Εργασία 2η

ΥΠΟΛΟΓΙΣΤΙΚΗ ΚΒΑΝΤΙΚΗ ΦΥΣΙΚΗ ΚΑΙ ΕΦΑΡΜΟΓΕΣ 15/07/2025

Ταμπαρέσκου Ιωάννα, 4453

Εισαγωγή

Στην εργασία αυτή μελετάμε τη χρήση τεχνικών μηχανικής μάθησης για τον διαχωρισμό σήματος και υποβάθρου σε ένα εξωτικό σενάριο Higgs που βασίζεται σε υπερσυμμετρικές θεωρίες. Συγκεκριμένα, εξετάζουμε ένα βαρύ μποζόνιο Higgs το οποίο διασπάται σε δύο W μποζόνια και ένα ελαφρύτερο Higgs, με τις τελικές καταστάσεις να περιλαμβάνουν λεπτόνια και b-κουάρκ. Λόγω της πολυπλοκότητας αυτών των διασπάσεων και της ανάγκης για αποδοτική επιλογή γεγονότων, χρησιμοποιούμε αλγορίθμους μηχανικής μάθησης με στόχο τη βελτιστοποίηση της ανάλυσης.

Η υπερσυμμετρία (SUSY) είναι μια από τις πιο μελετημένες θεωρίες πέρα από το Καθιερωμένο Πρότυπο (SM), καθώς προσφέρει λύσεις σε αρκετά θεωρητικά προβλήματα, όπως το hierarchy problem και την ενοποίηση δυνάμεων [1]. Στο πλαίσιο των επεκτάσεων του Higgs sector, όπως το MSSM, προβλέπεται η ύπαρξη πέντε μποζονίων Higgs: δύο ουδέτερα CP-even (h, H), ένα ουδέτερο CP-odd (A) και δύο φορτισμένα (H^+, H^-) [2].

Η διάκριση σήματος από παρασκήνιο σε τέτοια σενάρια είναι εξαιρετικά απαιτητική, καθώς τα σήματα είναι σπάνια και συχνά καλύπτονται από μεγάλο υπόβαθρο που προέρχεται από κοινές διαδικασίες του SM. Παραδοσιακές μέθοδοι με βάση στατιστικά cuts δεν επαρκούν πλέον, και έτσι στρεφόμαστε σε πιο εξελιγμένες τεχνικές, όπως η μηχανική μάθηση (machine learning - ML), οι οποίες μπορούν να ανιχνεύσουν πολύπλοκες συσχετίσεις μεταξύ μεταβλητών [3].

Στην παρούσα εργασία χρησιμοποιούμε και συγκρίνουμε διαφορετικούς ταξινομητές, όπως Random Forest και XGBoost, και στη συνέχεια εκπαιδεύουμε ένα νευρωνικό δίκτυο σε περιβάλλον TensorFlow ή PyTorch. Η ανάλυση βασίζεται σε δεδομένα που περιλαμβάνουν τόσο μεταβλητές χαμηλού επιπέδου (π.χ. ορμές και ενέργειες σωματιδίων), όσο και υψηλού επιπέδου (π.χ. invariant masses, angular separations), και αξιολογείται η συνεισφορά κάθε ομάδας στη συνολική απόδοση του μοντέλου.

Η χρήση τεχνικών μηχανικής μάθησης στα πειράματα σωματιδιακής φυσικής είναι πλέον κοινή πρακτική, με εφαρμογές σε διάφορα στάδια της ανάλυσης – από την επιλογή γεγονότων μέχρι και την επανακατασκευή σωματιδίων [4, 5]. Σκοπός της παρούσας εργασίας είναι η κατανόηση και αξιολόγηση τέτοιων μεθόδων σε ένα συγκεκριμένο, εξωτικό φυσικό σενάριο.

