Aristotle University of Thessaloniki
Faculty of Sciences
Department of Physics
MSc Computational Physics

Computational Quantum Mechanics

Application of machine learning methods for signal and background separation in an exotic Higgs scenario

by Eleftheria Pigi Miliou
07/2025

1 Introduction

This study investigates machine learning methods to optimize signal-background separation in a supersymmetric Higgs boson search at the LHC. In the SUSY scenario, the Higgs sector contains five bosons (unlike the single Higgs in the Standard Model). We focus on a heavy Higgs boson decaying via $H^0 \rightarrow W^+W^-h$, with subsequent decays to leptons $W^\pm \rightarrow l^\pm + v$ and b-quarks ($h \rightarrow bb$). We evaluate three classification methods, K-Nearest Neighbors (KNN), Random Forest, and Artificial Neural Networks (ANN) using both low-level detector measurements and high-level derived variables to optimize signal identification. The goal is to assess their performance in distinguishing between signal (1) and background (0) classes.

2 Data Preprocessing

All script files begin by reading the dataset and verifying the data types. During this step, an element initially classified as an object was identified and corrected by replacing it with zero, as indicated by the string value. The cleaned data was then separated into three distinct categories based on the given instructions:

- Classification label: The first column (1 for signal, 0 for background).
- Low-level features (Columns 121): Raw detector-level quantities
- High-level features (Columns 2228): Derived physics variables

This separation allowed for targeted analysis of how different feature sets influence model performance.

3 KNN Classification

We first applied the KNN classifier, starting with the low-level features and repeating the process for high-level features. After splitting the data into training and test sets (75/25 ratio) and performing feature scaling, we evaluated the model with varying numbers of neighbors (k). The confusion matrices and accuracies were recorded as follows:

Table 1: KNN Classification Accuracy

Number of neighbors (k)	Low-Level Accuracy	High-Level Accuracy
1	0.537	0.615
3	0.554	0.657
5	0.547	0.659
10	0.566	0.662
20	0.575	0.674
50	0.595	0.660

For low-level features, accuracy improved as k increased, peaking at 0.595 for k = 50. High-level features consistently outperformed low-level ones, with the best accuracy (0.674) at k = 20. This suggests that high-level features contain more discriminative information for this task. Random Forest Classification

4 Random Forest Classification

Next, we implemented the Random Forest classifier, varying the number of estimators (n). The results were:

Table 2: Random Forest Classification Accuracy

Number of estimators (n)	Low-Level Accuracy	High-Level Accuracy
10	0.567	0.670
50	0.601	0.678
100	0.604	0.680

Low-level features achieved moderate accuracy (0.604 at n = 100), while high-level features again showed stronger performance, reaching 0.680 at n = 100. The results indicate that ensemble methods like Random Forest can marginally outperform KNN, particularly with high-level features.

5 Artificial Neural Network (ANN) Classification

The ANN was structured with an input layer, one hidden layer, and an output layer, trained using the Adam optimizer. We tested different batch sizes and epochs, recording the following accuracies:

Table 3: ANN Performance Across Feature Types

Parameters	Low-Level Feature				High-Level Feature			
	Trial 1	Trial 2	Trial 3	Trial 4	Trial 1	Trial 2	Trial 3	Trial 4
Batch	8	16	24	32	8	16	24	32
Epochs	200	150	150	200	200	150	150	200
Accuracy	0.558	0.588	0.576	0.588	0.702	0.688	0.687	0.701

Table 4: ANN Performance Comparison

Feature Type	Best Configuration		Accuracy	
	Batch	Epochs	Value	Δ vs Low-Level
Low-Level	16	150	0.588	-
High-Level	8	200	0.702	+0.114

Low-level features achieved modest accuracy (best: 0.588), whereas high-level features consistently exceeded 0.68, peaking at 0.702. The ANN's performance aligns with the trend observed in other models: high-level features yield better classification, likely due to their engineered or derived nature.

6 Discussion

The experiments reveal three consistent findings across all models. First, high-level features systematically outperformed low-level features by 811% accuracy in all classifiers, confirming their superior discriminative power for this physics task. This aligns with expectations, as high-level variables encode derived physical quantities that directly correlate with signal signatures. Second, model comparisons show the ANN achieved peak performance (70.2% accuracy with high-level features), demonstrating its capacity to learn complex feature interactions. The Random Forest proved more robust than KNN (67.8% vs. 66.0% accuracy for high-level features), though both were surpassed by the ANN. Third, hyperparameter analysis revealed critical thresholds: KNN accuracy plateaued beyond k = 20 neighbors, Random Forest gains diminished after n = 50 estimators, and ANN performance peaked with smaller batches (816), suggesting larger batches may hinder gradient updates for this dataset. These trends emphasize that while high-level features dominate performance, optimal hyperparameters vary significantly by algorithm.

7 Conclusions

Our analysis demonstrates that high-level features consistently outperform raw detector measurements across all classifiers, with ANNs achieving the highest discrimination power (70.2% accuracy). This confirms that physically-derived variables better capture signal characteristics in this SUSY Higgs search. Future work could explore feature importance analysis and hybrid architectures combining both feature types.