# Study of Di-Higgs Boson Sensitivity in the Mode with One Light Lepton and Three Hadronically Decaying Tau Leptons – Summer Student Report

José Gabriel García

*CERN Summer Student 2024*

Supervisor: André Sopczak

*IEAP CTU in Prague*

**Abstract**

Machine Learning techniques have proven to have the potential to improve the separation of signal and background events in High Energy Physics (HEP) data. This study focuses on the optimization of the $HH$ production measurement in the analysis channel with one light lepton and three hadronically decaying tau leptons. The analysis is performed with simulated data and the results are expressed as expected sensitivity.

# Contents



# 1 Introduction

## 1.1 Process of Interest

Since the discovery of the Higgs boson in 2012, significant progress has been made in studying its interactions with other particles, however, its self-coupling has not been measured yet. This measurement remains one of the most important challenges in modern particle physics. One particularly intriguing process that probes its self-coupling is the production of Higgs boson pairs $HH$ (Figure 1), which provides insights into the structure of the Higgs potential. Among the various decay modes, the final state with one lepton and three hadronically decaying tau leptons ($1l3\tau$) is of great interest. It could contribute additional sensitivity. This analysis is a case study focusing on this channel.

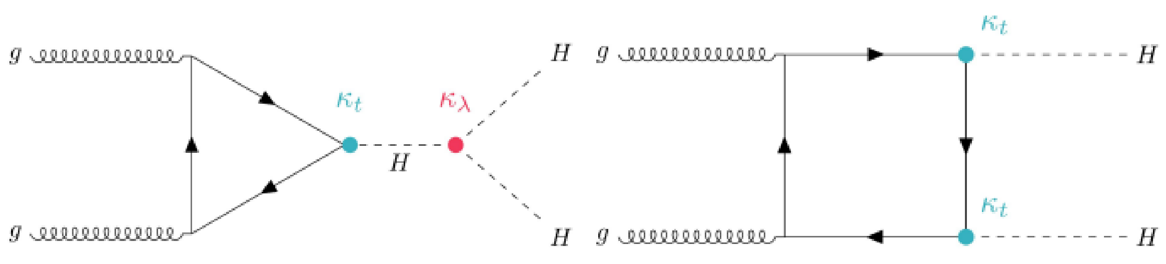


Figure 1: Feynman diagrams for $HH$ production

### 1.2 Machine Learning Methods

Owing to the complex decay modes, it is hard to separate $HH$ signal events from background events that have similar signatures. After applying selection cuts on measured variables, the signal to background ratio is quite small in the $1l3\tau$ channel. Machine Learning (ML) techniques are applied to obtain a classifier that assigns each event a probability of being $HH$ signal. This study addresses a new channel as a CERN summer student project. Further details on this project are given in Ref. [1].

## 2 Preselection and Expected Numbers of Events

Simulated datasets are used, as given in Table 1. Preselection criteria are applied to separate $HH$ events in the $1l3\tau$ signal region. The object definition are given in Tables 2, 3, and 4.

The numbers of events that pass the preselection are listed in Table 5. The classes used for this analysis are the signal class $HH$ and the common background classes $Z + jets$, *tops*, diboson ($VV$) and *other*.

## 3 Separation of Signal and Background

Figure 2 shows some of the features used for the signal and background separation.

## 4 Significance

Significance is a common metric used in particle physics to quantify the certainty of the presence of a signal. The significance $S$ can be quite complex to be calculated, however, there is a widely used approximation

$$S = \frac{s}{\sqrt{s+b}}, \tag{1}$$

where $s$ is the number of expected signal events and $b$ is the number of expected background events. An alternative approximation is

$$S = \frac{s}{\sqrt{b}}. \tag{2}$$

Higher accuracy is obtained using the statistical method implemented in TRExFitter [2].