Μεθοδολογία

Η εργασία βασίζεται στο σύνολο δεδομένων HIGGS, το οποίο περιλαμβάνει 28 αριθμητικά χαρακτηριστικά ανά δείγμα, με στόχο την ταξινόμηση γεγονότων ως σχετιζόμενων ή μη με το μποζόνιο Higgs. Τα πρώτα 21 χαρακτηριστικά είναι μετρήσεις χαμηλού επιπέδου και τα υπόλοιπα 7 παράγωγα χαρακτηριστικά υψηλού επιπέδου. Ο στόχος είναι η δυαδική ταξινόμηση (1: Higgs, 0: background).

Αρχικά πραγματοποιήθηκε προεπεξεργασία των δεδομένων, περιλαμβάνοντας τυποποίηση (standardization) και τυχαία ανάμιξη. Το σύνολο χωρίστηκε σε εκπαίδευση και έλεγχο με αναλογία 80:20 και διατήρηση ισορροπίας κλάσεων.

Για την ταξινόμηση χρησιμοποιήθηκαν τέσσερις αλγόριθμοι: XGBoost, Random Forest, SVM και ένα νευρωνικό δίκτυο με τρία κρυφά στρώματα και dropout. Κάθε μοντέλο εκπαιδεύτηκε και αξιολογήθηκε σε τρεις περιπτώσεις: πλήρες σύνολο χαρακτηριστικών, μόνο χαμηλού επιπέδου και μόνο υψηλού επιπέδου.

Η απόδοση των μοντέλων μετρήθηκε μέσω accuracy, precision, recall, F1-score και AUC. Τα αποτελέσματα παρουσιάζονται συγκριτικά, με στόχο την αξιολόγηση τόσο της αποτελεσματικότητας των μοντέλων όσο και της σημασίας των διαφορετικών τύπων χαρακτηριστικών.

