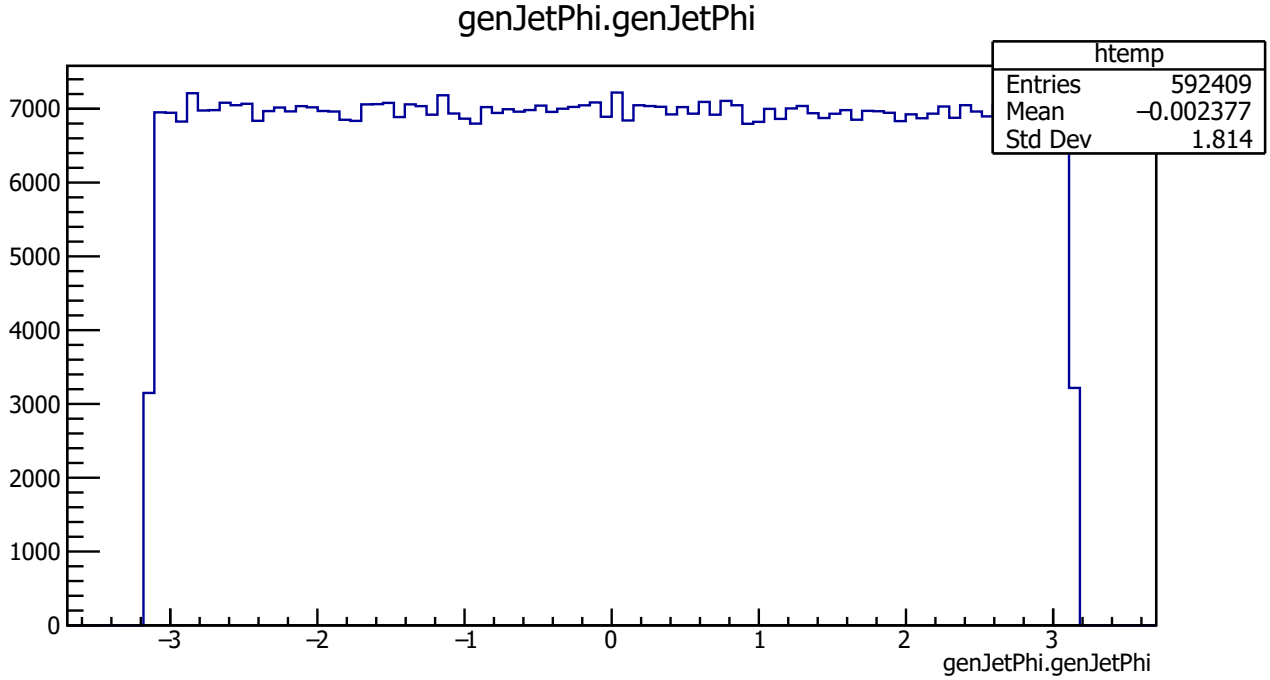
Με αυτή την μεταβλητή περιγράφεται η ψευδωκύτητα του πίδακα πριν αυτός συναντήσει τον ανιχνευτή.



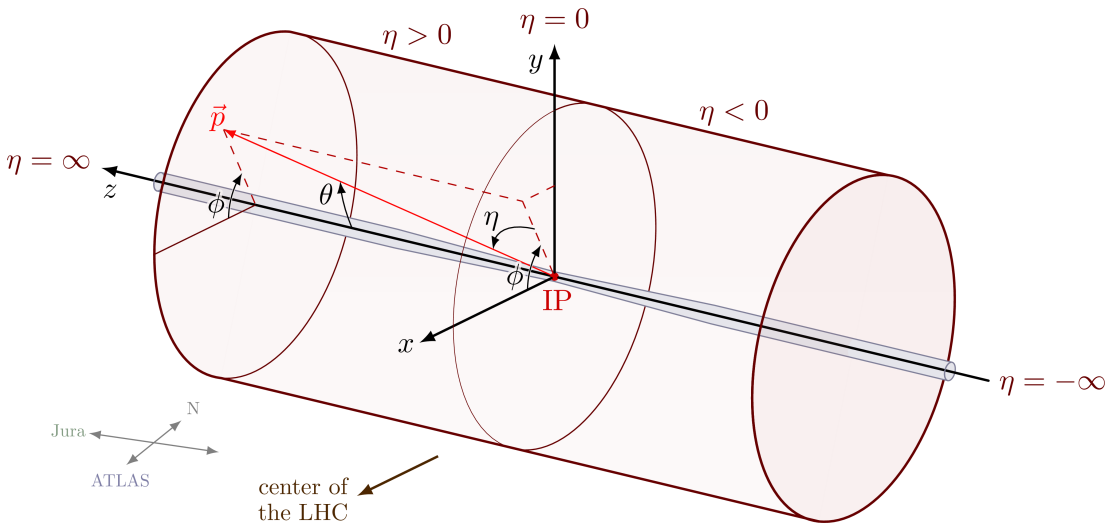
Σχήμα 3.3: Απεικόνιση της μεταβλητής genJetEta με το λογισμικό ROOT

Η μεταβλητή genJetPhi

Αντίστοιχα, με αυτή την μεταβλητή, δίνεται η αζιμούθια γωνία φ του πίδακα πριν αυτός περάσει στις ανιχνευτικές διατάξεις.



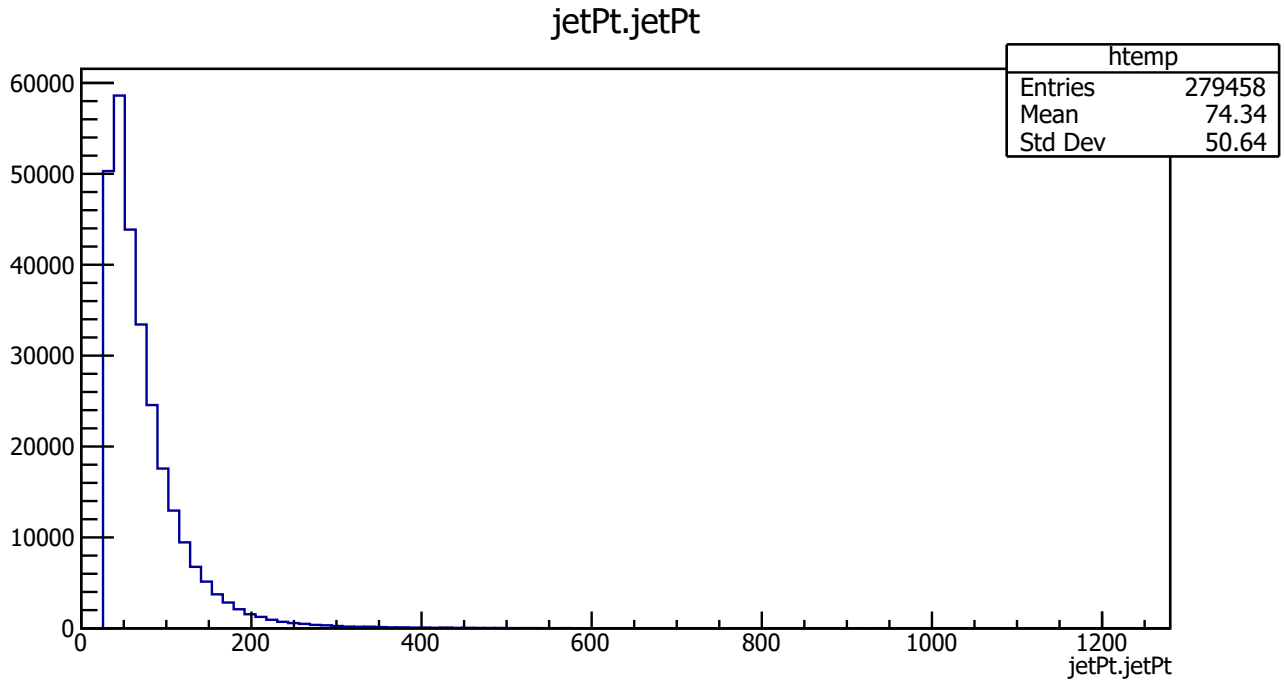
Σχήμα 3.4: Απεικόνιση της μεταβλητής genJetPhi με το λογισμικό ROOT



Σχήμα 3.5: Οι μεταβλητές της ορμής p, της αζιμούθιας φωνίας φ και της ψευδωκύτητας η στον κύλινδρο του πειράματος CMS (πηγή εικόνας: tikz.net)

Η μεταβλητή jetPt

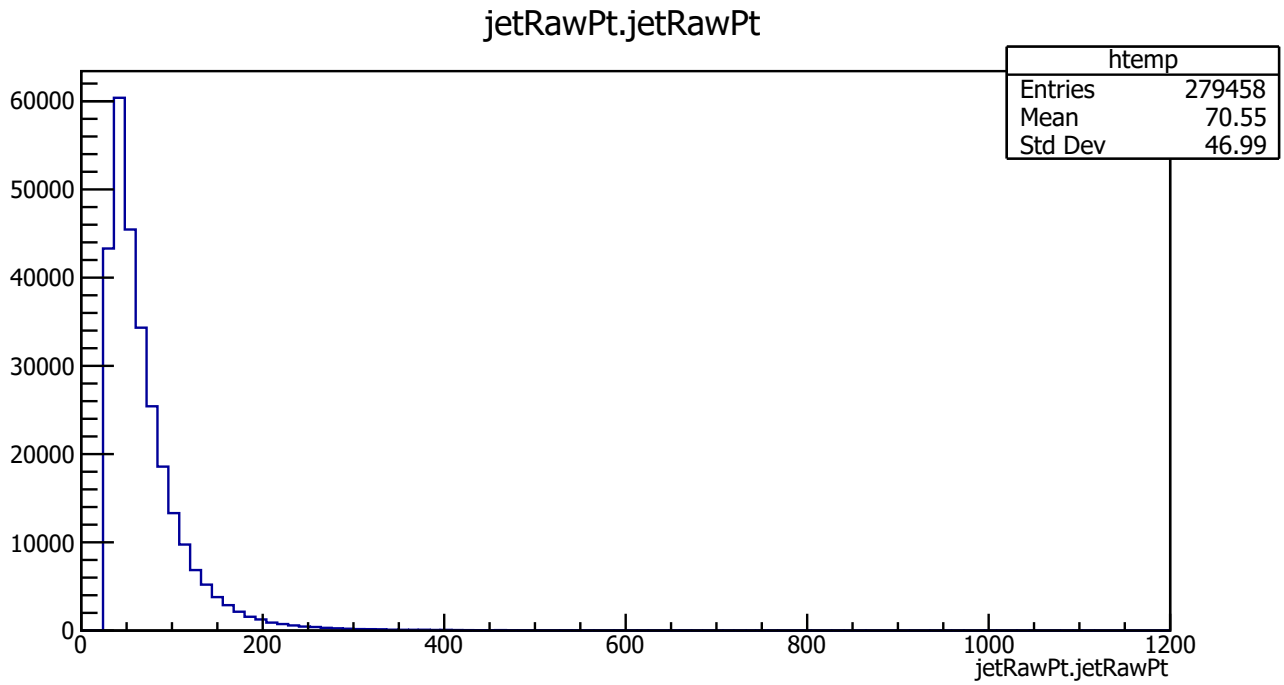
Η συγκεκριμένη μεταβλητή μας δίνει την διορθωμένη Pt του jet, σύμφωνα με την μέθοδο calibration που χρησιμοποιείται σήμερα. Η μεταβλητή αυτή αποκτά ιδιαίτερη σημασία για την τελική σύγκριση της μεθόδου που επιχειρείται στην παρούσα εργασία σε σχέση με την υπάρχουσα μέθοδο.



Σχήμα 3.6: Απεικόνιση της μεταβλητής jetPt με το λογισμικό ROOT

Η μεταβλητή jetRawPt

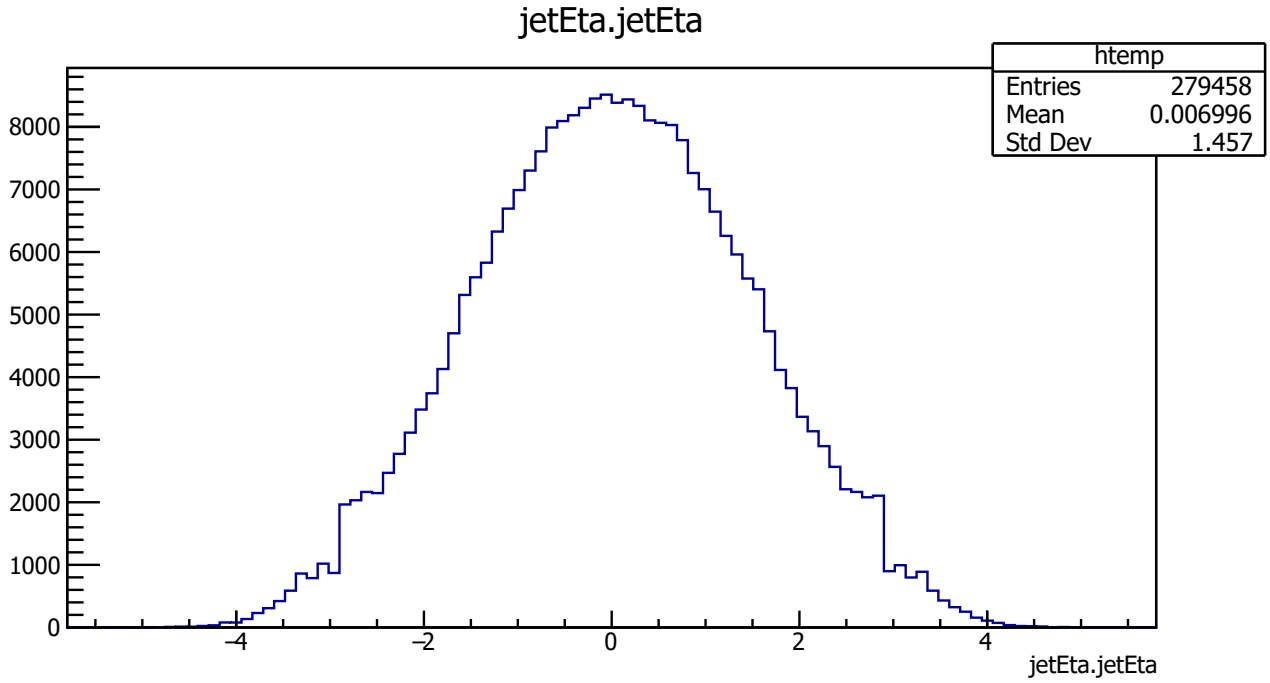
Η μεταβλητή αυτή μας δίνει την εγκάρσια ορμή του πίδακα P_t όπως αυτή μετράται από τον ανιχνευτή.



Σχήμα 3.7: Απεικόνιση της μεταβλητής jetRawPt με το λογισμικό ROOT

Η μεταβλητή jetEta

Η μεταβλητή αυτή μας δίνει την ψευδωκύτητα που μετράει ο ανιχνευτής για κάθε jet.

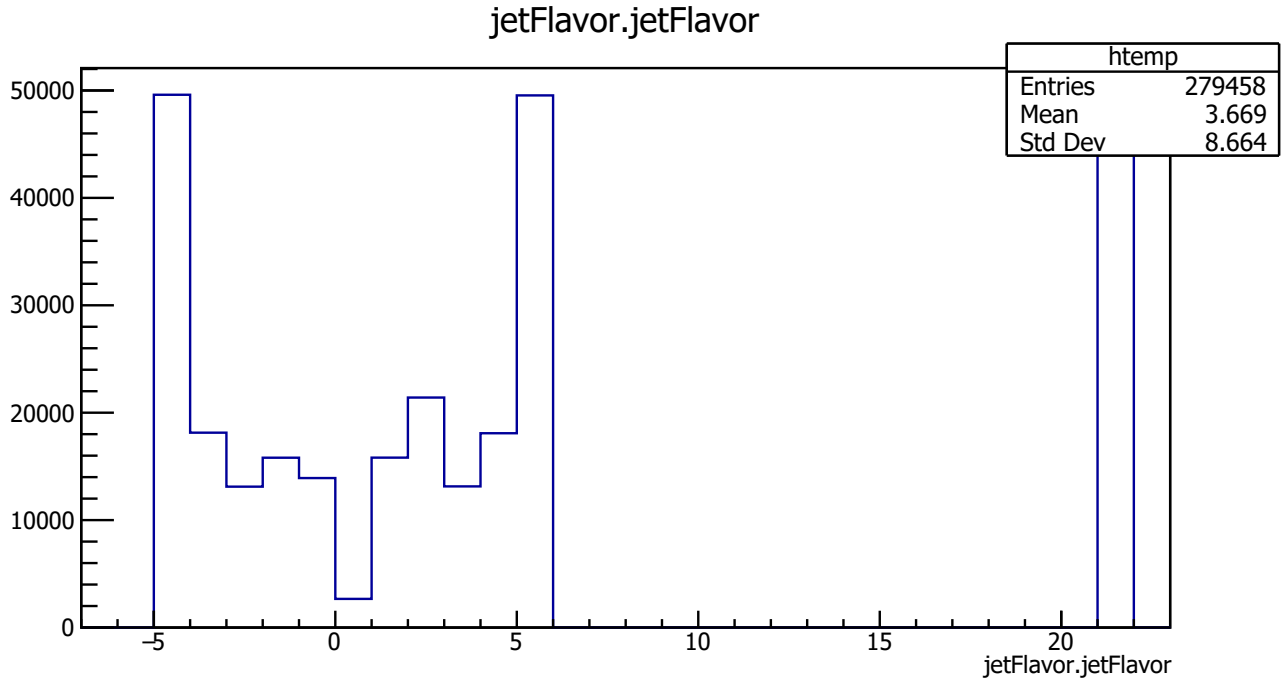


Σχήμα 3.8: Απεικόνιση της μεταβλητής jetEta με το λογισμικό ROOT

Η μεταβλητή jetFlavor

Η συγκεκριμένη μεταβλητή επιστρέφει το είδους σωματιδίου που έχει αντιστοιχίσει (tag) στον πίδακα η προσομοίωση. Από 1 έως 5 είναι τα κουάρκ, με 1 να αντιστοιχεί στο up, 2 στο

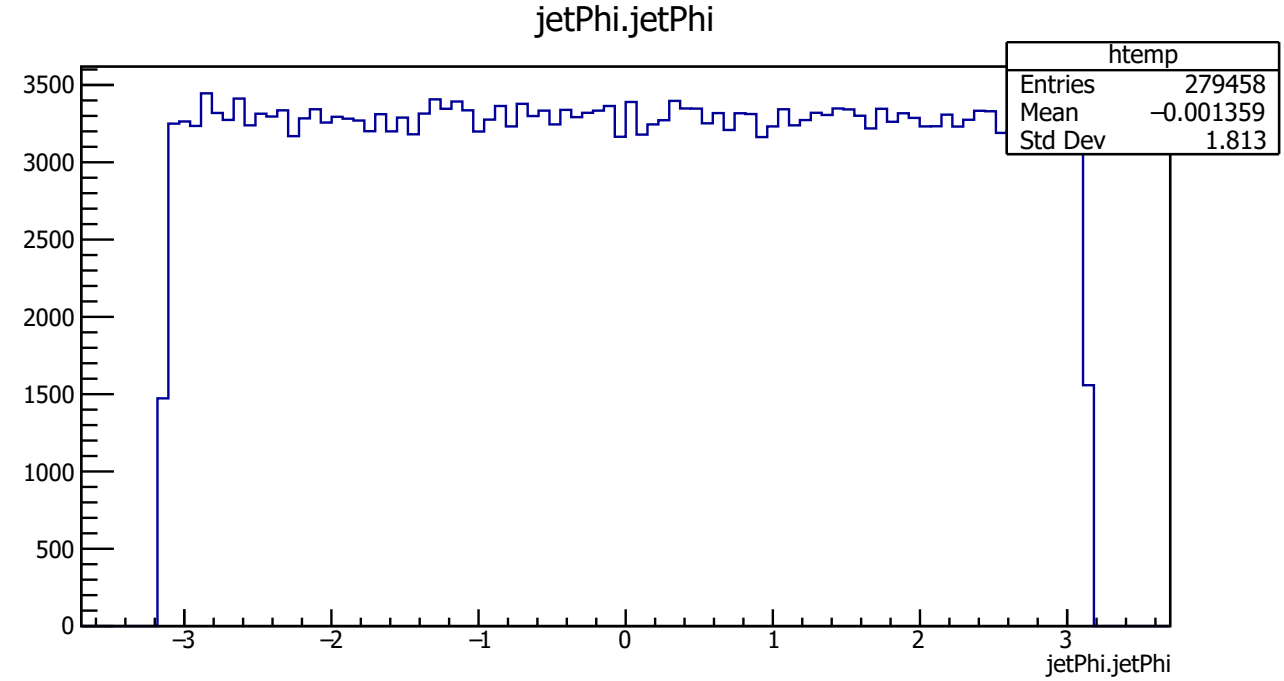
down, 3 στο strange, 4 στο charm και 5 στο bottom. Από -1 έως -5 είναι τα αντίστοιχα αντι-κουαρκ. Στο 21 είναι το πλήθος τον πιδάκων από γλουόνια.