| Process | Files |
| --- | --- |
| SM HH | 600854, 600856, 600858 |
| Zjets | 364100, 364101, 364103, 364104, 364105, 364106, 364107, 364108, 364109, 364110, 364111, 364112, 364113, 364114, 364115, 364116, 364117, 364118, 364119, 364120, 364121, 364122, 364123, 364124, 364125, 364126, 364127, 364128, 364129, 364130, 364131, 364132, 364133, 364134, 364135, 364136, 364137, 364138, 364139, 364140, 364141 |
| DYMLowZ | 364198, 364199, 364200, 364201, 364202, 364203, 364204, 364205, 364206, 364207, 364208, 364209, 364210, 364211, 364212, 364213, 364214, 364215 |
| Wjets | 364156, 364157, 364158, 364159, 364160, 364161, 364162, 364163, 364164, 364165, 364166, 364167, 364168, 364169, 364170, 364171, 364172, 364173, 364174, 364175, 364176, 364177, 364178, 364179, 364180, 364181, 364182, 364183, 364184, 364185, 364186, 364187, 364188, 364189, 364190, 364191, 364192, 364193, 364194, 364195, 364196, 364197 |
| $t\bar{t}$ | 410470 |
| tops | 410644, 410645, 410646, 410647, 410658, 410659, 304014, 410080, 410081 |
| $V\gamma$ | 700011, 700012, 700013, 700014, 700015, 700016, 700017 |
| ttW | 700168 |
| ttZ | 410156, 410157, 410218, 410219, 410276, 410277, 410278 |
| ttH | 346343, 346344, 346345 |
| 3V | 364242, 364243, 364244, 364245, 364246, 364247, 364248, 364249 |
| VH | 342284, 342285 |
| VV | 363355, 363356, 363357, 363358, 363359, 363360, 363489, 364250, 364253, 364254, 364255, 364283, 364284, 364287 |
| tV | 410560 |
| ptVV | 410408 |
| ZZ | 363355, 363356 |
| WZ | 363357, 363358 |
| lnuqqWZ | 363489 |
| WW | 363359, 363360 |
| jjVV | 364283, 364284, 364287 |
| flVV | 364250, 364253, 364254, 364255 |
| Rare $tt$ | 410397, 410398, 410399 |

Table 1: Signal and background processes and the Data Set ID (DSID)

| **Feature** | **Criterion** |
|---|---|
| Identification | JetID RNN Medium / Loose |
| $p_T$ [GeV] | $> 20$ |
| $\|\eta\|$ | $< 2.5$ |
| Crack region $1.37 < \|\eta\| < 1.52$ | vetoed |
| $N_{\text{track}}$ | 1 or 3 |
| Charge | $\pm 1$ |
| Electron veto | passEleBDT |
| Muon overlap removal | passMuonOLR |

Table 2: $\tau_{\text{had}}$ definition

| **Feature** | **Criterion** |
|---|---|
| Identification | LooseLH / TightLH |
| Isolation | PLV Loose / PLV Tight |
| $\|\eta\|$ cut | $(\|\eta\| < 1.37) \,\|\|\, (1.52 < \|\eta\| < 2.5)$ |
| $d_0$ significance cut | 5 |
| $z_0$ cut | 0.5 mm |

Table 3: Electron definition

| **Feature** | **Criterion** |
|---|---|
| Identification | Loose / Medium |
| Isolation | PLV Loose / PLV Tight |
| $\|\eta\|$ cut | $< 2.5$ |
| $\|d_0\|/\sigma_{d_0}$ | 3 |
| $z_0$ cut | 0.5 mm |

Table 4: Muon definition

| | Before Preselection | After Preselection |
|---|---|---|
| 3V | 11788 | 673 |
| DYMLowZ | 848 | 0 |
| flVV | 343644 | 2528 |
| jjVV | 28392 | 155 |
| lnuqqWZ | 2816 | 6 |
| ptVV | 2379 | 42 |
| Rare tt | 1131 | 0 |
| tops | 11662 | 17 |
| ttbar | 82231 | 63 |
| ttH | 123369 | 3122 |
| ttW | 19955 | 41 |
| ttZ | 6973 | 15 |
| tV | 2234 | 7 |
| Vgamma | 9732 | 10 |
| VH | 269 | 2 |
| VV | 391376 | 2707 |
| Wjets | 43718 | 56 |
| WW | 5387 | 6 |
| WZ | 7206 | 6 |
| Zjets | 58907 | 61 |
| ZZ | 3931 | 6 |
| HH signal | 115399 | 7421 |

