Discrimination between the final state of $t\bar{t}H$ and $t\bar{t}b\bar{b}$ using neural network

R P Putri¹, S Sumowidagdo², R Primulando³ and T Mart¹

¹Department of Physics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia
²Research Center for Physics, Indonesian Institute of Sciences (LIPI), Puspiptek Area, Serpong 15134, Indonesia
³Department of Physics, Faculty of Information Technology and Sciences, Catholic University of Parahyangan, Bandung 40141, Indonesia

Corresponding author’s email: suharyo.sumowidagdo@lipi.go.id

Abstract. After the discovery of the Higgs boson in 2012 at the Large Hadron Collider (LHC), the effort to understand the detailed properties of the Higgs bosons started. Of particular importance is study of the Higgs coupling to the top quark. This coupling can be studied through the associated production of Higgs boson with top-antitop quark pair, $t\bar{t}H$. This process however suffers from the indistinguishable background $t\bar{t}b\bar{b}$, since the Higgs boson decays predominantly into bottom anti-bottom quark pair, $b\bar{b}$. This study presents systematic approach of using machine learning (ML), specifically neural network method to distinguish between the process $t\bar{t}H$ and $t\bar{t}b\bar{b}$. Using input variables of kinematic variables (momentum), we found a signal efficiency of 46.7 % for signal events that have passed the preselection criteria. We conclude that the currently used input variables are not sufficient to discriminate between signal and background events, and we suggest that inclusion of input variables calculated from the fully reconstructed event could provide stronger discrimination between signal and background.

Keywords: Higgs, LHC, neural network

1. Introduction

On the 4th of July 2012, the ATLAS and CMS collaboration announced the discovery of Higgs Boson with a mass of 125 GeV predicted by the electroweak symmetric breaking mechanism of the Standard Model (SM) Particle Physics. Since then, studying the properties of Higgs Boson has motivated many physicists in the LHC. So far, the measurable nature of the Higgs Boson has shown remarkable consistency to that predicted by SM. But until now SM has not been able to explain the entire observation of physical phenomena.

One such example is that it is still difficult to explain that coupling of Higgs Boson with fermion is either a scalar, pseudoscalar, or a pseudoscalar-scalar mixture. This work uses coupling of Higgs with the heaviest fermion, top quark. This is due to the Yukawa Higgs coupling which is proportional to the fermion mass resulting in the greatest effect of Higgs interaction with the top quark. The background reaction contaminates the search for $t\bar{t}H$. From the decay, $t\bar{t}b\bar{b}$ is the most dominant particle
background. Particle resulting in the final state is complex, making it difficult to distinguish Higgs particles.

Recently, Machine Learning (ML) has become a developed method in many sectors, especially in the classification of large data. ML was ever used by the ATLAS Collaboration [1] in the search for Higgs particles. The search still considers Higgs as a scalar particle, and is inclusive using all background reactions. Therefore, experimentalists need suggestion regarding the best method to distinguish Higgs particles from background.

In this work, we will discriminate Higgs particles from the background using ML Neural Network (NN). But this work only focus with the $b\bar{b}$ background which is the most dominant. The Feynman diagrams of both reactions are shown in figure 1. The novelty of this work is to use the semileptonic final state particles and one neutrino to minimize the degree of freedom lost in the detector. This work will be part of a larger project, which discriminates the scalar Higgs particle with Pseudoscalar Higgs particle.

2. Event generation, detector simulation and event selection

The initial process was generate events using Monte Carlo Generator (MC) [2] from the collision of $p\bar{p}$ with the energy of the center of mass is $\sqrt{s} = 13$ TeV. There are 8 particles which are the result of the final state with each particle having their respective properties of 4-momentum, mass, angle $\theta$ and $\phi$. The generation of the events has a file format of "lhe". The events will be as the input for the detector. On this process, we use delphes simulation detector. Delphes [3] is a detector response, which can only detect the particle jet, electrons, muons, W, and Z. The response form the framework that simulate the response of the collider experiment. Delphes simulates the response of a detector consisting of an electromagnetic calorimeter (ECAL) and a hadronic calorimeter (HCAL), so that delphes can detect and distinguish final state particles whether they are jet or lepton. Event from Delphes simulation process then formed a file with format "lhco". Formation of Event and detector simulation generates 50000 events.

On Event Selection of this single lepton channel, the transverse momentum used must be greater than the threshold momentum, which is $P_T > 25$ GeV for jets and electrons. We use a maximum value of $\eta$ is 2.0 for electrons and 2.5 for jets. When there are extra electrons or muons particles, the extra particle needs to be identified whether it comes from decay W, t, $t\bar{t}$, or H. So when there is an electron or second muon that has $P_T < 10, \eta > 2$ event will be used for the next process.

Figure 1. Feynman diagrams for $t\bar{t}H$ signal process (left) and $t\bar{t}b\bar{b}$ background process (right).
3. Neural network construction

Machine Learning is Artificial Intelligent which receive input data and using statistical analyst to predict output. On this research, we used Neural Network (NN) which builds output signal that is transmitted to the other layer. NN consists of one layer input, one or more hidden layers, and one output layer. A weight is parameter which influences input to the next layer until producing the output. This equation describes output signal of NN.

\[
Output = f(n) = \frac{1}{1 + e^{-n}} \tag{1}
\]

\[
n = w \cdot x + b \tag{2}
\]

\(w\) is weight, \(x\) is input value or \(n\) from previous layer, and \(b\) is bias factor which only one for every layer.