Σχήμα 3.9: Η μεταβήτή jetFlavor με το λογισμικό ROOT. Φαίνεται η κατανομή σε είδος κάθε πί-δακα.

Η μεταβλητή jetPhi

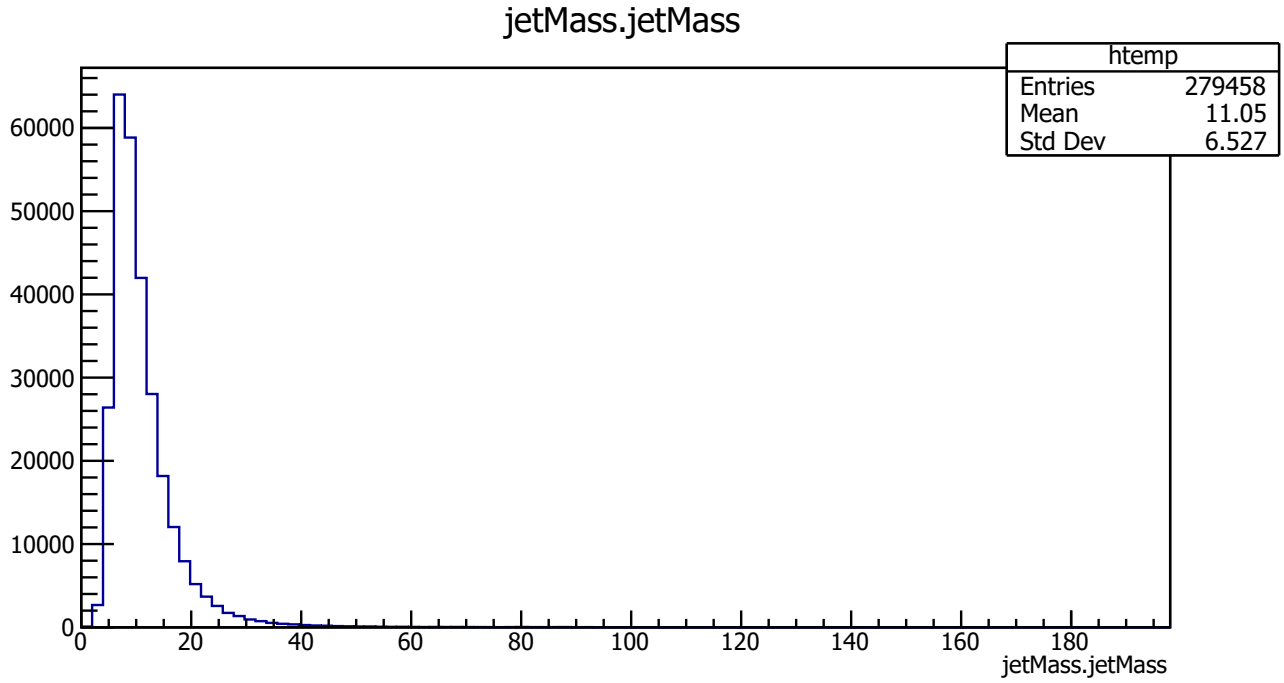
Με αυτή τη μεταβλητή μας δίνεται η αζιμούθια γωνία φ κάθε πίδακα όπως αυτή μετράται από τον ανιχνευτή.



Σχήμα 3.10: Απεικόνιση της μεταβλητής jetPhi με το λογισμικό ROOT

Η μεταβλητή jetMass

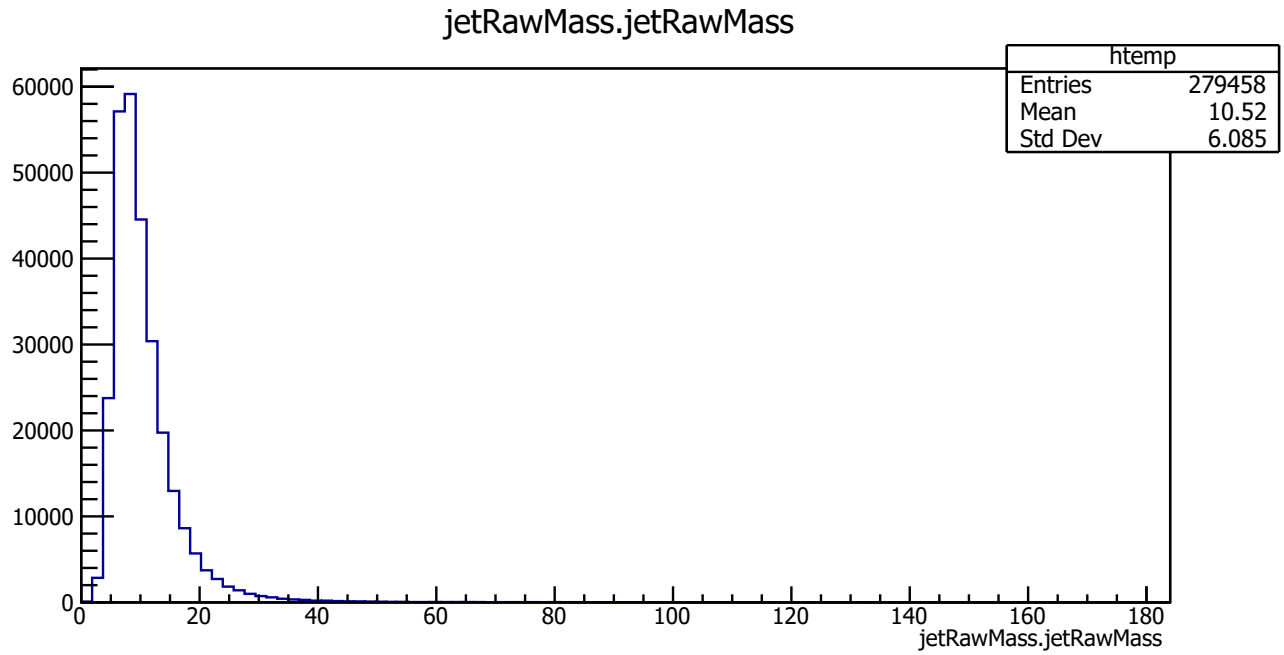
Η μεταβλητή αυτή μας δίνει την πραγματική μάζα του πίδακα, πριν δηλαδή εισέρθει στις ανιχνευτικές διατάξεις.



Σχήμα 3.11: Απεικόνιση της μεταβλητής jetMass με το λογισμικό ROOT

Η μεταβλητή jetRawMass

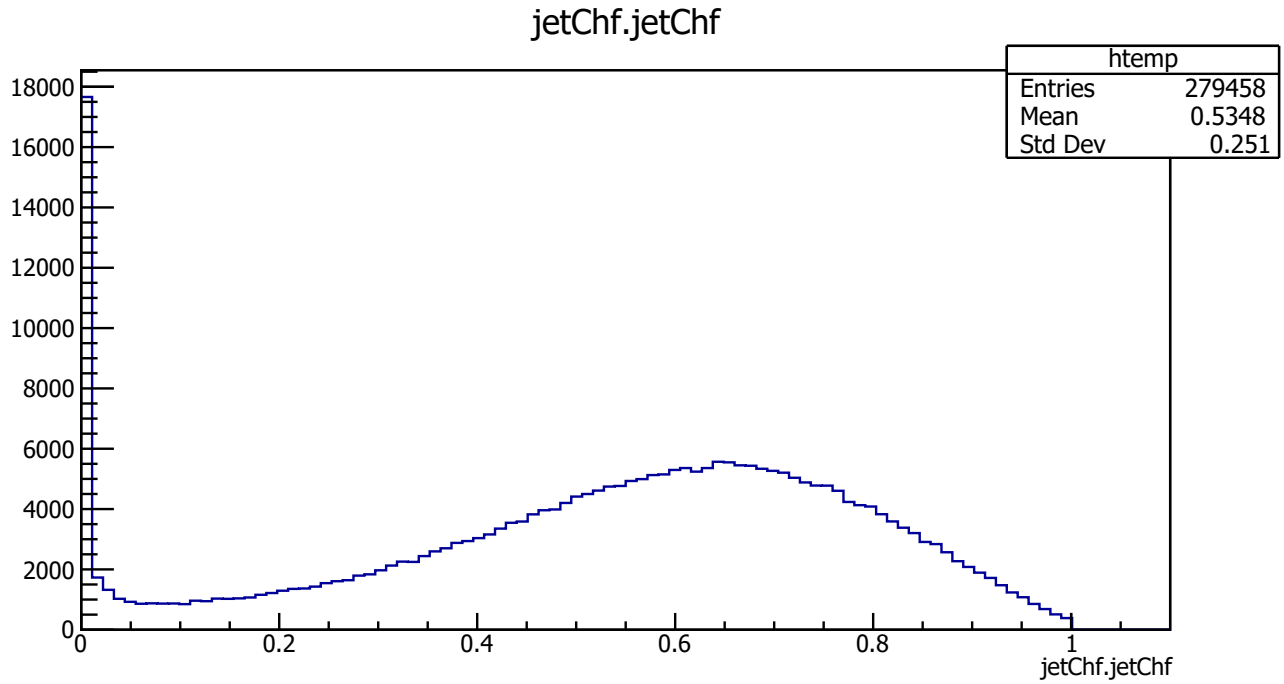
Η μεταβλητή αυτή μας δίνει την μάζα του πίδακα που μετράει ο ανιχνευτής.



Σχήμα 3.12: Απεικόνιση της μεταβλητής jetRawMass με το λογισμικό ROOT

Η μεταβλητή jetChf

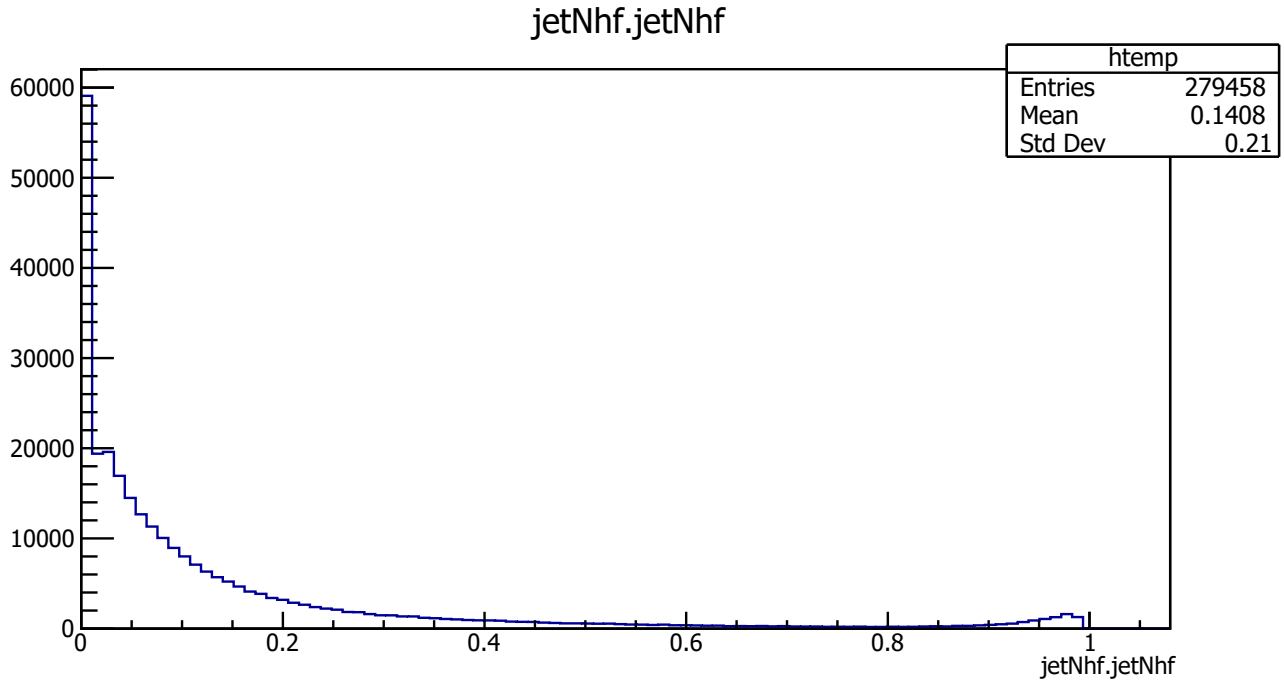
Η μεταβλητή αυτή μας δίνει το ποσοστό φορτισμένων αδρονίων στον πίδακα.



Σχήμα 3.13: Απεικόνιση της μεταβλητής jetChf με το λογισμικό ROOT

Η μεταβλητή jetNhf

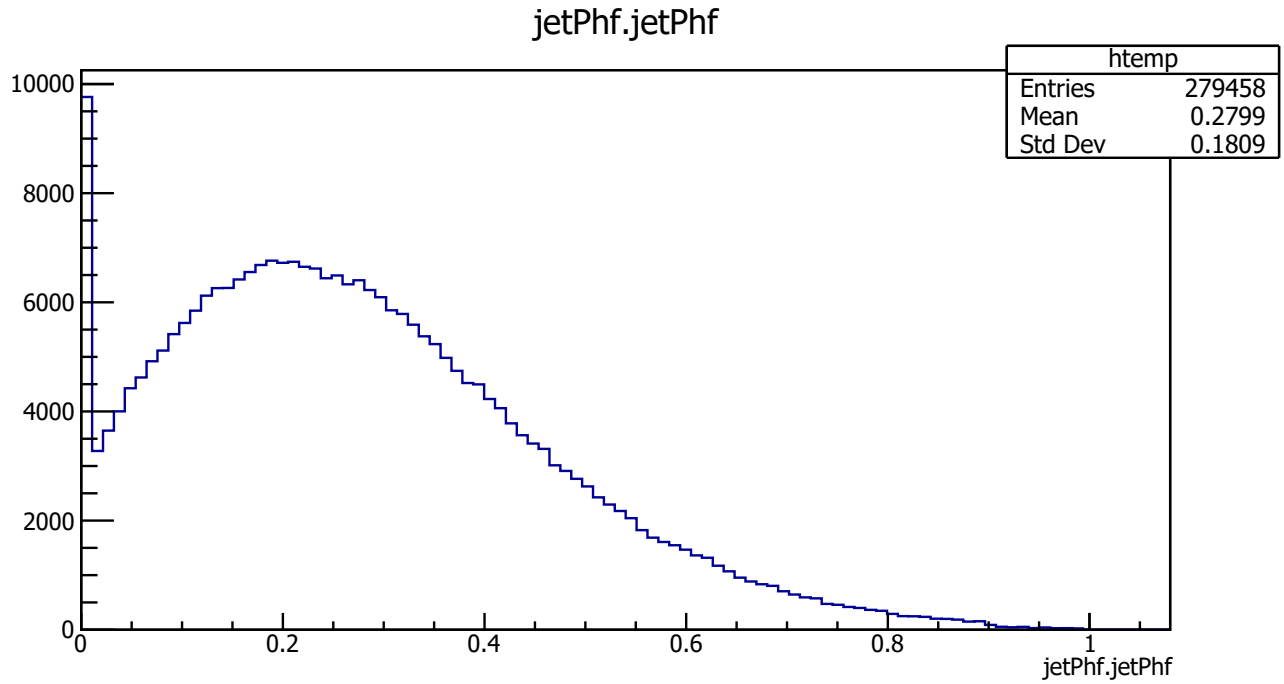
Η μεταβήτή αυτή μας δίνει το ποσοστό των ουδέτερων αδρονίων στον πίδακα.



Σχήμα 3.14: Απεικόνιση της μεταβλητής jetNhf με το λογισμικό ROOT

Η μεταβλητή jetPhf

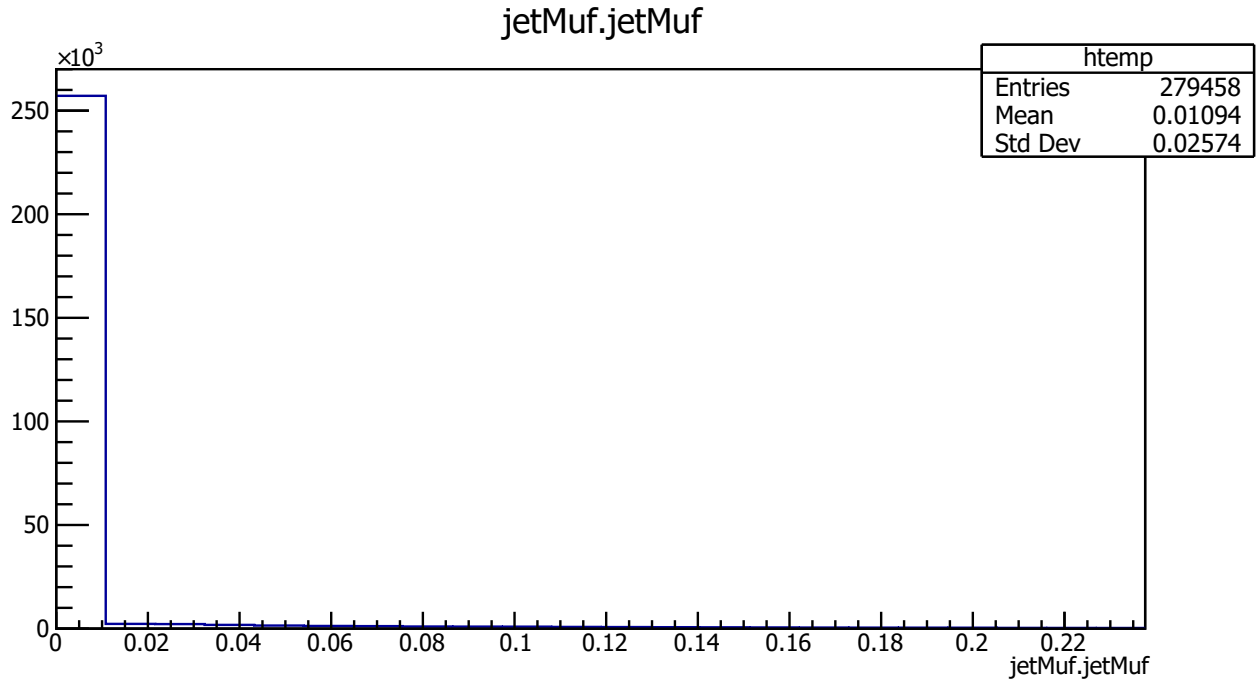
Η μεταβλητή αυτή μας δίνει το ποσοστό των φωτονίων στον πίδακα.



Σχήμα 3.15: Απεικόνιση της μεταβλητής jetPhf με το λογισμικό ROOT

Η μεταβλητή jetMuf

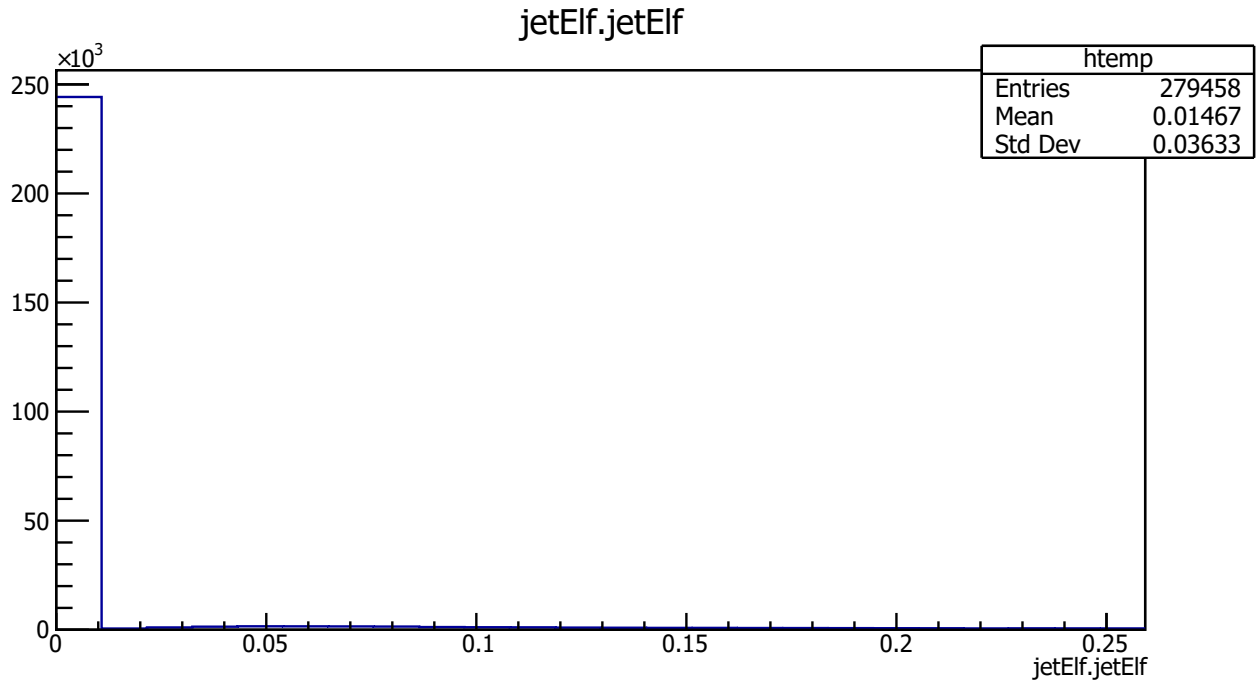
Με αυτή τη μεταβλητή παίρνουμε το ποσοστό μιονίων στον πίδακα.