Table 5: Signal and background events before and after preselection

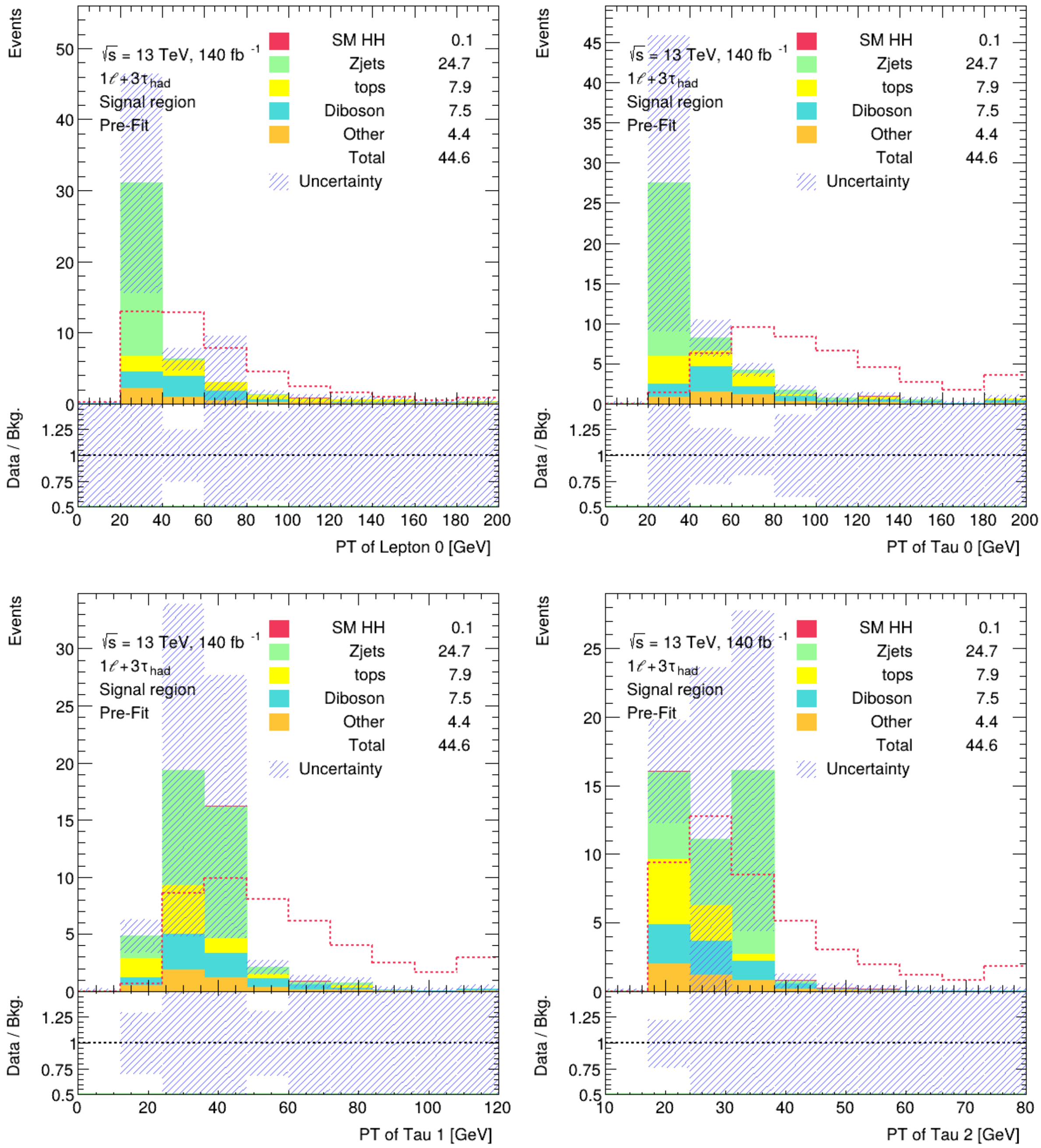


Figure 2: Leading lepton, leading tau, subleading tau and third tau transverse momentum for signal and background

# 5 Binary Network

## 5.1 Inputs and Functionality

The most common Machine Learning algorithm employed in this analysis is the Multi-Layer Perceptron (MLP). It consists of a number of layers where each layer contains some number of nodes. Each node has a connection to every node in the previous layer as shown in Figure 3. The first (left-most) layer is called the input layer and in this case takes in physical observables for each event. The final layer is called the output layer and it is used to obtain a probability that an event belongs to a signal class. Layers in between are called hidden layers.

Each node applies a linear transformation to the data passed to it, where each connection is assigned a weight $w_i$. To be able to model more complex situations, the node then applies a non-linear function called activation function so that the effect of the node can be described as

$$A\left(\sum_{i=1}^{N} w_i \cdot x_i\right), \tag{3}$$

where $A$ is the activation function and $x_i$ are the inputs from previous layers [3].

