Classification of Signal and Background in Search for Di-Higgs Boson at Large Hadron Collider

Rano Marufova

Advisor: Suyog Shrestha

Introduction

In this research, we leveraged Machine Learning (ML) techniques, specifically classification algorithms, to address a fundamental challenge in particle physics: separating signal and background events. Classification is a supervised learning algorithm that assigns events to predefined categories based on given input variables (features). Our focus was on classifying simulated Di-Higgs boson (HH) process as the signal and non-Di-Higgs boson (non-HH) processes as the background, wherein each input variable represented a detector measurement. We augmented this dataset by engineering new variables informed by our understanding of the Higgs boson and its predicted properties.

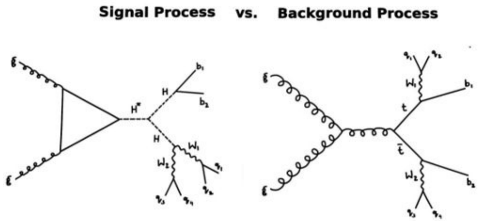

Figure 1

We employed three different classification algorithms available in the Toolkit for Multi-Variate Analysis (TMVA) developed by CERN: Boosted Decision Trees (BDT), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). We systematically optimized these algorithms by varying the input features and the algorithm's hyperparameters, which resulted in the maximum values of signal significance for each algorithm.

Data Sample and Input Features

Variables are obtained from simulated Di-Higgs (signal) and top-quark (background) processes, where each variable represent a corresponding measurement in the detector or is a function of the said measurements. The measurements typically correspond to the momentum of the particle (pT) and the angular coordinates of the detector (eta, phi).
From the detector measurements, we have 93 variables, 13 of which were composite variables and we engineered 16 new variables using our knowledge of particle physics.
Figure 2 shows the variables that have have distinguishing features between signal and background.

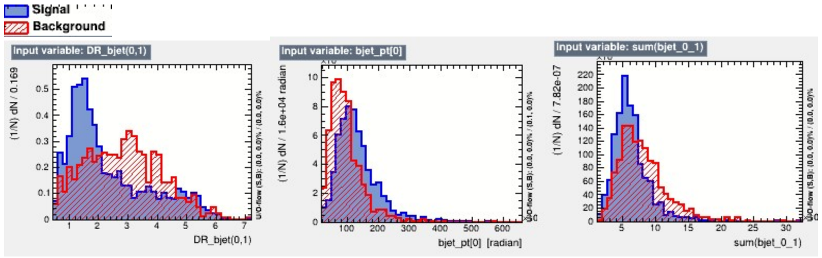

Figure 2

Optimization

Background refers to any observed event or measurement that is not directly related to the phenomenon or particle of interest but arises from other sources or processes. **Signal**, on the other hand, refers to the specific event or measurement that represents the presence or characteristics of the particle or phenomenon under study. It is the desired outcome of the experiment and carries the information researchers are interested in. In this research our event of interest is Di-Higgs.
Our simulated data has the measurements based on the Feynman diagrams shown in Figure 1. Simulated measurements are used to train a machine learning model for classification.

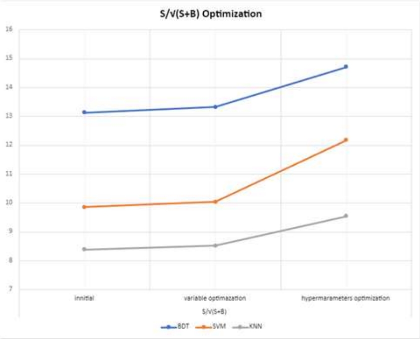

Figure 3

Variable optimization involves selecting the most relevant input variables, often through feature selection or dimensionality reduction.

Hyperparameter optimization aims to find the best settings for model parameters, such as number of trees or maximum depth, to maximize performance.
We were able to reach higher separation efficiency by optimizing our algorithms.

Final Results

After trying out different methods of optimization we were able to achieve the maximum S/√(S+B) of 14.7141 by optimizing the BDT algorithm. SVM and KNN algorithms had S/√(S+B) of 12.1838 and 9.5439 respectively.

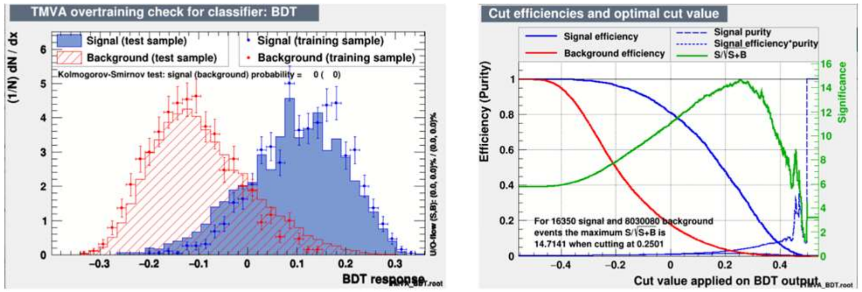

Figure 4

Conclusion and Outlook

- We did not manage to test algorithms such as DNN and MLP, possibly due to incompatibility with our operating system, highlighting the need for further investigation to benchmark their performance.

- We found that KNN proved to be less efficient in our classification problem, leading us to consider its exclusion from future research endeavors.

- We intend to transition our code from C++ to Python to benefit from enhanced implementation capabilities.

- We need to work further on optimization of our algorithms to achieve maximum significance (S/√(S+B)

☆ **We want to express gratitude to the Tolls Fellows Program for providing financial support for this project.**