Σχήμα 3.16: Απεικόνιση της μεταβλητής jetMuf με το λογισμικό ROOT

Η μεταβλητή jetElf

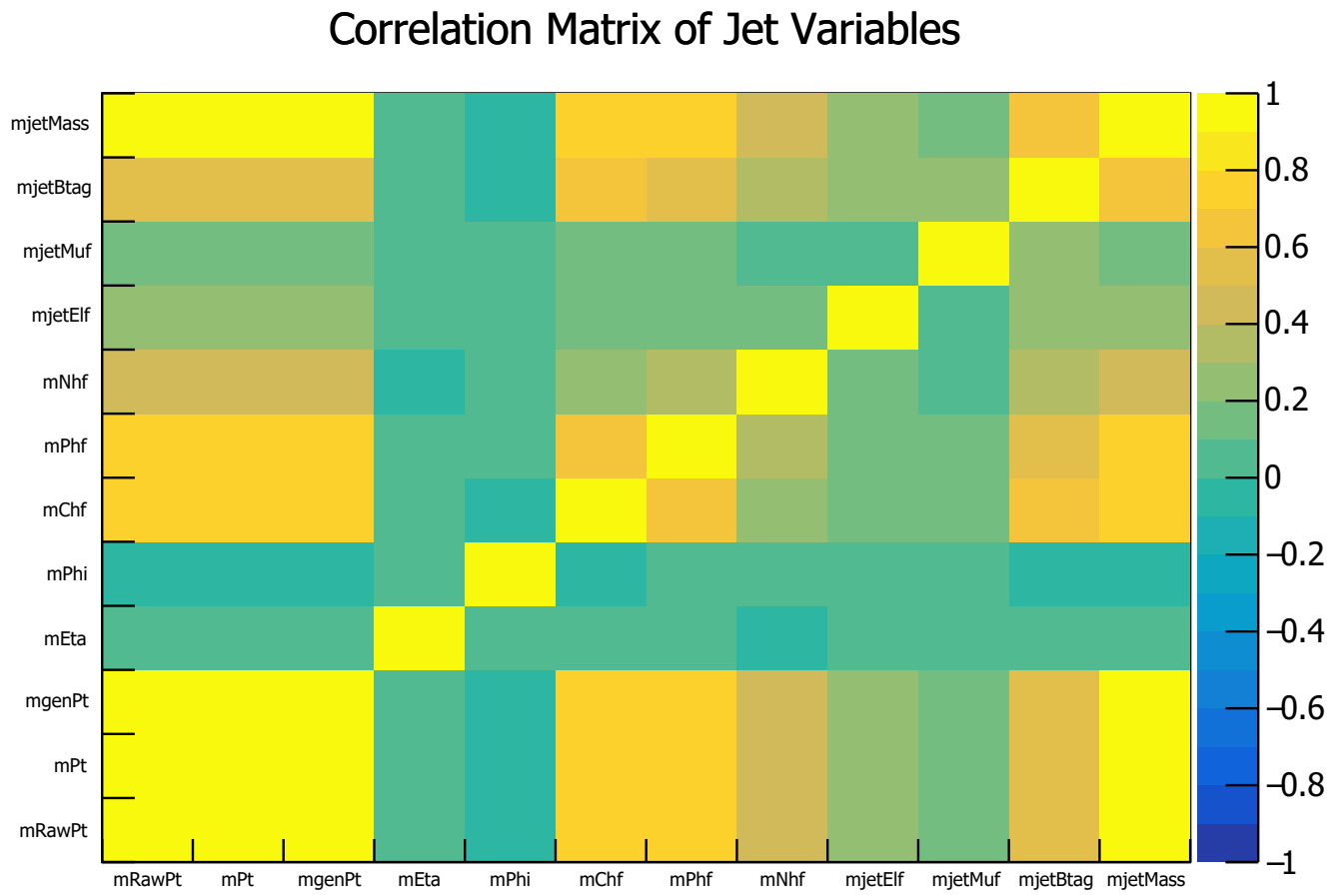
Με αυτή τη μεταβλητή παίρνουμε το ποσοστό ηλεκτρονίων στον πίδακα.



Σχήμα 3.17: Απεικόνιση της μεταβλητής jetElf με το λογισμικό ROOT

Συσχέτιση μεταξύ των μεταβλητών

Πολλές φορές είναι βοηθητικό να μελετήσουμε έναν πίνακα συσχέτισης (correlation matrix) των μεταβλητών μας. Ένας τέτοιου είδους πίνακας είναι της μορφής $2x2$ και έχει στήλες και γραμμές τις μεταβλητές του προβλήματος μας. Κάθε κελί περιέχει έναν αριθμό, ο οποίος εκφράζει πόσο εξαρτάται η μεταβλητή i από την μεταβλητή j. Συγκεκριμένα αν στο κελί έχουμε την τιμή 1 σημαίνει απόλυτη συσχέτιση μεταξύ των μεταβλητών, -1 σημαίνει απόλυτη αντισυσχέτιση μεταξύ τους ενώ 0 σημαίνει ότι δεν υπάρχει συσχέτιση. Για τις μεταβλητές του προβλήματος μας προκύπτει ο εξής πίνακας.



Σχήμα 3.18: Ο πίνακας συσχέτισης για τις μεταβλητές του προβήματος μας. Με πορτοκαλί/κίτρινο και μπλε έχουμε τα ζευγάρια των ισχυρότερα συσχετισμένων ή αντισυσχετισμένων μεταβλητών.

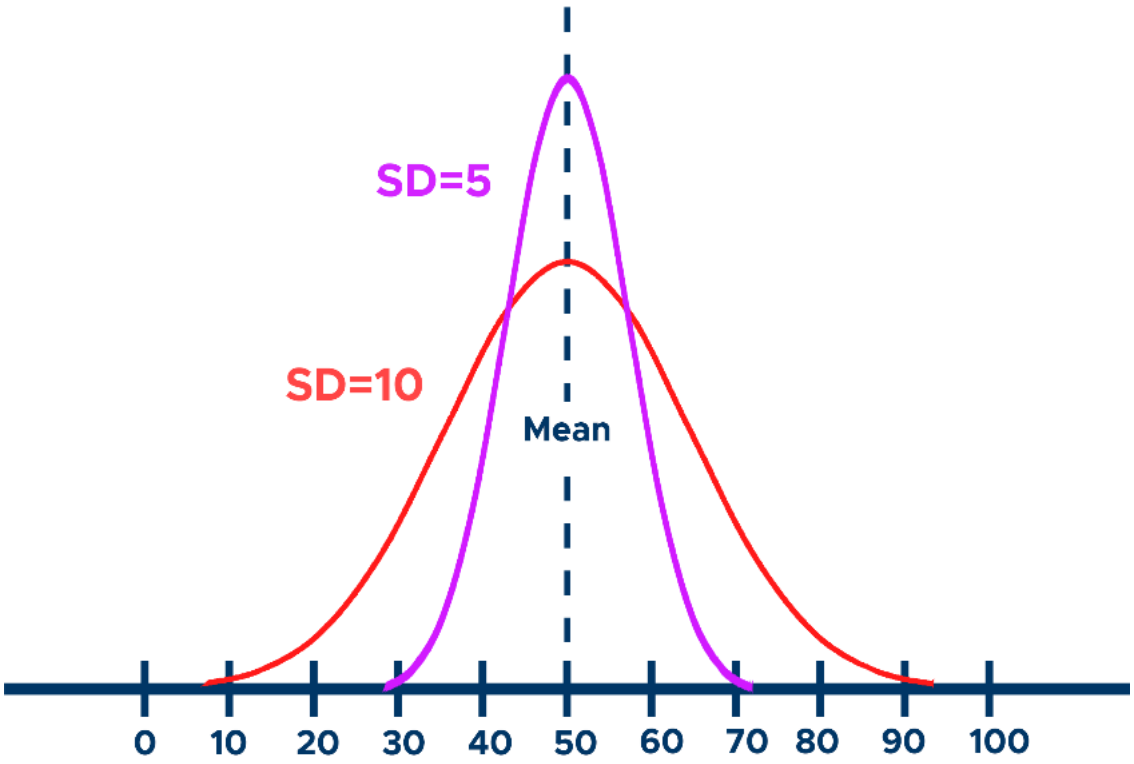
Από τον παραπάνω πίνακα μας ενδιαφέρει άμεσα η γραμμή της μεταβλητής mgenPt, καθώς αυτή θα αποτελέσει και τον στόχο πάνω στον οποίο θα εκπαιδευθούν οι αλγόριθμοι μας (καθοδηγούμενη μηχανική μάθηση). Παρατηρούμε ότι η μεταβλητή/στόχος εμφανίζει έντονη συσχέτιση με μεταβλητές όπως η mRawPt ή η mRawMass, καθώς και έντονη αντισυσχέτιση με μεταβλητές όπως η mMass και η mjetElf. Μηδενική έως καθόλου συσχέτιση εμφανίζεται μεταξύ της μεταβλητής mgenPt και των mRawPhi, mjetMuf κ.α.

Παράμετροι στατιστικής ανάλυσης δεδομένων

Η μελέτη της απόκρισης (Response) του ανιχνευτή βασίζεται στον λόγο της μετρούμενης ορμής με την πραγματική ορμή του πίδακα. Αν μετρούσαμε με απόλυτη ακρίβεια την πραγματική εγκάρσια ορμή κάθε πίδακα, η απόκρισης του ανιχνευτή θα είχε τη μορφή μίας συνάρτησης δέλτα στη μονάδα. Αφού δεν έχουμε τέλεια ανίχνευση της ορμής, οι αποκρίσεις θα ακολουθούν γκαουσιανές κατανομές. Σε αυτές τις κατανομές θα εφαρμόσουμε γκαουσιανό fit. Η απόδοση κάθε παλινδρομητή (BDTs, NNs, DNNs) κρίνεται από τις εξής παραμέτρους από τις κατανομές που προκύπτουν για το Response κάθε μεθόδου μετά την εκπαίδευση της:

- Μέση Τιμή (mean): Αποτελεί την πιο κοινή τιμή σε μια συλλογή δεδομένων. Είναι ένα μέτρο της κεντρικής τάσης μιας κατανομής πιθανότητας και αναφέρεται επίσης και ως αναμενόμενη τιμή (expected value).

- Διασπορά (standard deviation): Εκφράζει πόσο συγκεντρωμένες είναι οι τιμές γύρω από τη μέση τιμή. Μεγάλη διασπορά σημαίνει ότι υπάρχουν πολλά δεδομένα δεξιά και αριστερά της μέσης τιμής, ενώ, αντίθετα, μικρή διασπορά σημαίνει ότι θα συναντήσουμε λίγα στοιχεία δεξιά ή αριστερά της μέσης τιμής, με τα περισσότερα να είναι συγκεντρωμένα σε αυτή.



Σχήμα 3.19: Παράδειγμα της μέσης τιμής (mean) σε μία κατανομή καθώς και του μεγέθους της διασποράς (standard deviation).

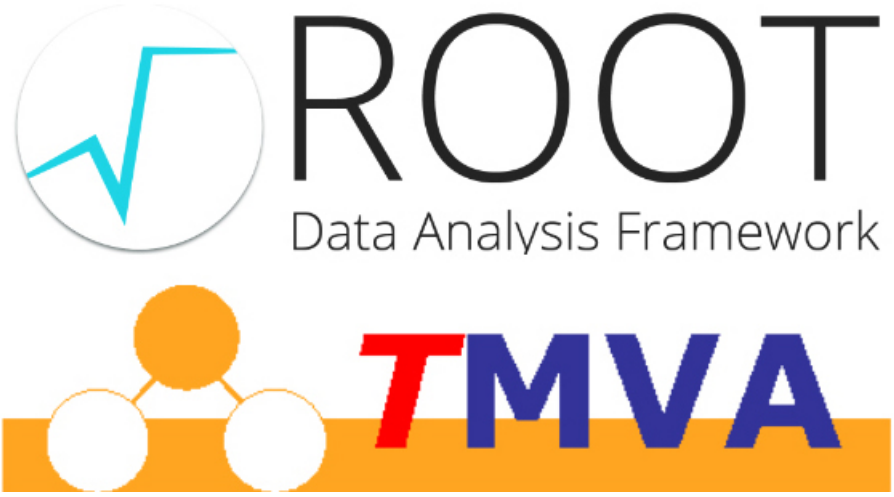
Τα εργαλεία που χρησιμοποιήθηκαν

Το βασικότερο εργαλείο που χρησιμοποιήσαμε για την ανάλυση και την επεξεργασία των δεδομένων ήταν το λογισμικό **ROOT**. Το ROOT, γραμμένο σε γλώσσα λογισμικού C++, αναπτύχθηκε στο CERN τη δεκαετία του 1990 και από τότε χρησιμοποιείται ευρέως για την ανάλυση δεδομένων για φυσική υψηλών ενεργειών από επιστήμονες σε όλο τον κόσμο, καθώς και σε πολλές άλλες εφαρμογές όπως αστρονομία κλπ. Το ROOT μας δίνει επίσης μεγάλες δυνατότητες για σχεδιασμό γραφικών παραστάσεων (ιστογράμματα, fitting plots κλπ.) και λεπτομερέστερη παρουσίαση των αποτελεσμάτων μας. Λόγω του τρόπου συμπίεσης και αποθήκευσης των δεδομένων το ROOT μπορεί να αποθηκεύσει μεγάλο όγκο πληροφοριών σε μικρό υπολογιστικό χώρο, ενώ η δομή των δεδομένων σε TTrees δίνει την δυνατότητα για εύκολη πρόσβαση σε μεγάλο όγκο δεδομένων, με ταχύτητα και ευκολία από πολλούς διαφορετικούς χρήστες σε διαφορετικά λειτουργικά συστήματα. Ένα TTree είναι σε θέση να αποθηκεύσει όλα τα είδη δεδομένων, όπως αντικείμενα, πίνακες ή ιστογράμματα. Όταν χρησιμοποιούμε ένα TTree, γεμίζουμε τις διακλαδώσεις (Branches) και τα φύλλα (Leafs) με δεδομένα τα οποία εγγράφονται στο δίσκο όταν το δέντρο είναι γεμάτο. Είναι σημαντικό να συνειδητοποιήσουμε ότι κάθε αντικείμενο δεν γράφεται μεμονωμένα, αντιθέτως τα αντικείμενα συλλέγονται και γράφονται ομαδικά. Σε αυτό το

σημείο το TTree εκμεταλλεύεται τη συμπίεση και παράγει ένα πολύ μικρότερο αρχείο από ό,τι αν τα αντικείμενα ήταν γραμμένα μεμονωμένα.

Για την εφαρμογή των μεθόδων της μηχανικής μάθησης στην επεξεργασία των δεδομένων μας χρησιμοποιήσαμε το **TMVA** (Toolkit for MultiVariate Analysis), το οποίο είναι ενσωματωμένο στο ROOT. Το TMVA προσφέρει την δυνατότητα ταυτόχρονης χρήσης μίας πληθώρας αλγορίθμων (BDTs, Neural Networks, Fischer κα) για training, testing, performance evaluation και εφαρμογής των μεθόδων ταξινόμησης και regression.Το TMVA έχει σχεδιαστεί ειδικά για τις ανάγκες των εφαρμογών της φυσικής υψηλών ενεργειών (High Energy Physics), αλλά δεν περιορίζεται σε αυτές.

Η εισαγωγή των δεδομένων και ο προσδιορισμός των παραμέτρων στα λογισμικά ROOT και TMVA, τα ιστογράμματα των αποτελεσμάτων και τα λοιπά σχεδιαγράμματα, καθώς και οι υπόλοιπες τροποποιήσεις έγιναν σε προγράμματα (scripts) γραμμένα στη γλώσσα C++. [8, 9]



3.2 Επεξεργασία

Από τα γεγονότα στους πίδακες

Όλη η ανάλυσή μας ξεκινάει από τα δεδομένα της Monte Carlo προσομοίωσης τα οποία είναι αποθηκευμένα στο αρχείο: "jetsCalibs_TT_TuneCUETP8M2T4_13TeV-powheg-pythia8.root". Μέσα σε αυτό το αρχείο είναι αποθηκευμένα σε μορφή TTree τα δεδομένα που θέλουμε να αναλύσουμε. Πριν προχωρήσουμε στον κορμό της επεξεργασίας οποιουδήποτε δεδομένου αυτής της εργασίας, χρειάστηκε να γίνει μία βασική επεξεργασία στο παραπάνω αρχείο ROOT. Όπως γίνεται αντιληπτό ήδη από την εισαγωγή, θα χρειαστεί να μελετήσουμε τους αδρονικούς πίδακες με τα αντίστοιχα χαρακτηριστικά τους. Τα Branches όμως στο TTree του αρχικού αρχείου έχουν δεδομένα από events που καταγράφει ο ανιχνευτής και όπως εξηγήθηκε στο Κεφάλαιο 1, σε ένα event μπορεί να εμφανιστεί από κανένα μέχρι αρκετά παραπάνω από ένα jets. Πρώτο βήμα λοιπόν ήταν η κατασκευή ενός νέου αρχείου ROOT που θα "ανοίξει" το προηγούμενο και θα αποτελείται πλέον από το σύνολο των Jet της προσομείωσης Monte Carlo (με τα αντίστοιχα χαρακτηριστικά τους), εύκολα και άμεσα προσβάσιμων για ανάλυση. Αυτό το αρχείο, πάνω στο οποίο θα εκτελεστούν και οι περισσότερες από τις παρακάτω διαδικασίες ονομάστηκε "matchJetTree.root". Η αρχική ομαδοποίηση των εναποτιθέμενων στον ανιχνευτή ενεργειών σε πίδακες κατά την εξομοίωση, έγινε με τη χρήση του αλγορίθμου anti-kt, όπως αντίστοιχα και στις πραγματικές μετρήσεις του CMS. Ύστερα γίνεται μια ομαδοποίηση των reconstructed jets από τον ανιχνευτή στον χώρο $\eta - \varphi$ για την αντιστοίχιση τους με τον πραγματικό πίδακα, το particle jet δηλαδή που αποτελούταν από σταθερά αδρόνια πριν "χτυπήσει" στον ανιχνευτή. Το κριτήριο για αυτή την ομαδοποίηση ήταν ότι αν κάποιο reconstructed jet είχε:

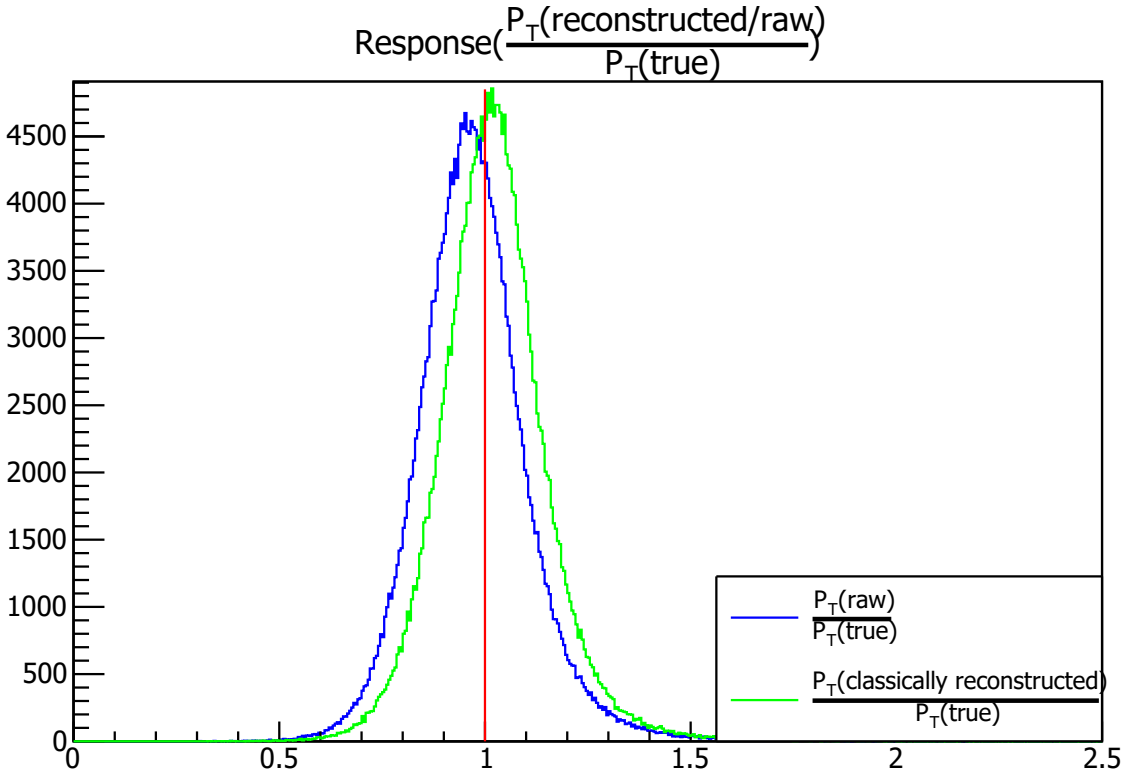
$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\varphi)^2} < 0.4$$

τότε το αντιστοιχούσαμε σε ένα "true jet"/particle jet και καταγράφαμε στο αντίστοιχο TBranch του TTree που θα αξιοποιήσουμε για την παρούσα εργασία όλα τα στοιχεία του που θα χρειαστούμε στην συνέχεια της μελέτης μας, όπως πραγματική εγκάρσια ορμή (genJetPt), κλασικά διορθωμένη εγκάρσια ορμή (jetPt), αζιμούθια γωνία (jetPhi) κλπ.