The training of the network is done with simulated events by updating the network weights in each epoch. This is done by minimizing a loss function which is calculated from a set of predictions and it is defined such that the higher the value of the loss function corresponds to a worse set of predictions. This function is then minimized in the training process, so it must be differentiable such that the effect of updating the network weights on the loss function can be calculated. A typical evolution of its value during a training process is shown in Figure 4. The weighted binary cross-entropy [5] is used as the loss function for all training processes. The reason that the validation loss is lower than the training loss is that the validation is always performed after each training step so that the network generally performs better. The importance of the input features has been studied and the 10 most important features are shown in Figure 5.

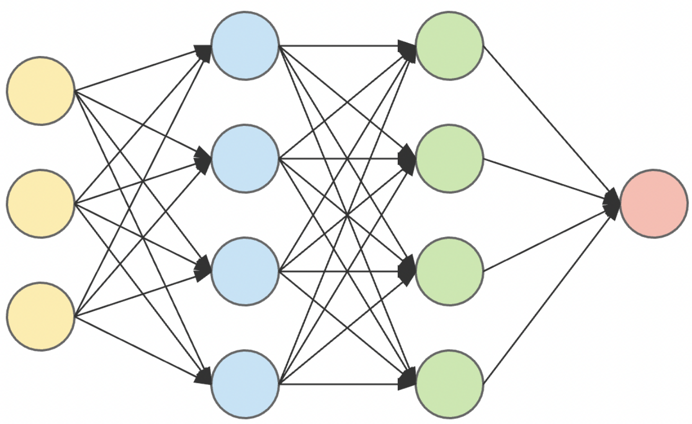

Figure 3: Schematic diagram showing the nodes, layers and connections in an MLP with four layers [4]

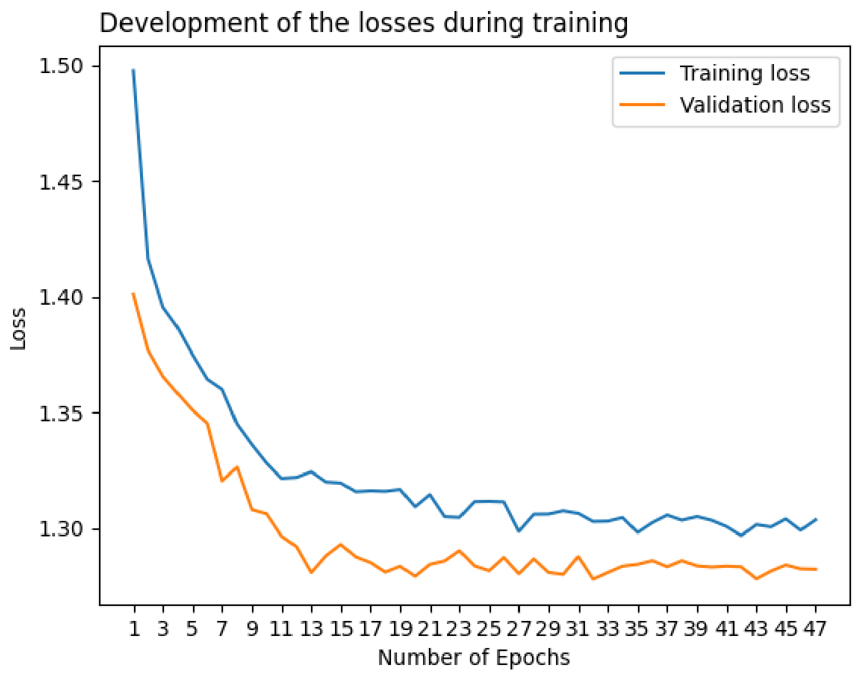


Figure 4: Typical behaviour of the value of a loss function during training and validation