Training and Testing was done as procedure of separating signals and backgrounds using the data. Basically the total data will be divided into data for training and testing. On this work, 2000 of 3119 total events that has been separated between signals and backgrounds were used for training as the first discrimination process to obtain the weight parameters of each variable. Then the remaining data, 1119 between the signal and the background were mixed and used in the testing process using the weight parameters of the training results to see the results of discrimination and efficiency when the data between the signal and the background has been mixed.

Based on the network architecture generated in figure 2, there are 36 variables used as input nodes with 1 bias node, then there is a hidden layer which is a procedure between input variables to output. On this work, we used 1 layer with nodes \(N + 5 = 41\) nodes to produce the greatest efficiency with an additional bias node in the hidden layer to normalize the output results. Then from the hidden layer, we use the “Sigmoid” neuron activation function type which produces only 1 output value between 0 and 1. Where the result that closer to 1 indicates the output is the signal and the result that closer to 0 indicates output is background. Then from the network architecture we get the number of weights or the number of connections from input to output as much as 1559 \(((36\text{ input} + 1\text{ bias node}) \times 41\text{ first hidden layer}) + ((41 + 1\text{ bias node}) \times 1))\). Where the weight value of each relation varies to indicates the gradient as the distinguishing line between the signal and the background.

![Image](image.png)

**Figure 2.** Neural Network architecture using 36 inputs, 41 hidden layers and 1 output.
4. Results and analysis
In this work, the event result of the event generation by Monte Carlo simulation is 50000 event of signals and 50000 event of background. The event is selected based on the parameters explained on the event selection section so that the number of the event becomes 1345 for background and 1774 for signal, so it can be calculated the efficiency value of selection event is 3.5 % for the signal and 2.7 % for the background.

Based on the mapping of the input variable, we get from the result of training that the best three variables sorted from the strongest to discriminate between the signal and the background reaction are Hadronic over Electromagnetic (HadOvEm), Transverse Momentum, and Mass. HadOvEm physically means the W and top which decay hadronically to b-jet and q-jet per electromagnetic energy in the calorimeter cell. This means that the difference between scalar HadOvEm and background HadOvEm is a variable with the most distinguishable distribution of data when compared to other variables. If viewed in more detail, the distribution of HadOvEm values on the background is greater than HadOvEm value on the signal.

There are transverse momentum and mass which becomes the next best discrimination variable. On the signal, there is $t\bar{t}$ which decays into Higgs then becomes $bb$, while on the background there is $t\bar{t}$ which decays into gluon which then becomes $bb$ so that there is a significant difference between a total of two b-jets from all four b-jets which is equal to the Higgs mass of the signal reaction and also the Gluon mass on the background reaction.

Figure 3 below describes an increasingly valuable signal area 1 and an increasingly valuable background reaction area 0, with the points as output of training and the bar as output of testing. Efficiency value is 46.7 % for testing and 73.5 % for training. Based on the efficiency value of training and testing, the value of training efficiency is greater than the testing efficiency which means that the area between the signal and the background reaction on the training is greater discriminated than the area in testing. Then figure 4 illustrates the relationship between the number of areas to be studied for the respective efficiency values on the signal as well as the background reactions that are connected with figure 3. For example, when cutting the left area of 0.1, it is found that the value of the reaction efficiency of the background is reduced more than the signal efficiency, so on figure 4 background is reduced more than the signal for the amount of data 1000 signals and 1000 background reactions. In addition, from figure 4 there is also a significance value of 22.36 when there is no cut area. Significance is described as a signal per square of the total event (signal and background).

![Figure 3](image-url)
5. Conclusion
The efficiency value resulting from event selection is quite small, i.e. only 3.5% on signal and 2.7% on the background which means less data will be processed in Machine Learning discrimination. Then the efficiency of the testing process on NN is also only 46.7%, which means that of the same amount of signals and background, the total of identified signals is only 46.7%. These results indicate that the Machine Learning Method (ML), especially Neural network used has not good enough yet to discriminate between the signal $t\bar{t}H$ and the background $t\bar{t}b\bar{b}$ because there are still many missing parameters for the purpose of a more complete analysis so that we need to generate more events and optimization of a more comprehensive top quark reconstruction to make the resulting variables significantly discriminate more between the signal and the background reaction.

References
[1] ATLAS Collaboration 2018 Observation of Higgs Boson Production in Association with A Top Quark Pair at the LHC with the ATLAS Detector arXiv:1806.00425v1 [hep-ex]
[2] Seymour M H and Marx M 2013 Monte Carlo Event Generators arXiv:1304.6677v1 [hep-ph]
[3] Favereau J d et al. 2014 DELPHES 3: A Modular Framework for Fast Simulation of A Generic Collider Experiment arXiv:1307.6346v3 [hep-ex]
Appendix

Input Variable of Final State Particle