Όπως αναφέραμε και στην εισαγωγή, στόχος της παρούσας εργασίας είναι η βελτίωση της απόκρισης του ανιχνευτή κατά την μέτρηση της ορμής του πίδακα. Η απόκρισης αυτή μπορεί να μετρηθεί πολύ απλά με την εξής σχέση:

$$Response = \frac{p_T(reconstructed)}{p_T(true)}$$

δηλαδή η απόκριση είναι ο λόγος της (εγκάρσιας) ορμής που μετράει ο ανιχνευτής, και αφού αυτή διορθωθεί, με την πραγματική (εγκάρσια) ορμή του πίδακα. Σχεδιάζοντας τα αποτελέσματα της παραπάνω σχέσης για όλα τα γεγονότα παίρνουμε μια απόκριση με τη μορφή γκαουσιανών κατανομών. Γίνεται αντιληπτό πως αν ο ανιχνευτής μετρούσε ακριβώς την πραγματική ορμή κάθε jet, τότε η απόκριση θα είχε μέση τιμή ίση με 1 και μηδενική διασπορά γύρω από αυτή, θα βλέπαμε δηλαδή την εικόνα μιας συνάρτησης δέλτα. Με βάση τα παραπάνω συγκρίνουμε κι εμείς τα αποτελέσματα κάθε μεθόδου βαθμονόμησης του ανιχνευτή και θέλουμε μέση τιμή απόκρισης όσο το δυνατόν πιο κοντά στη μονάδα και με τη μικρότερη διασπορά.



Σχήμα 3.20: Οι κατανομές για την απόκριση του ανιχνευτή στα raw δεδομένα από τις μετρήσεις (μπλέ) και οι αντίστοιχες κατανομές μετά την κλασική μέθοδο βαθμονόμησης (πράσινο). Στην πρώτη περίπτωση έχουμε μέση τιμή 0.973 (απόκλιση $\backsimeq 2.7\%$ από τη μονάδα) και διασπορά 0.137 ενώ στη δεύτερη μέση τιμή 1.023 (απόκλιση $\backsimeq 2.3\%$ από τη μονάδα) και διασπορά 0.133.

Σε κάθε μέθοδο που χρησιμοποιούμε παρακάτω, τα δεδομένα μας, καθώς προέρχονται από προσομοίωση, χωρίζονται σε δύο ίσα μέρη, το ένα για να χρησιμοποιηθεί για την εκπαίδευση (training sample) και το άλλο για τον μετέπειτα έλεγχο (test sample). Ο παραπάνω διαχωρισμός έγινε και για να ελεγχθεί το φαινόμενο της υπερ-εκπαίδευσης (overtrain).

3.2.1 Η τρέχουσα μέθοδος διόρθωσης της απόκρισης στο CMS

Όπως αναφέρθηκε παραπάνω, σκοπός της βαθμονόμησης των αδρονικών πιδάκων είναι ο συσχετισμός κατά μέσο όρο της ενέργειας που μετράται για τον πίδακα από τον ανιχνευτή με την πραγματική του. Η διόρθωση γίνεται μέσω ενός πολλαπλασιαστικού παράγονα C, ο οποίος πολλαπλασιάζεται με κάθε στοιχείο της raw ανιχνευμένης τετρα-ορμής $p_\mu^{raw} = (p_T, \eta, \varphi, E)$.

$$p_\mu^{cor} = C \cdot p_\mu^{raw} = (C \cdot p_T, \eta, \varphi, C \cdot E)$$

Ο πολλάπλασιασμός του παράγονα C μόνο με την ορμή και την ενέργεια του πίδακα μας εξασφαλίζει ότι δεν θα υπάρξει αλλαγή στη διεύθυνση ανίχνευσης του πίδακα. Ο υπολογισμός του παράγοντα C είναι πολύπλοκος και εξαρτάται από πολλούς παράγοντες, όπως η θέση ανίχνευσης του πίδακα στον ανιχνευτή, βαθμονομήσεις μέσω Monte Carlo, την σχετική και η απόλυτη ενέργεια του πίδακα κα. Συγκεκριμένα, η βαθμονόμηση μέσω Monte Carlo βασίζεται σε εξομοιώσεις που διορθώνουν την ενέργεια του ανακατασκευασμένου πίδακα ώστε να είναι ίση, κατά μέσο όρο, με την ενέργεια που παράγεται από τα jet σωματιδίων, υπολογισμένη πάλι με MC. Για να γίνει η παραπάνω εξομοίωση για την εκτίμηση της ενέργειας των παραγόμενων jet σωματιδίων χρησιμοποιούνται εξομοιώσεις γεγονότων QCD οι οποίες παράγονται με το λογισμικό PYTHIA. Ο παράγοντας C που θα καταλήξουμε υπολογίζεται συναρτήσει της εγκάρσιας ορμής p_T που μετράται στον ανιχνευτή και της ψευδωκύτητας η, είναι δηλαδή $C(p_T^{raw}, \eta)$.

Η βασική βαθμονόμηση Monte Carlo (MC) προκύπτει από την θεωρία της Quantum Chromo-Dynamics (QCD) και αντιστοιχεί δείγματα σε πίδακες με γεύση (flavour) αποτελούμενη από πληθώρα πιδάκων γλουονίων χαμηλής εγκάρσιας ορμής (p_T). Ο γενικός αλγόριθμος είναι ότι κάθε ανακατασκευασμένο jet αντιστοιχείται χωρικά στον χώρο $\eta - \varphi$ με έναν πίδακα σωματιδίων από MC, με την προϋπόθεση η τιμή

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\varphi)^2}$$

να είναι μικρότερη του 0.25. Σε κάθε bin εκτίμησης MC καταγράφονται η κάθετη ορμή $p_T^{generated}$, η απόκριση $R = \frac{p_T^{reco}}{p_T^{gen}}$ και η ορμή για τον ανιχνευτή p_T^{reco}. Η μέση διόρθωση για κάθε bin ορίζεται ως το αντίστροφο της μέσης απόκρισης:

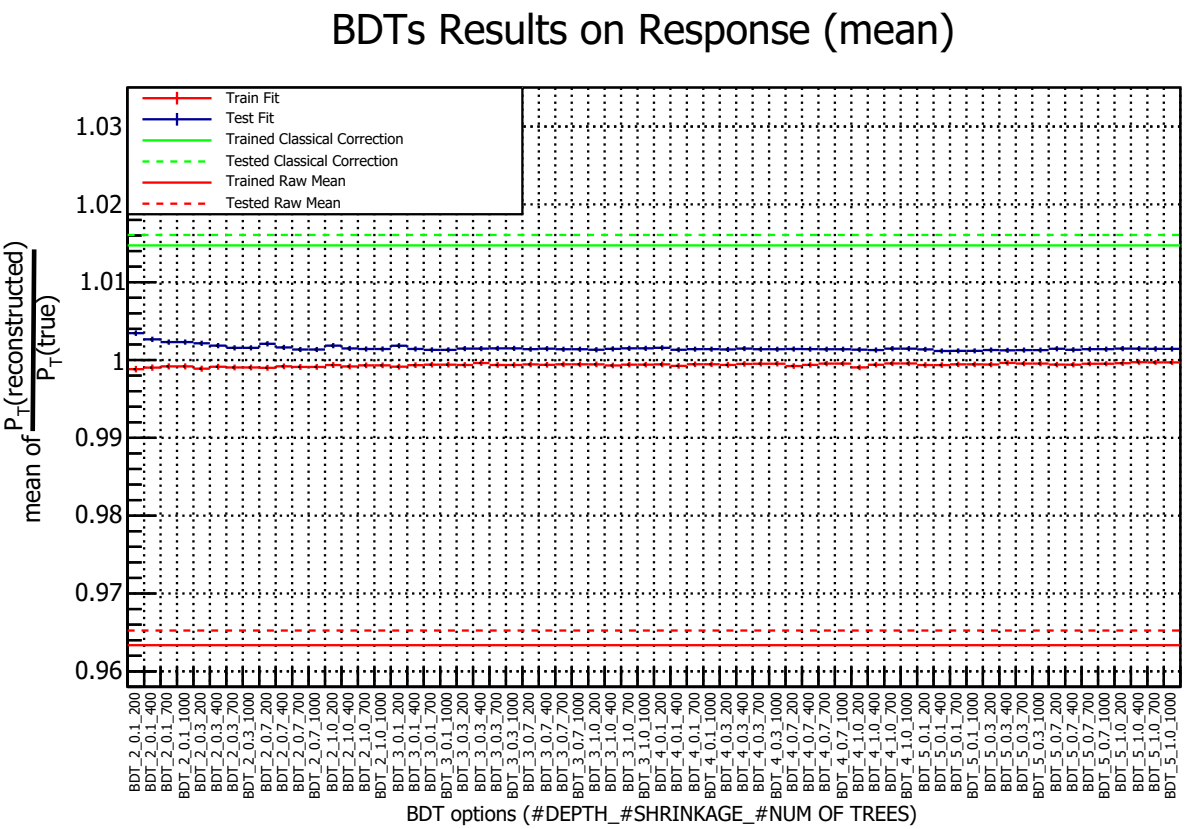
$$C_{MC}(p_T^{reco}) = \frac{1}{<R>}$$

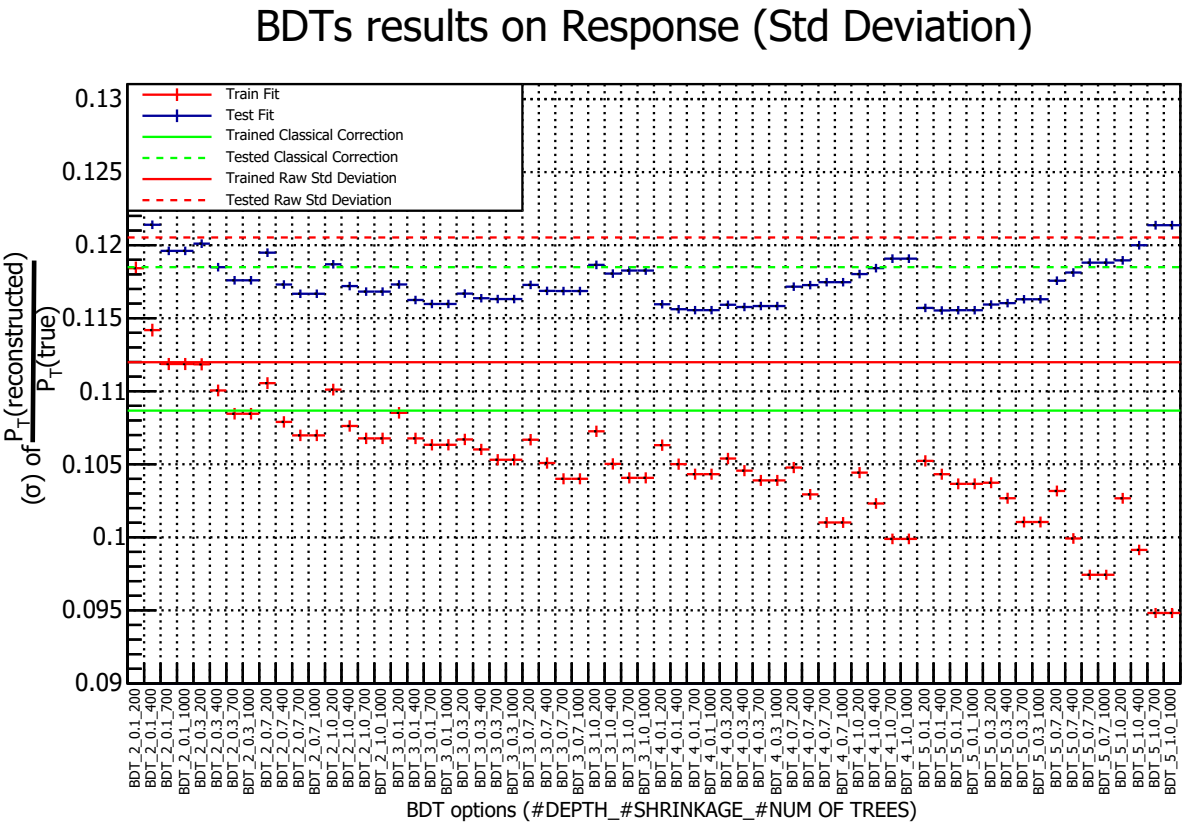
[10]

3.3 Boosted Decision Trees

Για την υλοποίηση των δέντρων απόφασης χρησιμοποιήσαμε την μέθοδο AdaBoost (όπως αυτή περιγράφεται στο 2.3) και αρχικά εκπαιδεύσαμε τα δέντρα για ορισμένες τιμές του μέγιστου βάθους κάθε δέντρου (MaxDepth:2 έως 5) και για διαφορετικά πλήθη δέντρων στο "δασος" κάθε run (NTrees=200/400/700/1000). Χρησιμοποιήσαμε επίσης το option "Shrinkage" με τιμές $0.1,\ 0.3,\ 0.7,\ 1.0$, το οποίο ορίζει τον ρυθμό εκμάθησης του αλγορίθμου GradBoost. Συνολικά η παραπάνω προσέγγιση μας δίνει $4 \times 4 \times 4 = 64$ διαφορετικά αποτελέσματα, τα οποία φαίνονται παρακάτω. Στον TMVA μέσω της κλάσης factory ορίζουμε πώς θε εκτελεστεί κάθε BDT με την αντίστοιχη από τις παραπάνω παραμέτρους. Ορίζουμε σαν στόχο εκπαίδευσης της παλινδρόμησης (Target) την μεταβλητή **mgenJetPt** . Έπειτα ορίζουμε την αναλογία δείγματος Εκπαίδευσης (Training) και Ελέγχου (Testing) που και στην περίπτωσή των BDT's είναι 50% Training – 50% Testing. Η ίδια διαδικασία ακολουθείται και για τις ρυθμίσεις των υπόλοιπων αλγορίθμων μάθησης.

```
dataloader -> AddVariable ("mRawPt", "Raw Pt",    'F');
dataloader -> AddSpectator ("mPt", "corrected Pt", 'F');
dataloader -> AddVariable ("mRawEta", "Raw Eta",    'F');
dataloader -> AddVariable ("mRawPhi", "Raw Phi", 'F');
dataloader -> AddVariable ("mChf", "Charged f", 'F');
dataloader -> AddVariable ("mPhf", "Photon f", 'F');
dataloader -> AddVariable ("mNhf", "Neutral f", 'F');
dataloader -> AddVariable ("mjetElf", "Electron f", 'F');
dataloader -> AddVariable ("mjetMuf", "Muon f", 'F');
dataloader -> AddVariable ("mjetMass", "jet Mass", 'F');
dataloader -> AddTarget ("mgenPt");
//
vector<double>grad_range{0.1,0.3,0.7,1.0};
vector<int>depth_range{2,3,4,5};
vector<int>nof_trees{200,400,700,1000};
//LOOP ON VECTORS
factory -> BookMethod (dataloader,TMVA::Types::kBDT, TString::
   Format("BDT_%d_%.1f_%d",item1,item2,item3),TString::Format("
   MaxDepth=%d:BoostType=Grad:Shrinkage=%.1f:Ntrees=%d",item1,
   item2,item3));
```





Σχήμα 3.21: Τα αποτελέσματα για τη μέση τιμή (mean) και τη διασπορά (sigma) μετά την εκπαίδευση των 64 διαφορετικών BDTs. Με κόκκινη συνεχόμενη και διακεκομμένη γραμμή είναι τα αποτελέσματα για τα δεδομένα όπως αυτά μετρούνται από τον ανιχνευτή πριν την επεξεργασία τους για το fit στα σύνολα εκπαίδευσης και ελέγχου αντίστοιχα. Την ίδια λογική έχουν και η πράσινη συνεχόμενη και διακεκομμένη γραμμή για τις τιμές της ορμής όπως αυτή προκύπτει μετά την επεξεργασία των raw δεδομένων με την κλασική μέθοδο που χρησιμοποιείται σήμερα στο CMS.

Τα παραπάνω κόκκινα σημεία αντιπροσωπεύουν τα αποτελέσματα από την εκπαίδευση των BDTs ενώ τα μπλέ τα αποτελέσματα από τον έλεγχο που έγινε, μαζί με τα σφάλματα στον υπολογισμό κάθε περίπτωσης. Όπως γίνεται αμέσως αντιληπτό, οι περισσότερες ρυθμίσεις στα BDTs καταφέρνουν να μας δώσουν αποτελέσματα καλύτερα από την κλασική μέθοδο για την μέση τιμή (πιο κοντά στη μονάδα), δεν ισχύει όμως το ίδιο και για τη διασπορά. Παρατηρούμε ότι όσο πληθαίνει το σύνολο των δέντρων απόφασης της μεθόδου (=>1000), μειώνεται το shrinkage (=>1.0) και αυξάνεται το βάθος (depth) του κάθε δέντρου (=>5), δηλαδή για τις όλο και πιο δεξιά τιμές, παρατηρείται έντονα το φαινόμενο του overtraining. Αυτό φαίνεται καθώς, ενώ ο αλγόριθμος εκπαιδεύεται πολύ καλά για το training set (κόκκινα σημεία), έχει εκπαιδευτεί πολύ συγκεκριμένα, αδυνατόντας να προσαρμόσει κατάλληλα "άγνωστα" σε αυτόν δεδομένα, όπως είναι σε αυτή την περίπτωση το testing set. Εξαιτίας αυτού βλέπουμε ότι τα κόκκινα σημεία εμφανίζουν βελτίωση, δηλαδή χαμηλότερη διασπορά, ενώ τα μπλε σημεία γίνονται συνεχώς χειρότερα, εμφανίζουν δηλαδή μεγαλύτερη διασπορά. Ο γενικότερος κανόνας (rule of thumb) που ακολουθείται σε τέτοιες περιπτώσεις ανάλυσης και θα ακολουθείται και στην συγκεκριμένη εργασία είναι ότι δεν θέλουμε τα αποτελέσματα της εκπαίδευσης με τα αποτελέσματα του ελέγχου να εμφανίζουν μεγαλύτερη διαφορά (απόσταση στο διάγραμμα) από αυτή που εμφανίζουν τα αποτελέσματα για τη μέση τιμή και τη διασπορά στο fit που έγινε στο testing set και training set των raw data (διαφορά/απόσταση κόκκινης διακεκομμένης και συνεχόμενης γραμμής) και των δεδομένων της κλασικής μεθόδου υπολογισμού της ορμής (διαφορά/απόσταση πράσινης διακεκομμένης και συνεχόμενης γραμμής). Σύμφωνα με τα παραπάνω, επιλέγουμε ως αποδοτικότερα δέντρα απόφασης αυτά με βάθος 2, shrinkage 0.7 και αριθμό δέντρων 400 και 700.