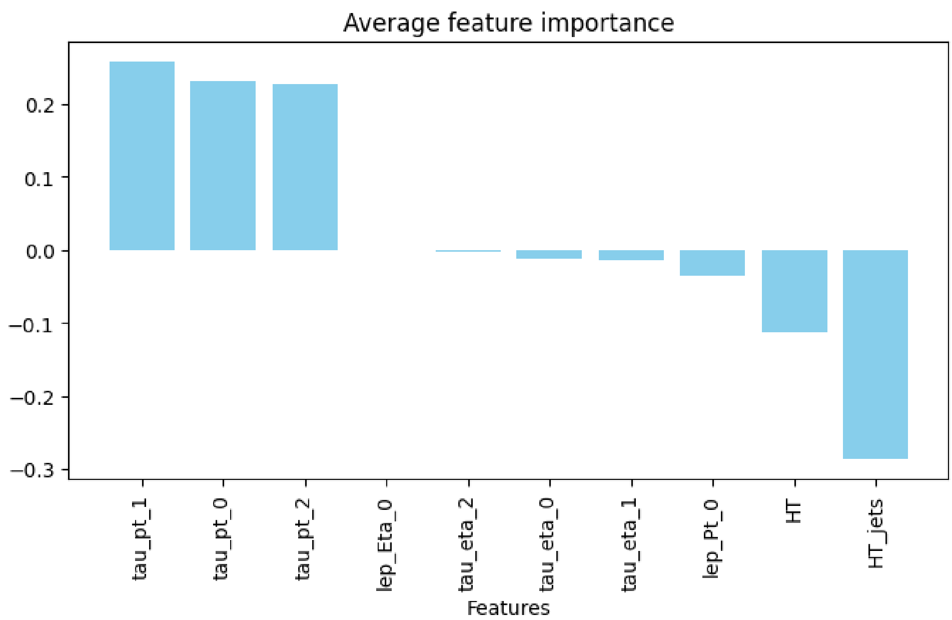


Figure 5: Ranking of the feature importance of the 10 most important features

## 5.2 Hyperparameter Optimization

Another important aspect of effectively using Machine Learning methods is to find the best parameters for the algorithms. In the case of the neural network discussed here the parameters that can be varied each time a new network is trained are architecture, i.e. the number of layers and nodes per layer, the learning rate, the rate of how much weights are updated during the training, the batch size, which describes how many events are processed before updating parameters and the maximum number of training steps allowed.

These parameters impact the performance of the neural network and an optimum has to be found. This is done by repeating the training process many times with different sets of parameters and monitoring the performance. The performance for each of the parameter sets is recorded.

Figure 6 shows the selected values for the dropouts of each layer in the hyperparameter optimization. A value of 0.2 dropout results in a better accuracy and a significant reduction in loss. Dropouts are a regularization technique that involves randomly deactivating a fraction of neurons during training. This prevents overfitting by forcing the network not to rely excessively on specific features. This improves the model's generalization by exposing it to more diverse combinations of connections. Figure 6 also shows dropout values that are too high, thus reducing the performance due to excessive information loss.

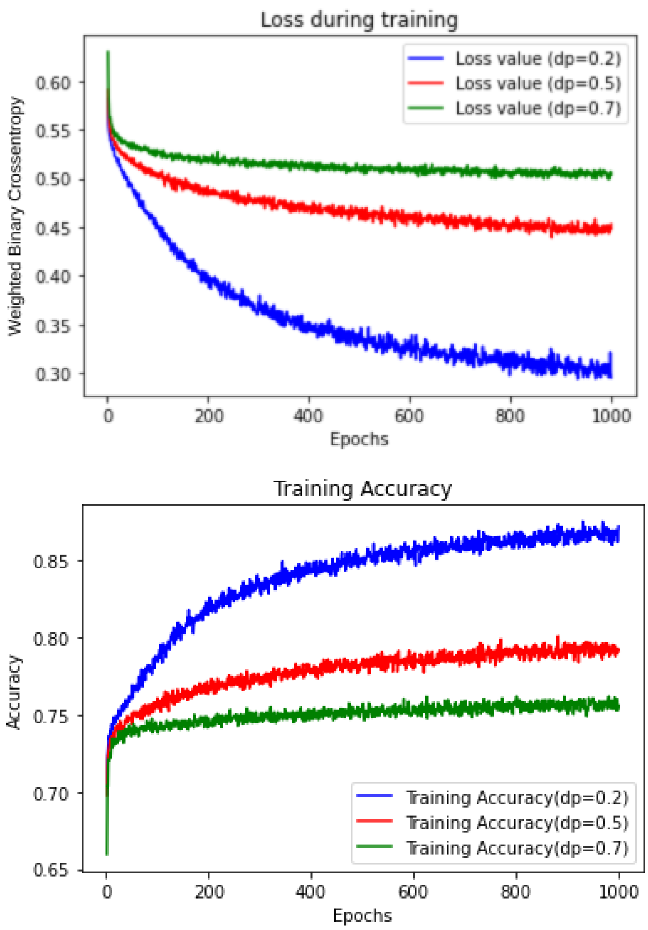


Figure 6: Dropout optimization for the Neural Network

# 6 Network Output

Figure 7 shows the network output assuming SM predictions. The red dotted line shows the background distribution normalized to the sum of the background events.