Αποτελέσματα

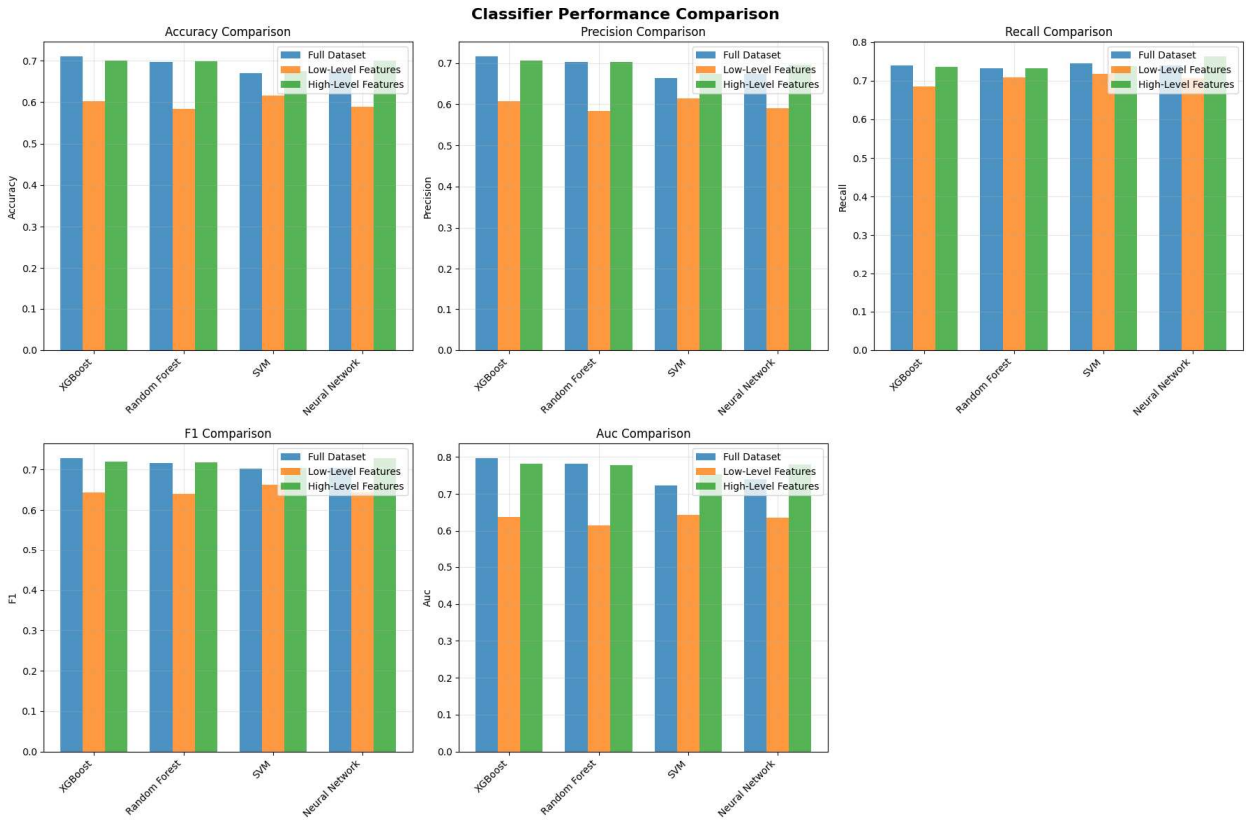
Παρακάτω παρατίθενται τα αποτελέσματα που προκύπτουν από τον κώδικα.

```
RESULTS COMPARISON
_____________________________________________________________

1. FULL DATASET RESULTS:
                accuracy  precision  recall      f1     auc
XGBoost           0.7100     0.7159  0.7399  0.7277  0.7961
Random Forest     0.6975     0.7030  0.7315  0.7170  0.7818
SVM               0.6694     0.6645  0.7446  0.7023  0.7220
Neural Network    0.6756     0.6732  0.7399  0.7049  0.7408

2. LOW-LEVEL FEATURES RESULTS:
                accuracy  precision  recall      f1     auc
XGBoost           0.6031     0.6074  0.6850  0.6439  0.6378
Random Forest     0.5831     0.5841  0.7088  0.6404  0.6148
SVM               0.6162     0.6145  0.7172  0.6619  0.6432
Neural Network    0.5894     0.5900  0.7076  0.6435  0.6366

3. HIGH-LEVEL FEATURES RESULTS:
                accuracy  precision  recall      f1     auc
XGBoost           0.7006     0.7056  0.7351  0.7200  0.7818
Random Forest     0.6981     0.7033  0.7327  0.7177  0.7778
SVM               0.6750     0.6732  0.7375  0.7039  0.7537
Neural Network    0.7012     0.6961  0.7625  0.7278  0.7798
```




```
SUMMARY ANALYSIS
_____________________________________________________________________

1. BEST PERFORMING CLASSIFIER:
   Full Dataset: XGBoost (AUC: 0.7961)
   Low-Level Features: SVM (AUC: 0.6432)
   High-Level Features: XGBoost (AUC: 0.7818)

2. FEATURE COMPARISON:
   Low-Level Features (1-21): Raw detector measurements
   High-Level Features (22-28): Derived physics quantities

   Average AUC - Low-Level: 0.6331
   Average AUC - High-Level: 0.7733
   → High-level features perform better on average

3. NEURAL NETWORK PERFORMANCE:
   Full Dataset: 0.7408
   Low-Level: 0.6366
   High-Level: 0.7798
```

Σχολιασμός Αποτελεσμάτων

Η σύγκριση των μοντέλων ταξινόμησης έδειξε ότι η καλύτερη συνολική απόδοση επιτεύχθηκε με το μοντέλο XGBoost, το οποίο πέτυχε AUC = 0.7961 όταν εκπαιδεύτηκε στο πλήρες σύνολο χαρακτηριστικών. Συγκριτικά, το Random Forest και το Neural Network παρουσίασαν ελαφρώς

χαμηλότερες επιδόσεις, με το τελευταίο να σημειώνει AUC = 0.7408. Αντιθέτως, όταν χρησιμοποιήθηκαν μόνο τα χαρακτηριστικά χαμηλού επιπέδου (Low-Level), παρατηρήθηκε αισθητή μείωση της ακρίβειας και της διαχωριστικής ικανότητας σε όλα τα μοντέλα, με καλύτερο το SVM (AUC = 0.6432), ενώ το Neural Network εμφάνισε τη χαμηλότερη απόδοση (AUC = 0.6366), γεγονός που υποδεικνύει ότι τα πρωτογενή δεδομένα του ανιχνευτή δεν είναι επαρκώς εκφραστικά από μόνα τους.