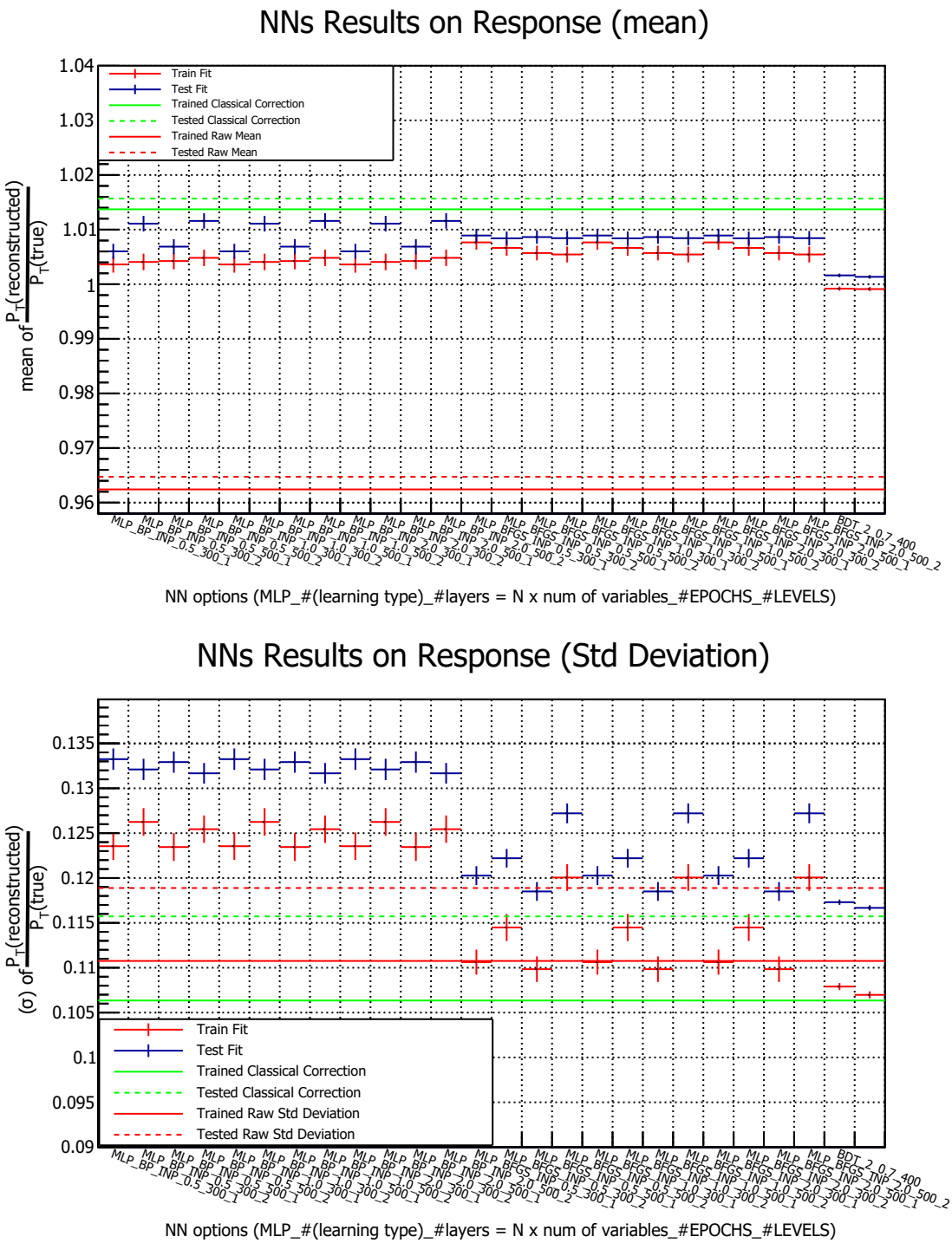
3.4 Νευρωνικά Δίκτυα (Neural Networks)

Η υλοποίηση των Νευρωνικών Δικτύων έγινε με την μέθοδο MLP (Multi Layer Perceptron). Για την μέθοδο εκπαίδευσης (TrainingMethod) δοκιμάστηκε η μέθοδος Back Propagation καθώς και η μέθοδος BFGS. Τα αντίστοιχα Νευρωνικά Δίκτυα εκπαιδεύτηκαν για $N \times 0.5$, $N \times 1$ και $N \times 2$ πλήθος κρυμμένων στρωμάτων (hidden layers), για 300 και 500 εποχές κάθε φορά και για βάθος 1 και 2 επίπεδα αντίστοιχα.

```
vector<double>inputs{0.5,1.0,2.0};
vector<int>cycles{300,500};
vector<int>nol{1,2};
.
.
.
// LOOP ON VECTORS
factory->BookMethod( dataloader, TMVA::Types::kMLP, TString::
   Format("MLP_BP_INP_%.1f_%d_%d",item1,item2,item3), TString::
   Format("!H:!V:VarTransform=Norm:NeuronType=sigmoid:NCycles=%d:
   HiddenLayers=N*%.1f:TestRate=6:TrainingMethod=BP:Sampling=1:
   SamplingEpoch=10:ConvergenceImprove=1e-6:ConvergenceTests=15:
   VarTransform=Norm:VarTransform=Norm",item2,item1) );

factory->BookMethod( dataloader, TMVA::Types::kMLP, TString::
   Format("MLP_BFGS_INP_%.1f_%d_%d",item1,item2,item3), TString::
   Format("!H:!V:VarTransform=Norm:NeuronType=sigmoid:NCycles=%d:
   HiddenLayers=N*%.1f:TestRate=6:TrainingMethod=BFGS:Sampling=1:
   SamplingEpoch=10:ConvergenceImprove=1e-6:ConvergenceTests=15:
   VarTransform=Norm",item2,item1) );
```

```
factory  ->  TrainAllMethods();
factory  ->  TestAllMethods();
factory  ->  EvaluateAllMethods();
```



Σχήμα 3.22: Τα αποτελέσματα για μέση τιμή (mean) και διασπορά (sigma) για τα παραπάνω option εκπαίδευσης των Νευρωνικών Δικτύων. Παρουσιάζονται μαζί η μέθοδος BackPropagation και η BFGS. Οι κόκκινες και πράσινες συνεχόμενες και διακεκομμένες γραμμές είναι ίδιες με την περίπτωση των BDTs. Στα 2 τελευταία (δεξιά) κελιά έχουν τοποθετηθεί τα βέλτιστα δέντρα απόφασης, όπως αυτά αποφασίστηκαν στην προηγούμενη παράγραφο, για σύγκριση.

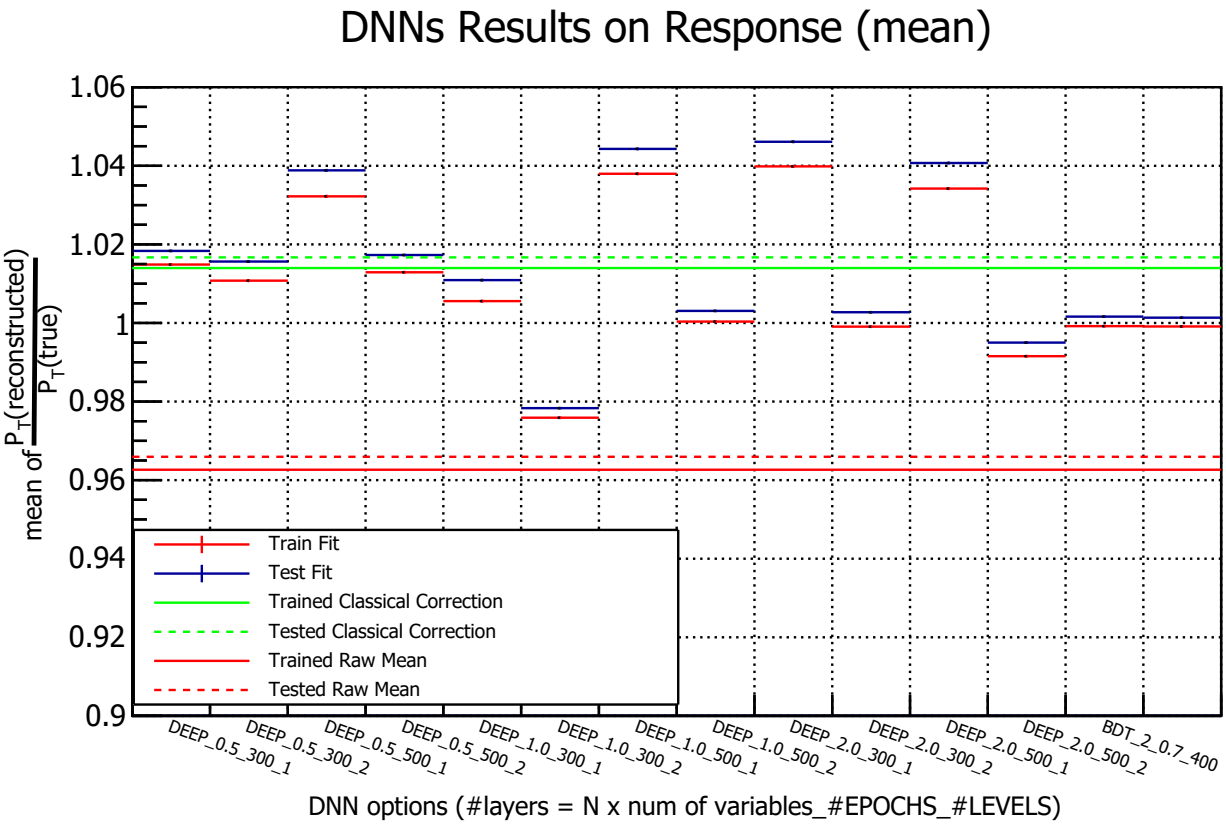
Η περίπτωση των νευρωνικών δικτύων θέλει μεγαλύτερη προσοχή για την καλύτερη περίπτωση ρύθμισεων του αλγορίθμου εκπαίδευσης. Για την μέση τιμή έχουμε και στις δύο περιπτώσεις αρκετά ικανοποιητικά αποτελέσματα, όμως στην περίπτωση του Back Propagation αλγορίθμου, όταν πάμε σε 2 επίπεδα νευρώνων, εμφανίζεται έντονο overtraining (μεγάλη διαφορά/απόσταση των σημείων). Επίσης, ενώ η μέθοδος BFGS υστερεί στην βελτίωση της μέση τιμής της κατανομής, μειώνει σημαντικά την διασπορά. Αυτό βέβαια δεν αρκεί, καθώς παρατηρείται έντονο overtraining. Σε κάθε περίπτωση, τα αποτελέσματα δεν είναι τόσο καλά όσο αυτά των ενισχυμένων δέντρων απόφασης (BDTs). Λαμβάνοντας υπόψιν τα παραπάνω σχόλια, αντιλαμβανόμαστε ότι οι πιο αποδοτικές ρυθμίσεις χωρίς την εμφάνιση overtraining είναι τα Δίκτυα της μεθόδου Back Propagation, με $N \times 1$ και $N \times 2$ πλήθος hidden layers, για 300 και 500 εποχές και βάθος 1 επίπεδο.

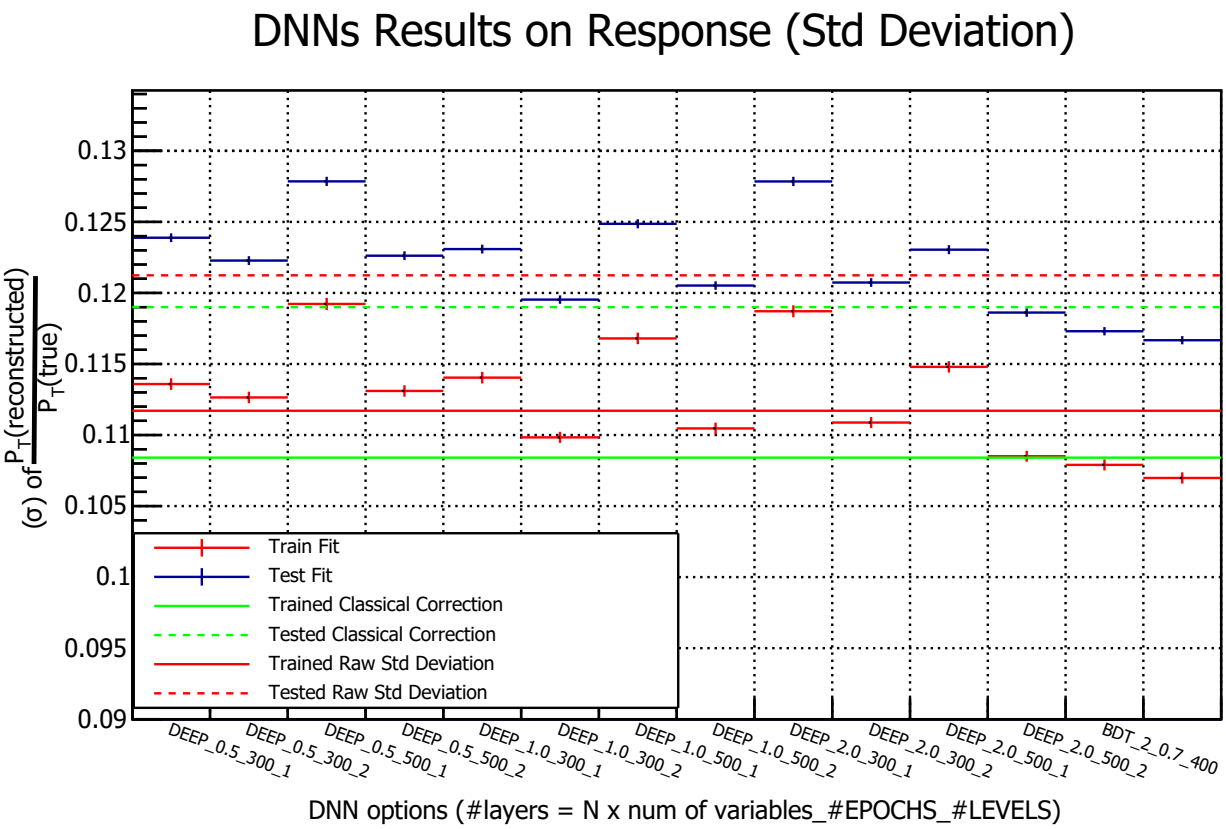
3.5 Deep Neural Networks

Οι ρυθμίσεις για την εκπαίδευση των DNN (Deep Neural Networks) δεν διαφέρουν από αυτές που είδαμε στο 3.3.2 για τα Νευρωνικά Δίκτυα (εδώ δεν υπάρχουν οι επιλογές για τις μεθόδους εκπαίδευσης BP και BFGS).

```
vector<double>inputs{0.5,1.0,2.0};
vector<int>cycles{300,500};
vector<int>nol{1,2};
// LOOP OVER VECTORS

factory->BookMethod(dataloader,TMVA::Types::kDL,"DNN",<options>)
    ;
```





Σχήμα 3.23: Τα αποτελέσματα για μέση τιμή (mean) και διασπορά (sigma) για τα παραπάνω option εκπαίδευσης των Deep Neural Networks. Οι κόκκινες και πράσινες συνεχόμενες και διακεκομμένες γραμμές είναι ίδιες με την περίπτωση των BDTs. Στα 2 τελευταία (δεξιά) κελιά έχουν τοποθετηθεί τα βέλτιστα δέντρα απόφασης, όπως αυτά αποφασίστηκαν στην προηγούμενη παράγραφο, για σύγκριση.

Η περίπτωση των Deep Neural Networks δίνει μια πολύ πιο καθαρή εικόνα για τα αποτελέσματα. Η καλύτερη απόδοση σχετικά με τη μέση τιμή ταυτίζεται με αυτή για την διασπορά και παρατηρείται για τα DNNs με $N \times 2$ και $N \times 1$ hidden layers, 500 εποχές και βάθος 2 επίπεδα, ενώ δεν εμφανίζεται σε καμία από τις περιπτώσεις υπερεκπαίδευση.

Κεφάλαιο 4

Διερεύνηση των αποτελεσμάτων

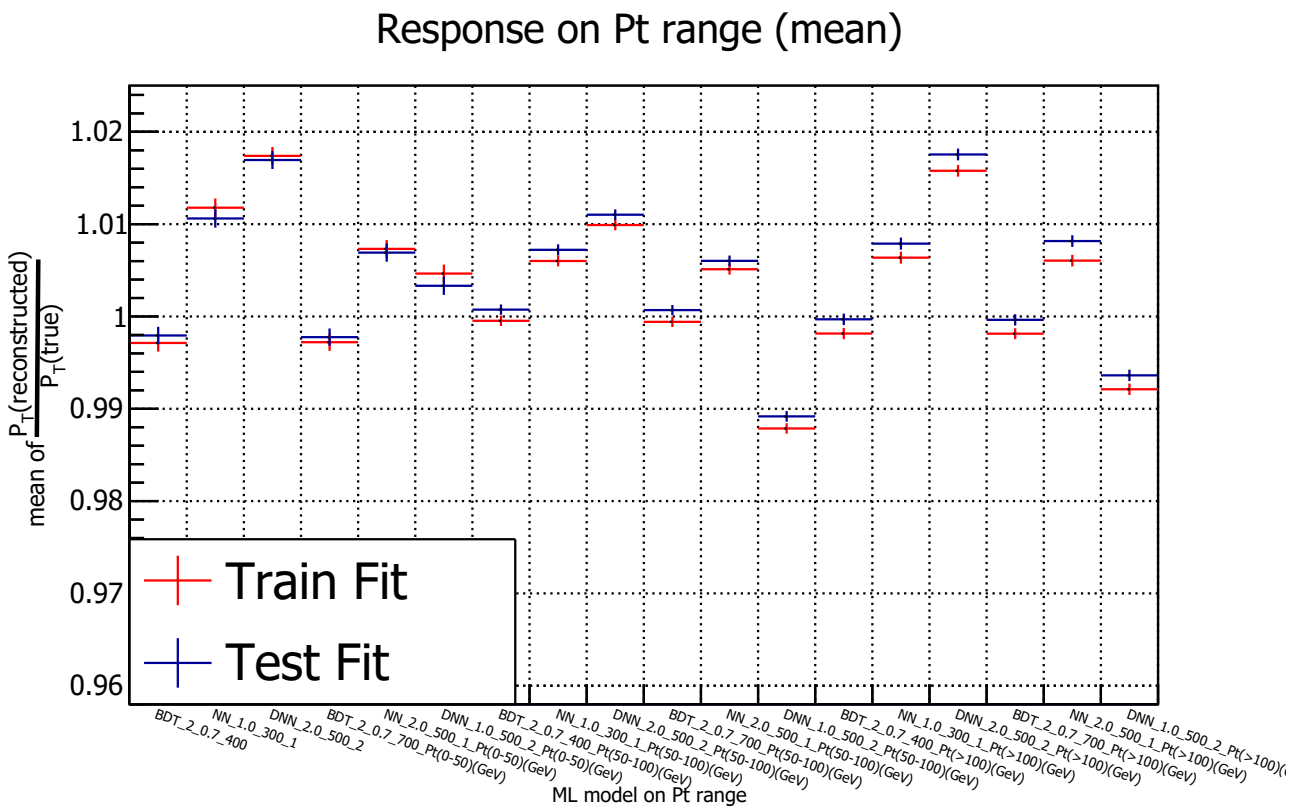
Καταλήξαμε στις κατάλληλες παραμέτρους εκπαίδευσης κάθε μεθόδου ώστε να έχουμε μέση τιμή της απόκρισης (response) κοντά στη μονάδα και όσο το δυνατόν μικρότερη διασπορά, λαμβάνοντας πάντα υπ όψιν να μην υπάρχει overtraining. Παρόλο που φαίνεται να έχει επιτευχθεί ο βασικός μας στόχος, πρέπει να ελέγξουμε ακόμα ορισμένες συμπεριφορές των αλγορίθμων.