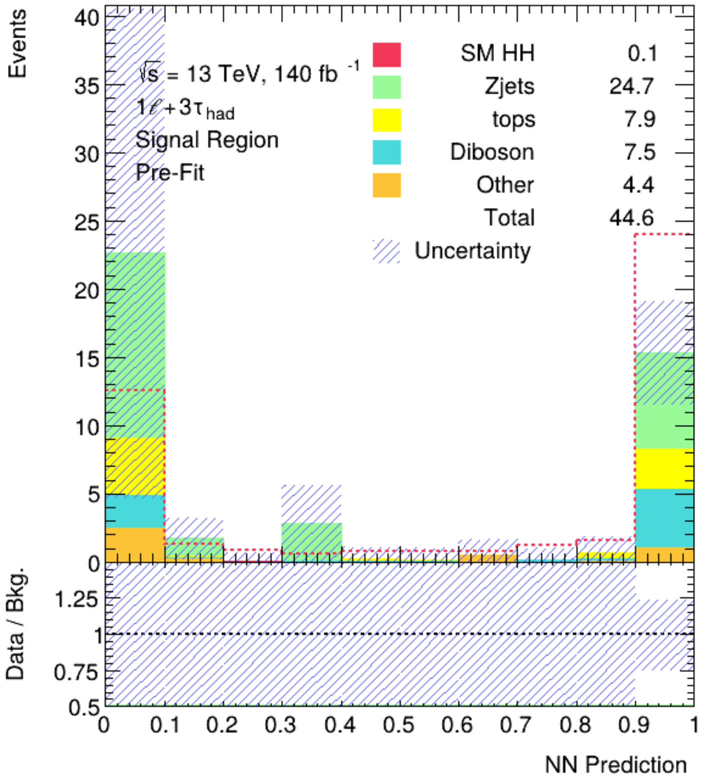


Figure 7: Neural Network prediction for $HH$ signal and background processes.

# 7 $\mu$ Parameter Expectation

The $\mu$ parameter is defined as

$$\mu = \frac{\sigma_{exp}}{\sigma_{SM}}, \tag{4}$$

where $\sigma_{exp}$ is the expected cross section for obtaining a signal with 95% Confidence Level (CL) and $\sigma_{SM}$ is the Standard Model predicted cross section. The simulation in this analysis uses the Standard Model expectation ($\mu = 1$). Using an Asimov fit in TRExFitter, the $\mu$ parameter for a 95% signal expectation is determined to be

$$\mu = 33.1.$$

## 8 Conclusions

It has been demonstrated that Machine Learning algorithms can be a powerful tool in analyzing particle physics data. In particular for the separation of signal and background processes for the $HH$ process. The performance of a binary classifiers has been evaluated. The most important features have been determined. The expected limit (median) at 95% CL is $\mu = 33.1$. Further studies should include a fake rate analysis and systematic uncertainties. The Machine Learning can also be further optimized. In this analysis, we also noted that the $W + jets$ process requires additional attention.

## Acknowledgements

I would like to extend my heartfelt appreciation to my supervisor, André Sopczak, for his exemplary guidance, support, and encouragement during the entirety of this project. I would also like to thank CERN for making this summer student project possible. The discussions, advice, help and support from Bartlomiej Henryk Zabinski is warmly acknowledged.

## References

1. Garcia, J. G. & Sopczak, A. *Study of Di-Higgs Boson Sensitivity in the Mode with One Light Lepton and Three Hadronically Decaying Tau Leptons - Summer Student Project* ATL-COM-PHYS-2024-1056. Geneva, 2024. https://cds.cern.ch/record/2920675.
2. *TRExFitter Documentation* https://trexfitter-docs.web.cern.ch/trexfitter-docs/.
3. Presperin, J. *Machine learning for ttH mechanism Higgs boson detection from CERN ATLAS data* Presented 14 Jun 2022 (2022). https://cds.cern.ch/record/2812400.
4. Dertat, A. *Applied Deep Learning - Part 1: Artificial Neural Networks* https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6.
5. *Tensorflow Weighted binary crossentropy* https://www.tensorflow.org/api_docs/python/tf/nn/weighted_cross_entropy_with_logits.