Αντιθέτως, η χρήση μόνο των χαρακτηριστικών υψηλού επιπέδου (High-Level) – τα οποία περιλαμβάνουν παραγόμενες φυσικές ποσότητες – οδήγησε σε αισθητά καλύτερα αποτελέσματα, με μέσο AUC = 0.7733. Ενδιαφέρον παρουσιάζει η συμπεριφορά του νευρωνικού δικτύου, το οποίο με τα high-level χαρακτηριστικά σημείωσε AUC = 0.7798, καλύτερη από αυτήν με το πλήρες σύνολο, φανερώνοντας ότι επωφελείται ιδιαίτερα από πληροφορία ενσωματωμένη σε φυσικές παραμέτρους. Συνολικά, τα high-level χαρακτηριστικά αποδεικνύονται πιο χρήσιμα από τα low-level, ενώ η ενσωμάτωσή τους σε πιο σύνθετα μοντέλα, όπως το XGBoost ή τα νευρωνικά δίκτυα, βελτιώνει σημαντικά την απόδοση. Τα ευρήματα αυτά αναδεικνύουν τη σημασία της επιλογής χαρακτηριστικών στη βελτιστοποίηση της ταξινόμησης, ειδικά σε δεδομένα φυσικής υψηλής ενέργειας.

Συμπεράσματα

Η ανάλυση των μοντέλων ταξινόμησης έδειξε ότι η καλύτερη απόδοση επιτυγχάνεται με το XGBoost, το οποίο πέτυχε AUC 0.7961 όταν εκπαιδεύτηκε με το πλήρες σύνολο χαρακτηριστικών. Τα high-level χαρακτηριστικά, που αποτελούνται από παραγόμενες φυσικές ποσότητες, αποδείχθηκαν πολύ πιο εκφραστικά και χρήσιμα σε σχέση με τα low-level πρωτογενή δεδομένα, καθώς η χρήση τους βελτίωσε σημαντικά την απόδοση όλων των μοντέλων (μέσο AUC ~0.77 έναντι ~0.63 για τα low-level). Ιδιαίτερα τα νευρωνικά δίκτυα ωφελήθηκαν από τα high-level χαρακτηριστικά, σημειώνοντας καλύτερη απόδοση με αυτά παρά με το πλήρες σύνολο, γεγονός που υποδηλώνει την ικανότητά τους να αξιοποιούν σύνθετες μη γραμμικές σχέσεις. Συνολικά, τα ευρήματα υπογραμμίζουν τη σημασία της επιλογής κατάλληλων χαρακτηριστικών και την προτίμηση σύγχρονων αλγορίθμων όπως το XGBoost για τη βελτιστοποίηση της ταξινόμησης σε δεδομένα φυσικής υψηλής ενέργειας.

Βιβλιογραφία

[1] H. Baer, X. Tata, *Weak Scale Supersymmetry: From Superfields to Scattering Events*, Cambridge University Press, 2006.

[2] M. Carena and H. E. Haber, *Higgs boson theory and phenomenology*, Prog. Part. Nucl. Phys. 50 (2003) 63, [hep-ph/0208209].

[3] P. Baldi, P. Sadowski, and D. Whiteson, *Searching for exotic particles in high-energy physics with deep learning*, Nature Communications 5, Article number: 4308 (2014).

[4] A. Radovic et al., *Machine learning at the energy and intensity frontiers of particle physics*, Nature 560, 41–48 (2018).

[5] J. Duarte et al., *Fast inference of deep neural networks in FPGAs for particle physics*, JINST 13 (2018) P07027.