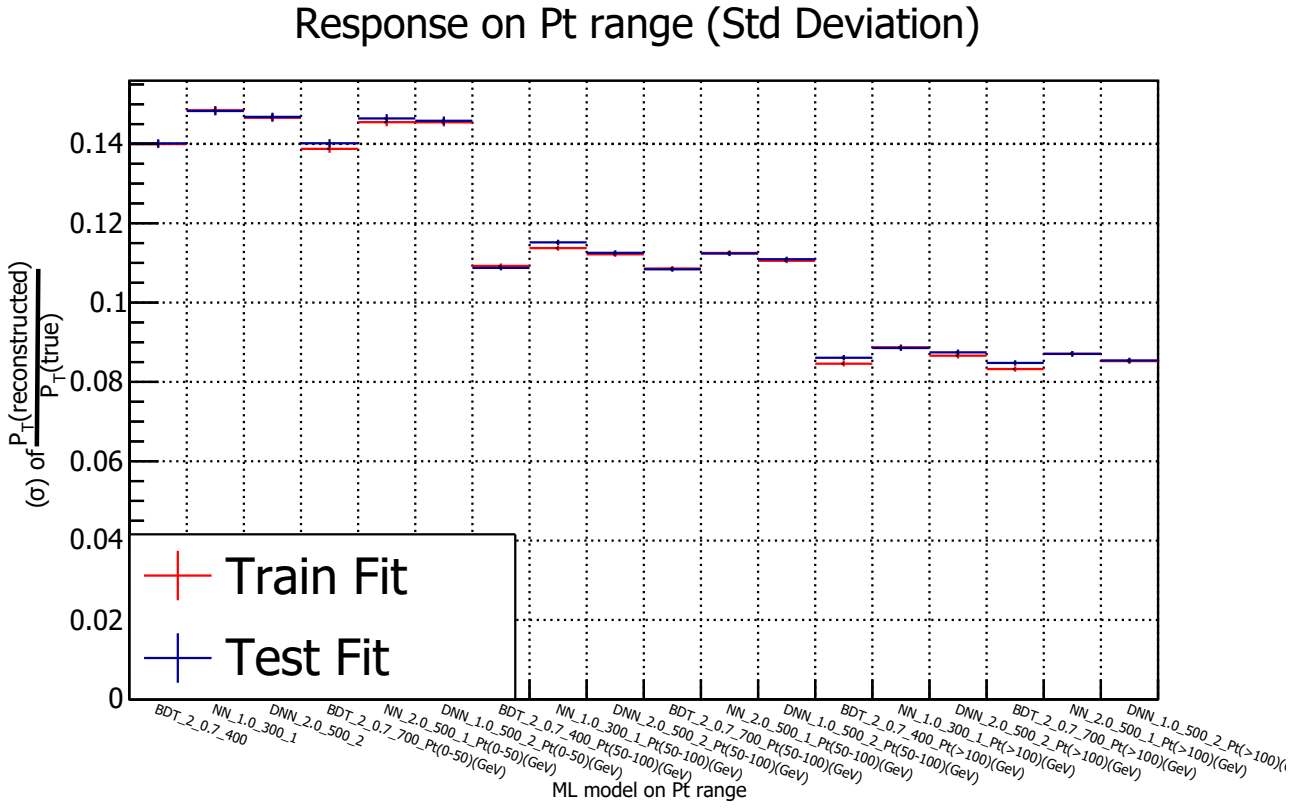
4.1 Απόκριση στον χώρο των ορμών και της ψευδωκύτητας η

Για να ελεγχθεί η απόδοση αλλά και η εγκυρότητα της εκπαίδευσης των μεθόδων μηχανικής μάθησης που δοκιμάσαμε στην παρούσα εργασία, χρειάζεται να ελέγξουμε την απόδοση των αλγορίθμων σε συγκεκριμένα διαστήματα της μεταβλητής της ορμής κάθε πίδακα (P_t) καθώς και της ψευδωκύτητας (η). Ο έλεγχος έγινε χρησιμοποιώντας τα δύα καλύτερα BDTs, τα δύο καλύτερα Neural Networks και τα δύο καλύτερα DNNs, όπως αυτά προέκυψαν από την ανάλυση του Κεφαλαίου 3.

Αρχικά διαλέγουμε 3 διαστήματα για την ορμή (genJetPt) κάθε πίδακα, ένα με χαμηλές τιμές, ένα με μεσαίες και ένα με υψηλές. Σε αυτά τα διαστήματα υπολογίστηκε εκ νέου η απόκριση (Response) της προσαρμογής των δεδομένων και θα συγκρίνουμε τις επιμέρους μέσες τιμές και διασπορές.

- genJetPt: [20,50] , [50,100] , [100,>100] [GeV]





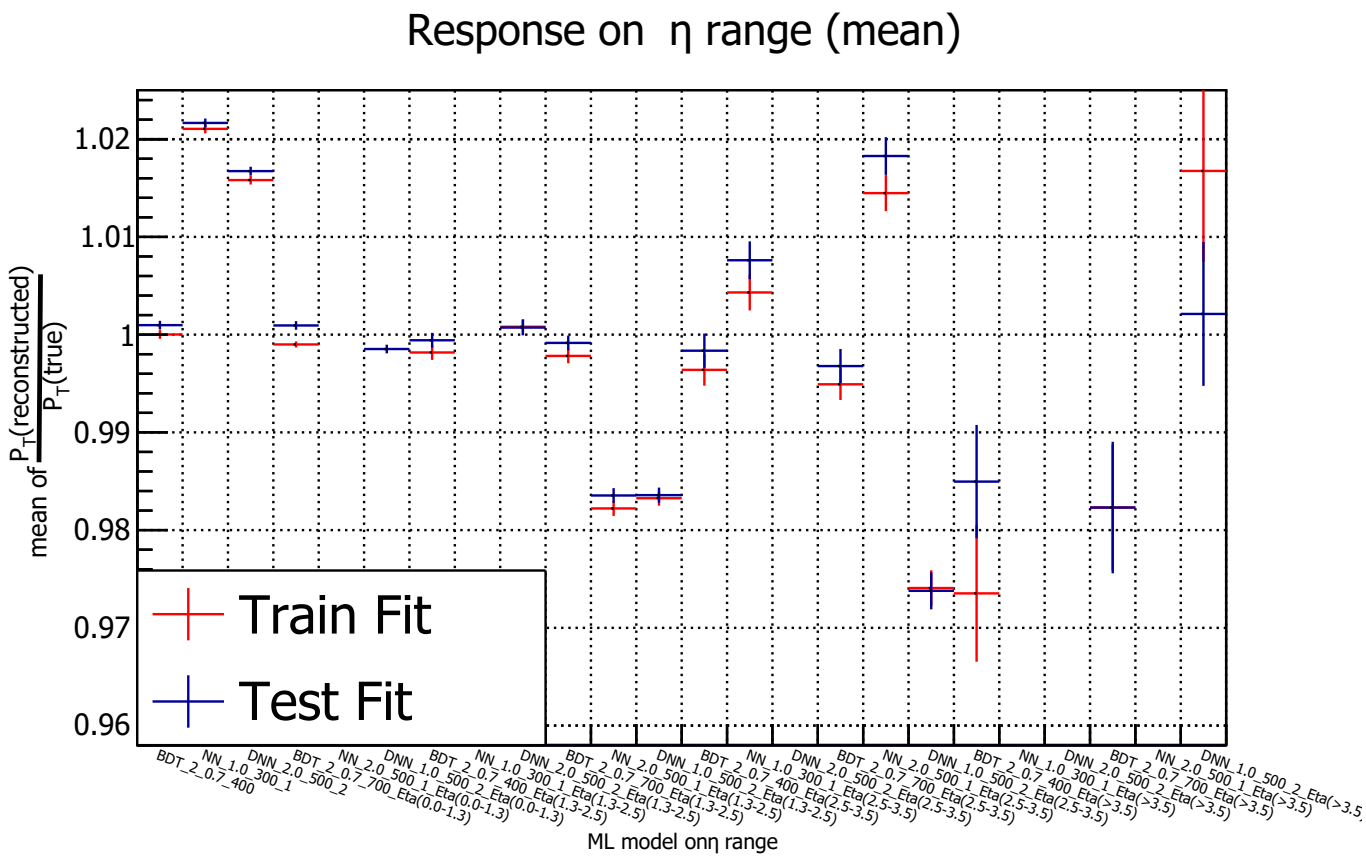
Σχήμα 4.1: Τα αποτελέσματα για μέση τιμή (mean) και διασπορά (sigma) για κάθε διάστημα ορμής του πίδακα. Αριστερά είναι οι μέσες τιμές και οι διασπορές για τις μικρότερες τιμές της μεταβλητής, στο κέντρο οι μεσαίες τιμές και δεξιά οι μεγαλύτερες. Οι κόκκινες και πράσινες συνεχόμενες και διακεκομμένες γραμμές είναι οι τιμές για τη μέση τιμή και τη διασπορά των δεδομένων (training set και testing set) για το σύνολο των τιμών της ορμής, όπως αυτές μετρώνται από τον ανιχνευτή (raw data) ή προκύπτουν απο την κλασική προσαρμογή των δεδομένων (classical correction).

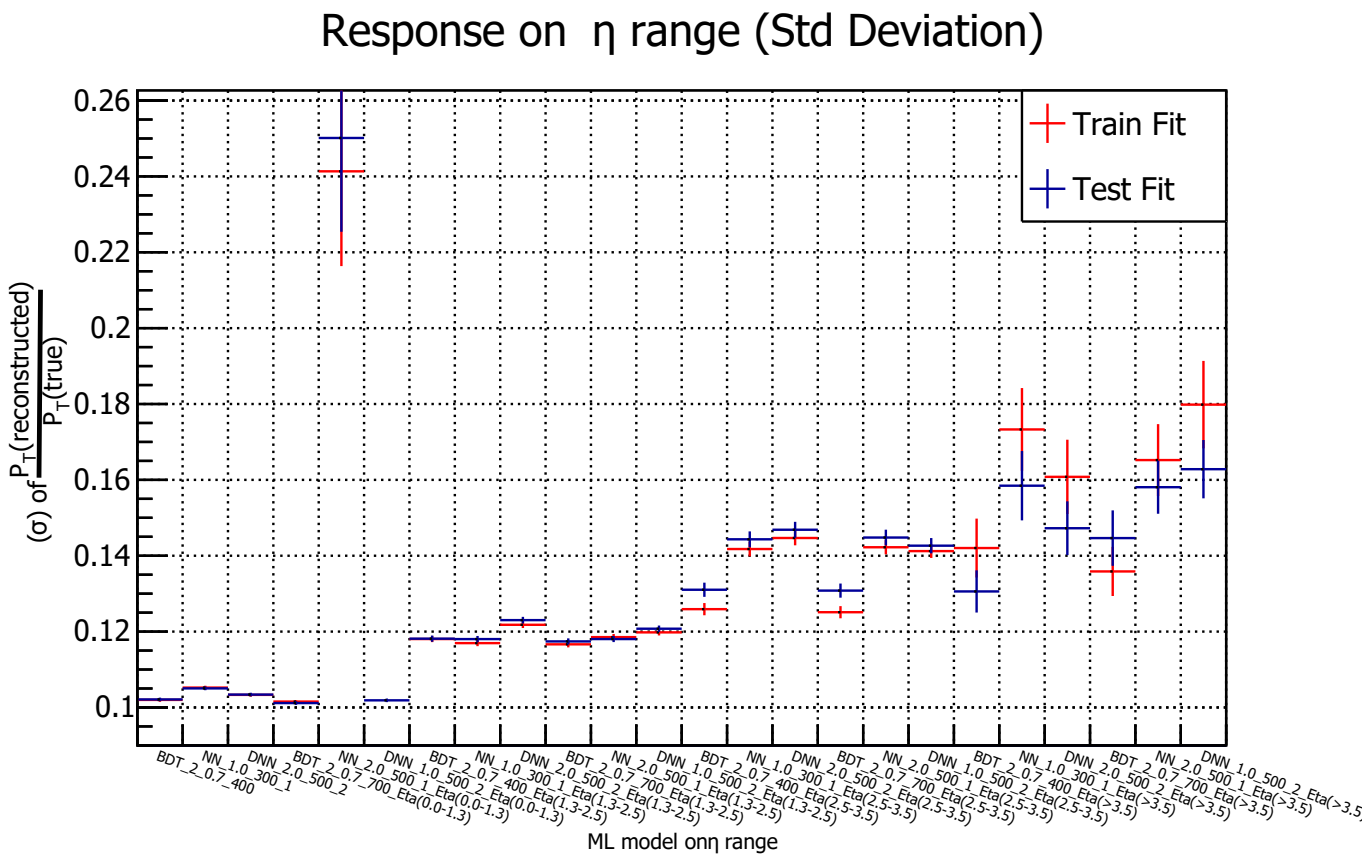
Παρατηρούμε ότι για μικρές τιμές της ορμής οι αλγόριθμοι μηχανικής μάθησης δεν καταφέρνουν να βελτιώσουν την απόδοση, αφού, αν και η μέση τιμή (mean) είναι πολύ πιο κοντά στη μονάδα από ότι στην κλασική μέθοδο, η διασπορά μεγαλώνει κατά πολύ. Για μεσαίες τιμές της ορμής οι αλγόριθμοι εμφανίζουν εξαιρετική απόδοση, καθώς όχι μόνο βελτιώνουν την μέση τιμή της απόκρισης, φέρνοντας την πολύ πιο κοντά στη μονάδα, αλλά μειώνουν αισθητά και την διασπορά της κατανομής του μεγέθους.

Για την ψευδωκύτητα η, η οποία περιγράφηκε αναλυτικά στο Κεφάλαιο 3.1, ακολουθούμε παρόμοια διαδικασία. Χωρίζουμε το σύνολο των τιμών της η σε 4 υποσύνολα, δύο για μικρές γωνίες, ένα για μεσαίες και ένα για μεγάλες. Συγκεκριμένα

- ψευδωκύτητα η: [0,1.3] , [1.3,2.5] , [2.5,3.5] , [>3.5]

Στα παραπάνω σύνολα λαμβάνουμε υπόψιν την απόλυτη τιμή της η, για να καλύψουμε όλο τον όγκο του κυλίνδρου του ανιχνευτή.



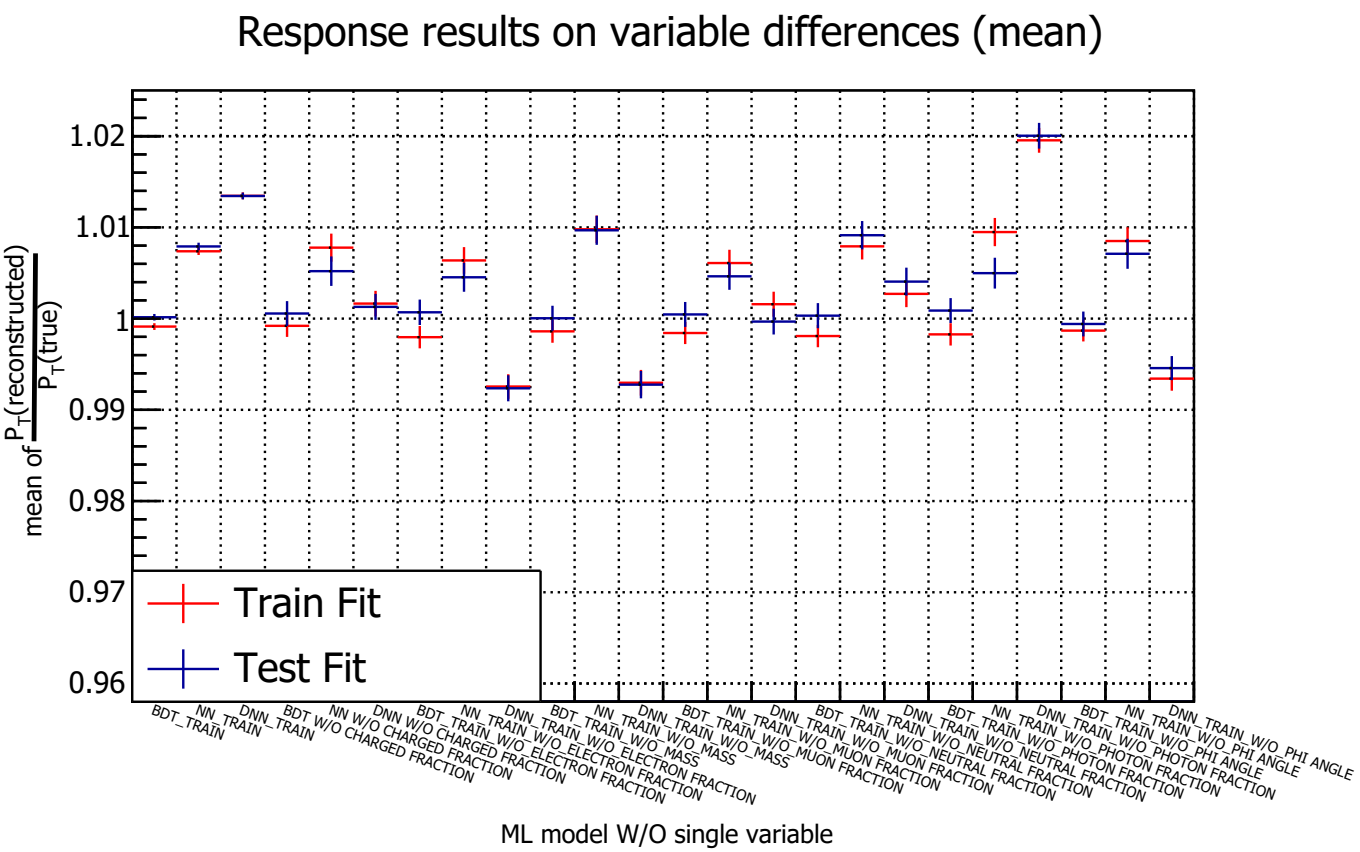


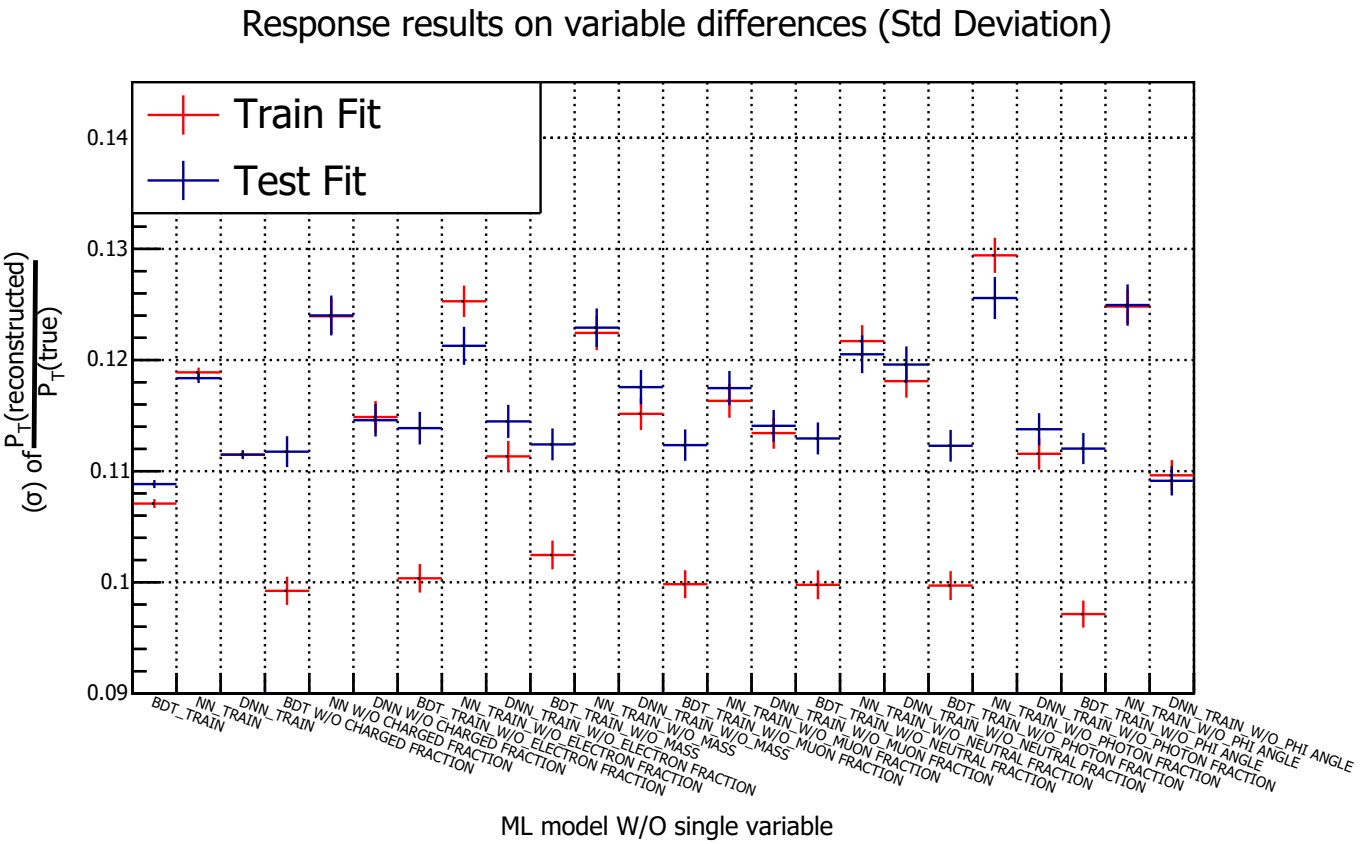
Σχήμα 4.2: Τα αποτελέσματα για μέση τιμή (mean) και διασπορά (sigma) για κάθε διάστημα της ψευδωκύτητας η του πίδακα, με τις πιο μικρές στα αριστερά, τις πιο μεγάλες στα δεξιά και στο κέντρο τις ενδιάμεσες. Οι κόκκινες και πράσινες συνεχόμενες και διακεκομμένες γραμμές ακολουθούν την ίδια λογική με παραπάνω.

Τα παραπάνω αποτελέσματα παρουσιάζουν ενδιαφέρον, καθώς παρατηρούμε ότι η εκτίμηση για την απόκριση (Response) φαίνεται να χειροτερεύει όσο αυξάνεται η τιμή της η. Δεν χρειάζεται όμως να μας παραξενεύει αυτό το αποτέλεσμα, καθώς είναι επιστημονικά αναμενόμενο. Πράγματι, σε αυτό το σύνολο τιμών για την ψευδωκύτητα αναμένουμε χειρότερη ποιότητα στις μετρήσεις, άρα και στα δεδομένα για την εκπαίδευση των αλγορίθμων, καθώς μεγαλύτερες τιμές του η σημαίνει μετρήσεις από τα άκρα (καπάκια) του κυλίνδρου, ή από στοιχεία κοντά σε αυτά, οπότε δεν έχουν την ίδια ακρίβεια με περιπτώσεις που ο πίδακας πέφτει κάθετα στον κύλινδρο-ανιχνευτή, δηλαδή όταν έχουμε μικρές τιμές της η (Κεφάλαιο 3.1).

4.2 Η σημασία κάθε μεταβλητής μεγέθους στο τελικό αποτέλεσμα

Ο έλεγχος των μεθόδων μηχανικής μάθησης που καταλήξαμε συνεχίζεται λαμβάνοντας υπόψιν αυτη τη φορά πώς θα επηρεαζόταν ο κάθε εκτιμητής (BDT, NN, DNN) αν αφαιρούσαμε κάποια από τις μεταβλητές του. Για να ελεγχθεί αυτό, εκπαιδεύσαμε κάθε έναν από τους βέλτιστους εκτιμητές που βρήκαμε στο Κεφάλαιο 3, αφαιρώντας κάθε φορά μία από τις μεταβλητές που είχαμε εισάγει αρχικά.





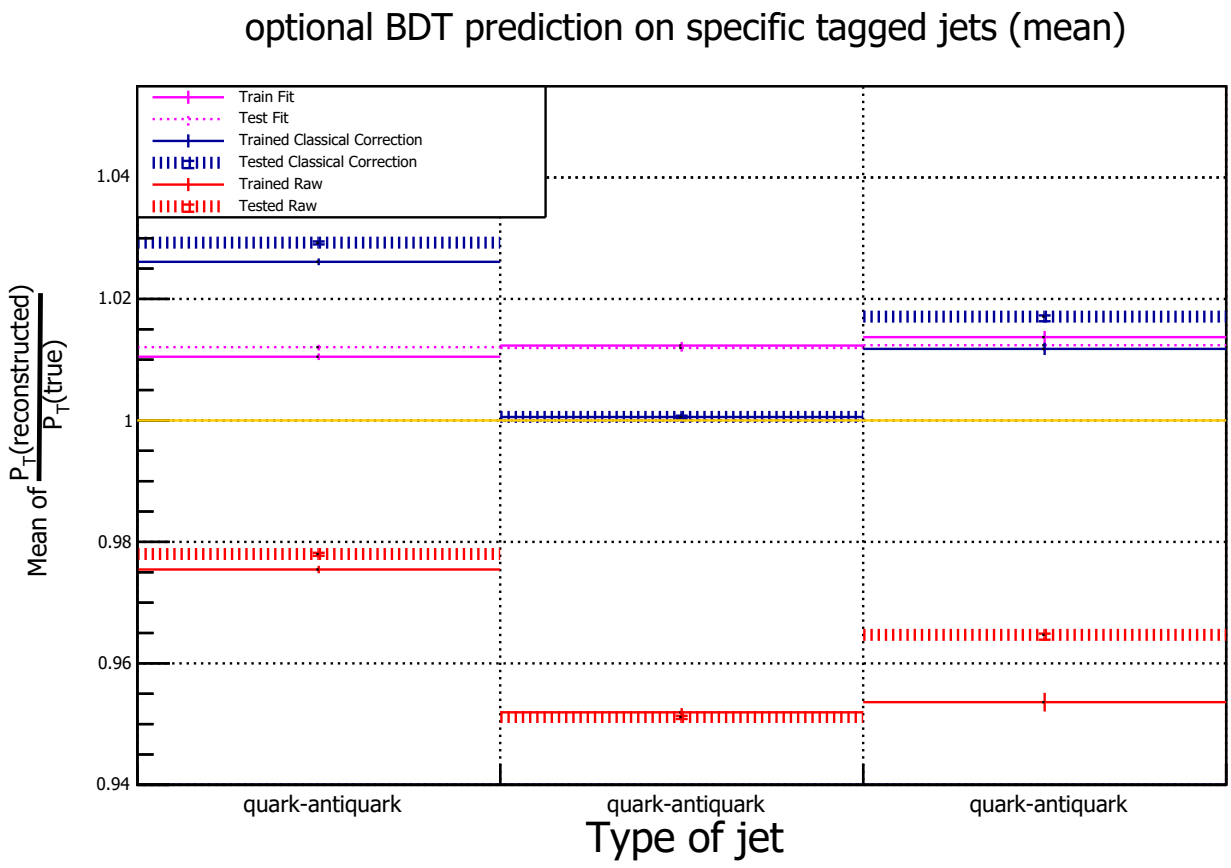
Σχήμα 4.3: Η συμπεριφορά των βέλτιστων εκτιμητών χωρίς κάποια από τις αρχικές μεταβλητές. Τα τρία πρώτα σημεία είναι η μέση τιμή και η διασπορά για το βέλτιστο BDT, Neural Network και Deep NN του Κεφαλαίου και προστίθενται για σύγκριση.

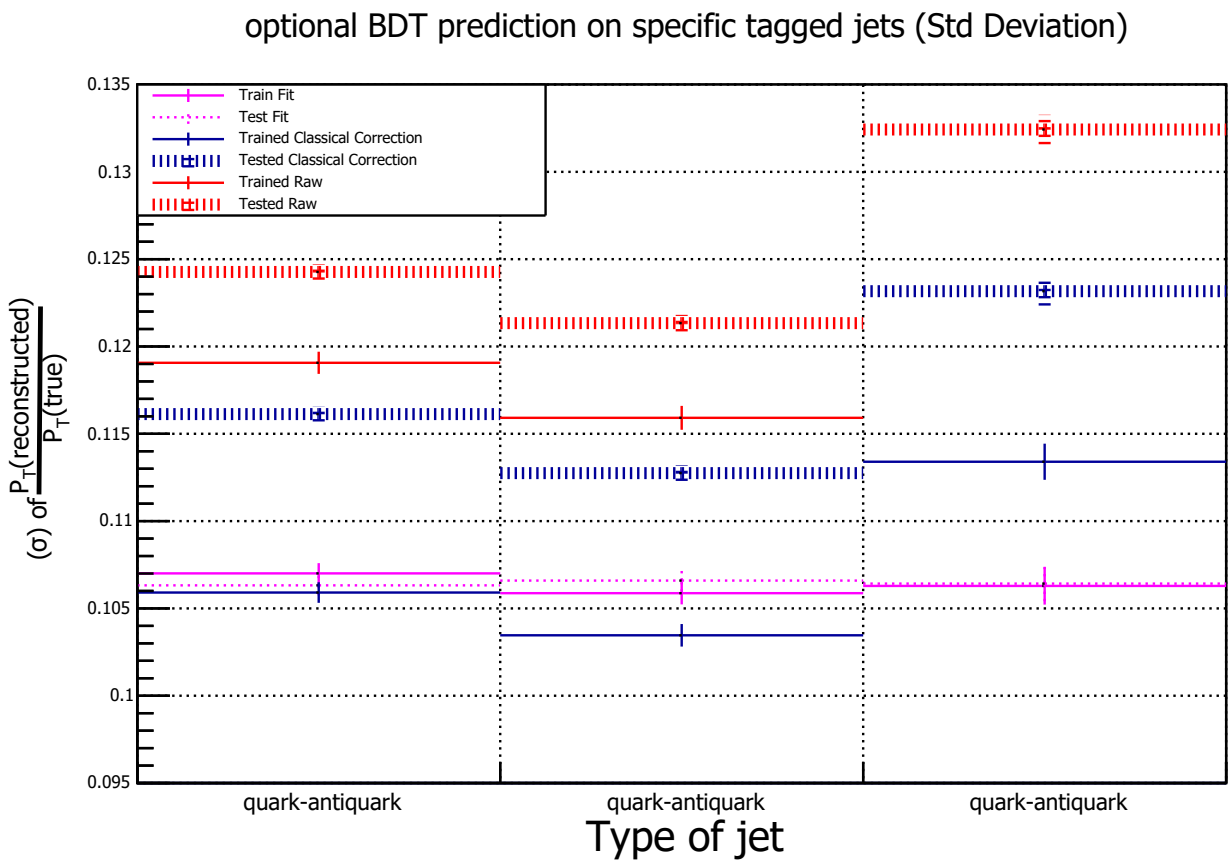
Ελέγχοντας τα νέα αποτελέσματα για τη μέση τιμή και τη διασπορά της απόκρισης σε κάθε περίπτωση, βλέπουμε πως με την έλλειψη κάποιας μεταβλητής, ο αντίστοιχος εκτιμητής είτε εμφανίζει χειρότερη μέση τιμή ή η διασπορά της απόκρισης, ή εμφανίζει υπερεκπαίδευση. Συμπεραίνουμε ότι τα αποτελέσματα είναι εμφανώς κατώτερα από την περίπτωση που συμπεριλαμβάνουμε όλες τις μεταβλητές.

Ο παραπάνω, αναγκαίος, έλεγχος επιβεβαιώνει τα αποτελέσματα του Κεφαλαίου 3 για τις βέλτιστες παραμέτρους που χρειάζεται κάθε διαδικασία μηχανικής μάθησης από αυτές που επιλέξαμε να μελετήσουμε (BDTs, NNs, DNNs) καθώς δεν εμφανίζεται κάποιο bias στα υποσύνολα των τιμών για την ορμή P_t και ψευδωκύτητα η και τα αποτελέσματα σε αυτά τα υποσύνολα ταιριάζουν απόλυτα με τα φυσικά χαρακτηριστικά του πίδακα και του πειράματος. Επίσης, ο τελευταίος έλεγχος έδειξε ότι αρχικά έγινε κατάλληλη επιλογή μεταβλητών και δεν υπάρχει κάποια περιττή μέτρηση του πίδακα στον ανιχνευτή που μπορεί να αγνοηθεί στην διαδικασία βελτίωσης της απόκρισης του τελευταίου.

4.3 Πίδακες γλουονίων και συγκεκριμένων γεύσεων κουαρκ

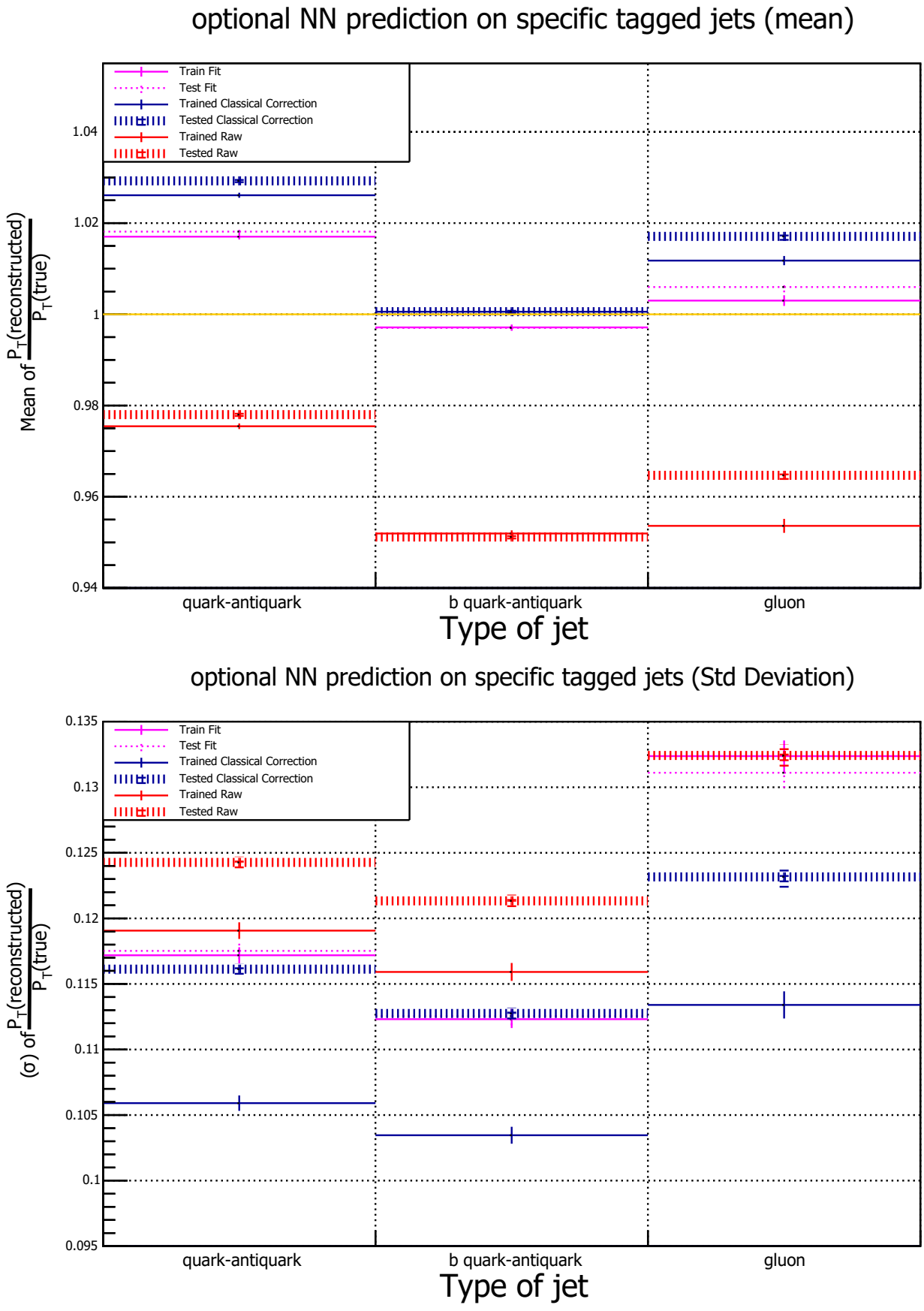
Τέλος, αξίζει να γίνει μια ιδιαίτερη διερεύνηση για τον τρόπο επιρροής των αποτελεσμάτων ανάλογα με τη γεύση των κουαρκ από τα οποία έχει αντιστοιχηθεί ότι προέρχεται κάθε πίδακας αλλά και των γλουονίων του. Η βασική μελέτη έγινε για όλα τα κουρκ εκτός του b κουαρκ, μαζί με τα αντίστοιχα αντικουαρκ, μια ξεχωριστή μελέτη για το b κουαρκ (συμπεριλαμβανομένου και του b αντικουαρκ) και τελευταία μια μελέτη για τα γλουόνια. Για κάθε μία από τις παραπάνω περιπτώσεις έγινε εκ νέου εκπαίδευση των βέλτιστων εκτιμητών από το Κεφάλαιο 3, αυτή τη φορά όμως, με τη βοήθεια της μεταβλητής jetFlavor μόνο για τους πίδακες που προέρχονται από τις παραπάνω κατηγορίες. Τα αποτελέσματα παρουσιάζονται παρακάτω.





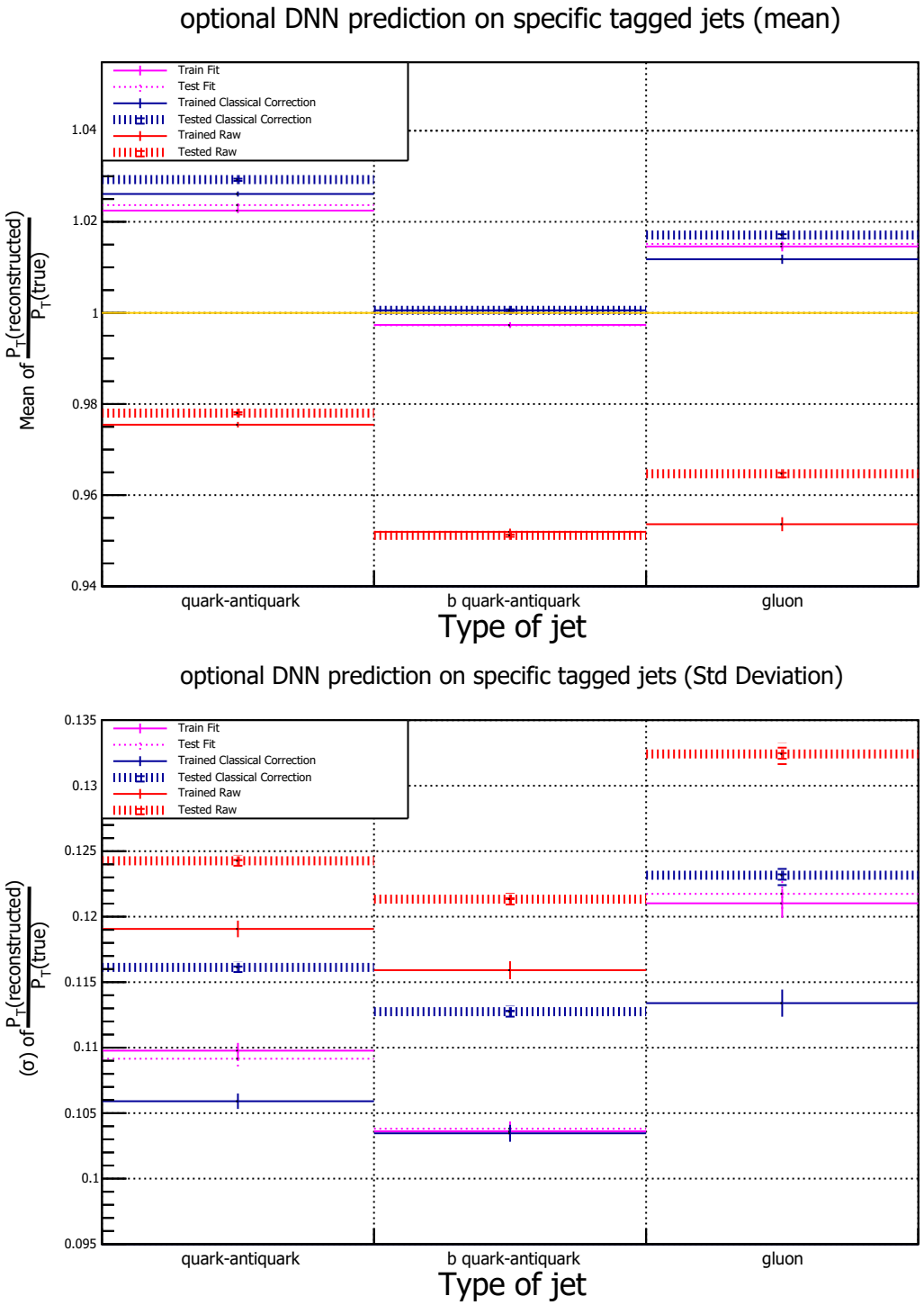
Σχήμα 4.4: Απόκριση του βέλτιστου Boosted Decision Tree για τις τρεις παραπάνω περιπτώσεις. Τα κόκκινα σημεία είναι για τα raw δεδομένα, με συνεχόμενη γραμμή αυτά από το training set και διακεκομμένη αυτά από το testing set. Την ίδια λογική ακολουθούν και τα μπλε σημεία που είναι οι αντίστοιχες τιμές μέσης τιμής (mean) και διασποράς (sigma) για την κλασική μέθοδο διόρθωσης της απόκρισης.

Η επιμέρους απόκριση των BDTs δεν μπορεί να θεωρηθεί καλύτερη από ότι αυτή της γενικής περίπτωσης, καθώς, παρότι οι νέες τιμές για την μέση τιμή και τη διασπορά της απόκρισης είναι καλύτερες, δηλαδή πιο κοντά στη μονάδα η πρώτη και μικρότερη η δεύτερη, εμφανίζεται και στις τρεις περιπτώσεις το φαινόμενο της υπερεκπαίδευσης, αφού τα αποτελέσματα για το fit εκπαίδευσης και ελέγχου (magenta συνεχόμενη και διακεκομμένη γραμμή) έχουν έντονα διαφορετική απόσταση από ότι οι αντίστοιχες τιμές για τα training και testing set (απόσταση διακεκομμένη με συνεχόμενη μπλε/κόκκινη γραμμής).



Σχήμα 4.5: Απόκριση του βέλτιστου Neural Network για τις παραπάνω περιπτώσεις κουαρκ και γλουονίων. Κόκκινα και μπλε σημεία ακολουθούν την ίδια λογική με παραπάνω.

Η περίπτωση του νευρωνικού δικτύου για την παλινδρόμηση στους πίδακες συγκεκριμένων γεύσεων και γλουονίων παρουσιάζει διαφορές από αυτή των BDTs. Αυτή τη φορά έχουμε βελτίωση της μέση τιμής της απόκρισης, αλλά καμία ουσιαστική βελτίωση στη διασπορά της νέας κατανομής. Ανεξαρτήτως αυτών, εμφανίζεται πάλι το φαινόμενο της υπερεκπαίδευσης.



Σχήμα 4.6: Απόκριση μετά την εκπαίδευση του βέλτιστου Deep Neural Network για τις τρεις παραπάνω περιπτώσεις. Κόκκινα και μπλε σημεία ακολουθούν την ίδια λογική με παραπάνω.

Τα Deep NNs, παρότι διαθέτουν αρκετές παραπάνω δυνατότητες από τα απλά νευρωνικά δίκτυα, παρουσιάζουν την ίδια εικόνα με τα τελευταία, επομένως δεν παρατηρείται κάποια ουσιαστική διαφορά ή βελτίωση. Στις 3 περιπτώσεις των αλγορίθμων εκμάθησης δεν παρουσιάστηκε κάποια βελτίωση της απόδοσης τους μελετώντας συγκεκριμένα είδη αδρονικών πιδάκων.

Κεφάλαιο 5

Συμπεράσματα και μελλοντικές προοπτικές

Η παρούσα εργασία ξεκίνησε με την φιλοδοξία να βελτιωθεί η βαθμονόμηση πιδάκων αδρονίων στο πείραμα CMS κάνοντας χρήση μεθόδων μηχανικής μάθησης. Η βελτίωση αυτή αφορούσε τόσο την προσπάθεια προσέγγισης της μονάδας για την μέση τιμή της κατανομής της απόκρισης , δηλαδή του λόγου $Response = \frac{p_T(reconstructed)}{p_T(true)}$, όσο και την μείωση της διασποράς γύρω από αυτή. Ένας τέτοιος στόχος θα μπορούσε να προσφέρει μεγάλη βελτίωση συνολικά στην ανάλυση των δεδομένων του πειράματος και στην εξέλιξη της μελέτης των υποατομικών σωματιδίων, καθώς προσφέρει έναν τρόπο βαθμονόμησης των πιδάκων χωρίς να απαιτεί έναν συγκεκριμένο παράγοντα διόρθωσης αλλά και μεγαλύτερη ακρίβεια. Για την επίτευξη αυτού του σκοπού μελετήθηκαν οι παράμετροι εκπαίδευσης τριών διαφορετικών μεθόδων μηχανικής μάθησης, ενός ενισχυμένου δέντρου απόφασης (BDT), ενός τεχνητού νευρωνικού δικτύου (ANN) και ενός νευρωνικού δικτύου βαθιάς μάθησης (DNN). Έγινε επίσης προσπάθεια εμπλουτισμού των μεταβλητών του πίδακα που αξιοποιούνται για την βαθμονόμηση, καθώς η κλασική μέθοδος βασίζεται μόνο σε κινηματικές μεταβλητές.

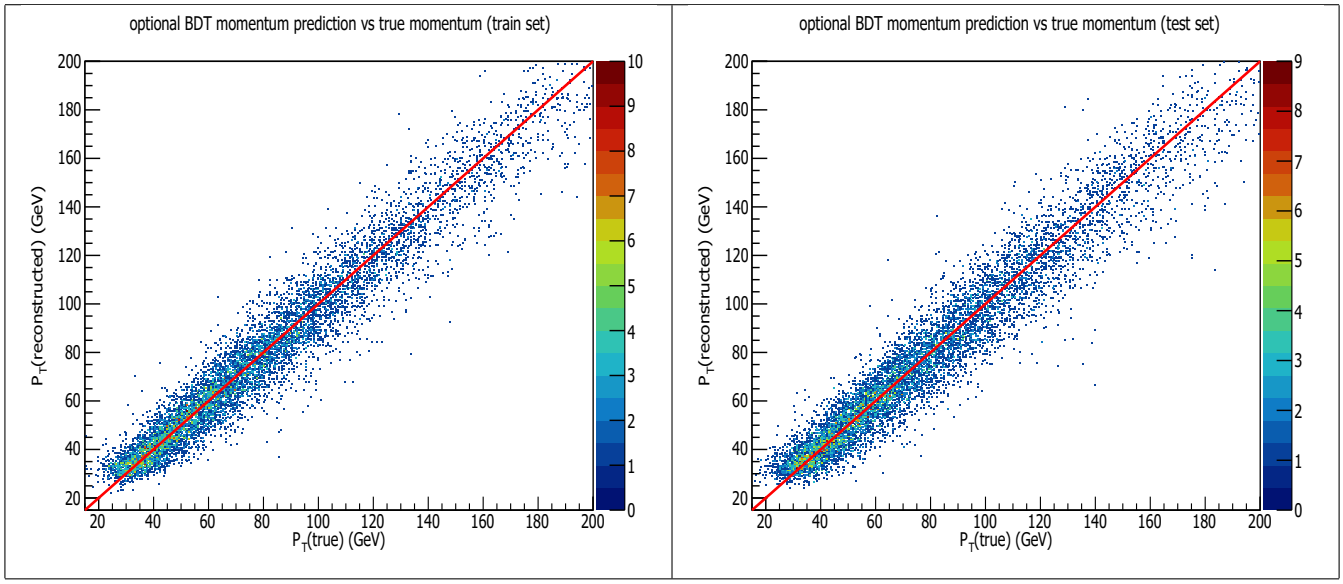
Η ανάλυση των τριών μεθόδων οδήγησε στην εύρεση ορισμένων κατάλληλων ρυθμίσεων ώστε να επιτευχθεί μία βελτίωση της μέσης τιμής και της απόκρισης χωρίς όμως να εμφανίζεται υπερεκπαίδευση του μοντέλου, όπως αυτή περιγράφεται στο Κεφάλαιο 2. Συγκεκριμένα, για τα BDTs καταλήξαμε ότι καλύτερα fit μας δίνει ο αλγόριθμος με δέντρα απόφασης βάθους 2, shrinkage 0.7 και 700 δέντρα (για τον αλγόριθμο AdaBoosting). Παρατηρήθηκε ότι όσο αυξανόταν ο αριθμός των δέντρων και το βάθος τους, εμφανιζόταν πολύ μεγάλο overtraining.

Για τα τεχνητά νευρωνικά δίκτυα παρατηρήθηκε μεγάλη διαφορά στην συμπεριφορά του αλγορίθμου Back-Propagation (BP) και του αλγορίθμου BFGS (περιγράφονται στο Κεφάλαιο 2). Όπως εξηγήθηκε και στο Κεφάλαιο 3, ενώ ο αλγόριθμος BP με 2 επίπεδα εμφανίζει έντονη υπερεκπαίδευση, όταν είναι ρυθμισμένος για 1 επίπεδο δίνει καλύτερη μέση τιμή για την απόκριση χωρίς υπερεκπαίδευση. Η μέθοδος BFGS για τα νευρωνικά δίκτυα, στην προσέγγιση της μέσης τιμής δεν δίνει όσο καλά αποτελέσματα δίνει η μέθοδος BFGS, ενώ η μικρή βελτίωση που εμφανίζει για την διασπορά της απόκρισης εμφανίζει μεγάλη αστάθεια και overtraining. Τελικά καταλήξαμε ως πιο αποδοτικές ρυθμίσεις αυτές για NNs της μεθόδου BP με 1 επίπεδο, $N \cdot 1$ και $N \cdot 2$ νευρώνες, για 300 και 500 εποχές.

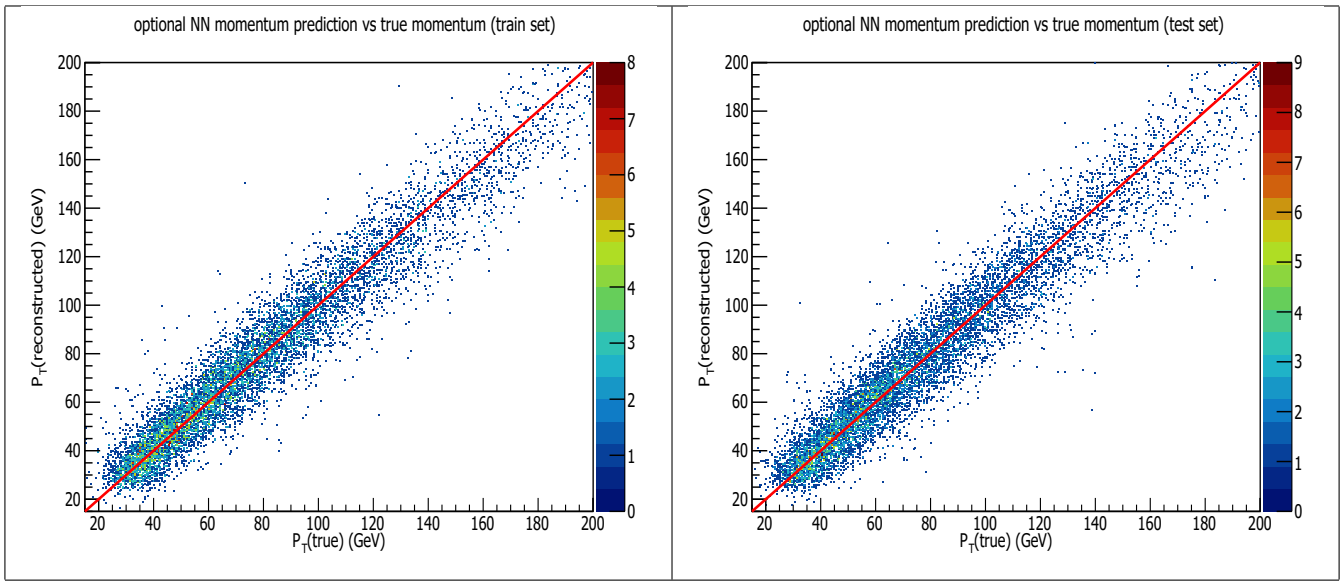
Η περίπτωση των νευρωνικών δικτύων βαθιάς μάθησης (BDTs) εμφανίζει πολύ θετικά αποτελέσματα, καθώς με λίγες ρυθμίσεις καταφέραμε να έχουμε εξαιρετική προσέγγιση της μονάδας για την μέση τιμή της απόκρισης, ενώ για τις ίδιες ρυθμίσεις, δηλαδή για 2 επίπεδα, $N \cdot 1$ και $N \cdot 2$ νευρώνες, 300 και 500 εποχές είχαμε και την μικρότερη διασπορά στην κατανομή χωρίς εμφάνιση overtraining.

Η απόδοση των αλγορίθμων που εκπαιδεύσαμε στα δεδομένα μας μπορεί να φανεί και στα παρακάτω διαγράμματα, όπου γίνεται ένα plot της τιμής της εγκάρσιας ορμής που προβλέπεται από τον καλύτερο από τα 3 είδη αλγορίθμων μηχανικής μάθησης που μελετήσαμε ($P_T(reconstructed)$) προς την πραγματική εγκάρσια ορμή του πίδακα ($P_T(true)$). Όπως η από-

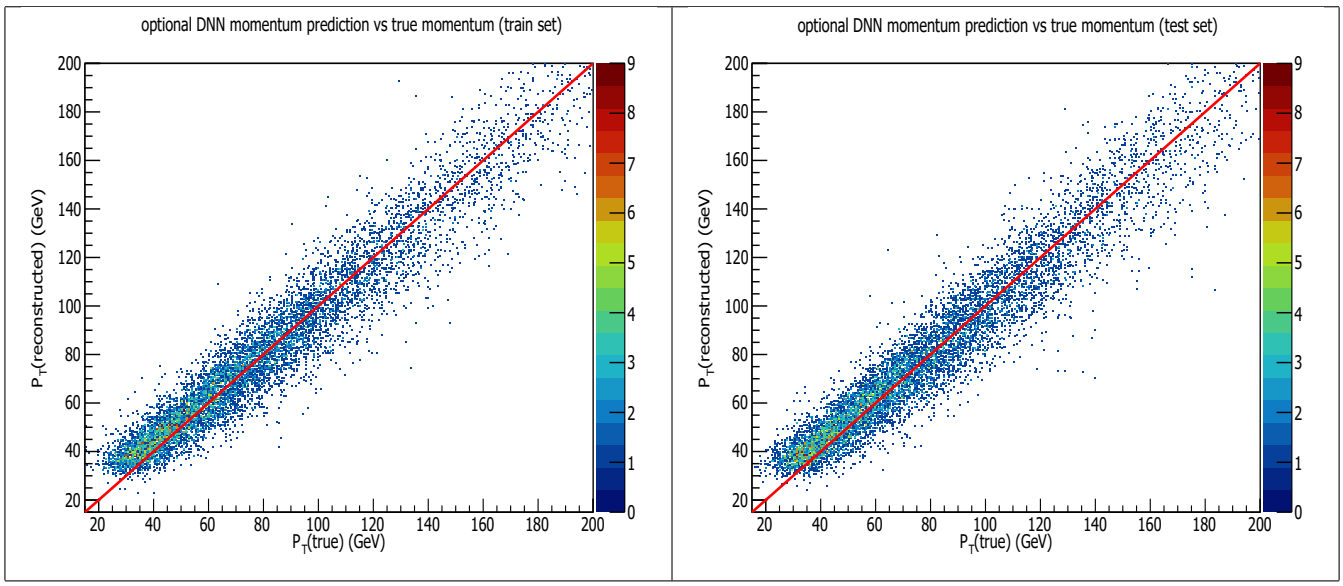
κριση $\frac{p_T(reconstructed)}{p_T(true)}$ αποτελεί τον λόγο αυτών των δύο μεγεθών και στην ιδανική περίπτωση θα ήταν μία συνάρτηση δέλτα στη μονάδα, έτσι και όταν σχεδιάζουμε γραφικά το ένα μέγεθος προς το άλλο, στόχος μας είναι η καλύτερη προσέγγιση μίας συνάρτησης $y = x$.



Σχήμα 5.1: Η ανακατασκευασμένη ορμή του πίδακα όπως αυτή προβλέπεται από το βέλτιστο δέντρο εκπαίδευσης της εργασίας προς την πραγματική ορμή. Η κόκκινη γραμμή είναι η γραφική παράσταση της συνάρτησης $y = x$.



Σχήμα 5.2: Η ανακατασκευασμένη ορμή του πίδακα όπως αυτή προβλέπεται από το βέλτιστο τεχνητό νευρωνικό δίκτυο της εργασίας προς την πραγματική ορμή. Η κόκκινη γραμμή είναι η γραφική παράσταση της συνάρτησης $y = x$.



Σχήμα 5.3: Η ανακατασκευασμένη ορμή του πίδακα όπως αυτή προβλέπεται από το βέλτιστο νευρωνικό δίκτυο βαθιάς μάθησης της εργασίας με προς την πραγματική ορμή. Η κόκκινη γραμμή είναι η γραφική παράσταση της συνάρτησης $y = x$.

Ο αρχικός σκοπός όμως αυτής της εργασίας επιβεβαιώνεται και από την περαιτέρω μελέτη που έγινε στο Κεφάλαιο 4. Είδαμε ότι η κλασική μέθοδος βαθμονόμησης παράγει έναν συντελεστή διόρθωσης συναρτήσει της εγκάρσιας ορμής και της ψευδωκύτητας. Ο έλεγχος των βέλτιστων αλγορίθμων μας σε διάφορα διαστήματα των συνόλων τιμής της εγκάρσιας ορμής και της ψευδωκύτητας η έδειξε ότι δεν υπάρχει κάποιο bias στις μεθόδους εκπαίδευσης και τα αποτελέσματα της εκπαίδευσης συνάδουν με την φυσική του πειράματος. Επίσης, στο Κεφάλαιο 4, δείξαμε ότι η επιλογή των μεταβλητών ήταν σωστή και η αφαίρεση κάποιας από αυτές θα μείωνε την ακρίβεια των προβλέψεων από τους αλγορίθμους. Επιπροσθέτως, η προσπάθεια μελέτης της ακρίβειας σε πίδακες συγκεκριμένης προέλευσης (γλουονίων, b-quark κλπ) δεν έδειξε βελτίωση των προβλέψεων σε κάποια συγκεκριμένη κατηγορία, κυρίως λόγω της εμφάνισης υπερεκπαίδευσης.

Τέλος, δεν θεωρούμε ότι η συγκεκριμένη εργασία κλείνει κάποια συζήτηση γύρω από την χρήση μεθόδων μηχανικής μάθησης για την προσπάθεια βελτίωσης της βαθμονόμησης πιδάκων αδρονίων στο πείραμα CMS, αλλά αντιθέτως επιχειρεί να την διευρύνει περαιτέρω. Πρώτο πιθανό βήμα θα ήταν η εφαρμογή των παραπάνω μοντέλων, που προέκυψαν μετά την εκπαίδευση των αλγορίθμων στα δεδομένα της εξομοίωσης, σε νέα, πραγματικά, δεδομένα αλλά και για τον έλεγχο της πρόβλεψης μάζας σωματιδίων εντός των ίδιων δεδομένων εξομοίωσης. Επίσης, είναι προφανές ότι η παρούσα εργασία δεν θα μπορούσε να μελετήσει όλες τις μεθόδους μηχανικής μάθησης με όλες τις παραμέτρους που εμπεριέχει η κάθε μία, επομένως χρειάζεται συνέχεια στην μελέτη της απόδοσης διάφορων μεθόδων μηχανικής μάθησης αλλά και σε συνδυασμούς μοντέλων εκπαίδευσης. Κλείνοντας κρίνεται χρήσιμη η μελέτη των συνόλων εκπαίδευσης αλλά και συγκεκριμένων χαρακτηριστικών των αλγορίθμων, στην προσπάθεια να μειωθεί η υπερεκπαίδευση εκεί που τα αποτελέσματα του training φαίνεται να κινούνται προς μία θετική κατεύθυνση βελτίωσης της απόκρισης, όπως για παράδειγμα στο κεφάλαιο 4.3.

Βιβλιογραφία

[1] D. Griffiths, “Introduction to elementary particles”, Wiley-VCH

[2] C. M. Bishop, “Pattern Recognition and Machine Learning”, Springer

[3] J.D. Kelleher, “Deep Learning”, MIT Press

[4] Κ. Κουσουρής, Σημειώσεις του Μαθήματος “Αναγνώριση Προτύπων και Νευρωνικά Δίκτυα”

[5] https://pdg.lbl.gov/2023/reviews/contents_sports.html

[6] Γ. Τσιπολίτης, Σημειώσεις του Μαθήματος “Τεχνολογία Ανιχνευτικών και Επιταχυντικών Διατάξεων”

[7] Γ. Τσιπολίτης, Σημειώσεις του Μαθήματος “Στοιχειώδη Σωματίδια Ι”

[8] TMVA 4 (Toolkit for Multivariate Data Analysis with ROOT) User’s Guide

[9] ROOT User’s Guide: 6 Release Cycle

[10] CMS Collaboration, Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS, arXiv:1107.4277v1 [physics.ins-det]

[11] https://home.cern/science/experiments

[12] “Χαρακτηρισμός αδρονικών πιδάκων από συγκρούσεις πρωτονίων στο πείραμα CMS του LHC”, Ζαχαροπούλου ΄Αννα, Μεταπτυχιακή Διπλωματική Εργασία, Εθνικό Μετσόβιο Πολυτεχνείο-ΕΚΕΦΕ “Δημόκριτος”, Αθήνα, Φεβρουάριος 2019

[13] “Hadronization of quarks and gluons into hadron jets at high energies”, V. G. Grishin, Soviet Physics Uspekhi, vol. 29 iss. 2 (1986)

[14] R.K Ellis, W.J. Stirling, B.R. Webber, “QCD and Collider Physics”, Cambridge University Press