An investigation of over-training within semi-supervised machine learning models in the search for heavy resonances at the LHC

Benjamin Lieberman[1], Joshua Choma[1], Salah-Eddine Dahbi[1], Bruce Mellado[1,2] and Xifeng Ruan[1]

[1]School of Physics and Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, Wits 2050, South Africa
[2]iThemba LABS, National Research Foundation, PO Box 722, Somerset West 7129, South Africa

E-mail: benjamin.lieberman@cern.ch

Abstract. In particle physics, semi-supervised machine learning is an attractive option to reduce model dependency in searches beyond the Standard Model. Over-training of the model must be investigated when using semi-supervised techniques to train machine learning models for searches for new bosons at the Large Hadron Collider. In the training of classification models, fake signals can be generated due to over-fitting. The extent of false signals generated in semi-supervised models requires further analysis and therefore the probability of such situations occurring must be quantified on a case-by-case basis. This investigation of $Z\gamma$ resonances is performed using toy Monte Carlo samples normalised to mimic ATLAS data in a background-plus-signal region. Performing multiple runs, using random toy Monte Carlo samples, the probability of false signals being produced through over-training is investigated. The distribution of significance, of fake signals being generated using semi-supervised techniques, is found to form the positive side of a normal distribution for all background rejections and can therefore be said to be under control.

1. Introduction

In 2012 the ATLAS and CMS collaborations reported on the observation of a Higgs boson with a mass of 125 GeV [1, 2]. The Standard Model (SM) was completed by the discovery of the Higgs boson. The SM however, is not able to explain a number of phenomena that display substantial experimental evidence, such as Dark Matter, the origin of neutrino mass, the matter-anti-matter asymmetry, and a number of theoretical problems. These experimental discrepancies with the SM motivate the search for new bosons.

A 2HDM+S model, where S is a singlet scalar, was used in Ref. [3, 4] to explain some features of the Run 1 Large Hadron Collider (LHC) data. Here the heavy scalar, H, decays predominantly into SS, Sh, where h is the SM Higgs boson. The model predicts the emergence of multi-lepton anomalies that have been verified in Refs. [5, 6, 7, 8], where a possible candidate of S has been reported in Ref. [9]. The model can elaborate on multiple anomalies in astrophysics, if it is complemented by a Dark Matter candidate [10]. It can be further extended to account for anomalies, including the anomaly reported by Fermilab, in the $g-2$ muon experiment [11, 12, 13]. For a full review of anomalies, see Ref. [14].

The above mentioned motivates for the searches of heavy scalar resonances. We choose to investigate the search of $H \to Z\gamma$ with $Z \to \ell\ell$ and $\ell = e, \mu$. This is done using semi-supervision with topological features, as suggested in Ref. [15]. Semi-supervised learning is a machine learning technique where a model is trained on partially labelled data in order to reduce training biases. In this paper we focus on the potential over-training entailed in the the use of semi-supervision when confronting side-bands and the signal region using a neural network.

1.1. $Z\gamma$ Dataset

In this study we use the simulated Higgs like heavy scalar decaying to $Z\gamma$ ($pp \to H \to Z\gamma$) events, where $Z \to e^+e^-$ or $Z \to \mu^+\mu^-$. The simulated $Z\gamma$ dataset was produced using truth particles and particle reconstruction by ATLAS full simulation. The objects of analysis are electrons, muons, photons, jets and b-jets. The simulated non-resonant $Z\gamma$ dataset is used as it is the dominant background, representing more than 90% of the total background. This is therefore an ideal dataset to evaluate the extent of false signals generated during the Machine Learning (ML) training as any signals found within the dataset are a product of over-training and/or fluctuations within the phase space. Further details on the $Z\gamma$ dataset, including production mechanisms, are described in Ref. [16]. The important features selected and used for this analysis are the invariant mass, $m_{\ell\ell\gamma}$; invariant di-jet mass, m_{jj}; pseudo-rapidity of leading and sub-leading jets, η_{j1}, η_{j2}; number of jets, N_j; number of leptons, N_ℓ; number of b-jets, N_{bj}; transverse energy, E_T^{miss}; transverse energy significance, $\sigma_{E_T^{miss}}$ and the following difference in the azimuthal angles, $\Delta\Phi(ForwardJets, E_T^{miss})$, $\Delta\Phi(LeadingJet, E_T^{miss})$, $\Delta\Phi(LeadingJet, Z\gamma)$, $\Delta\Phi(Z\gamma, E_T^{miss})$. Example feature distributions are shown in Figure 1.

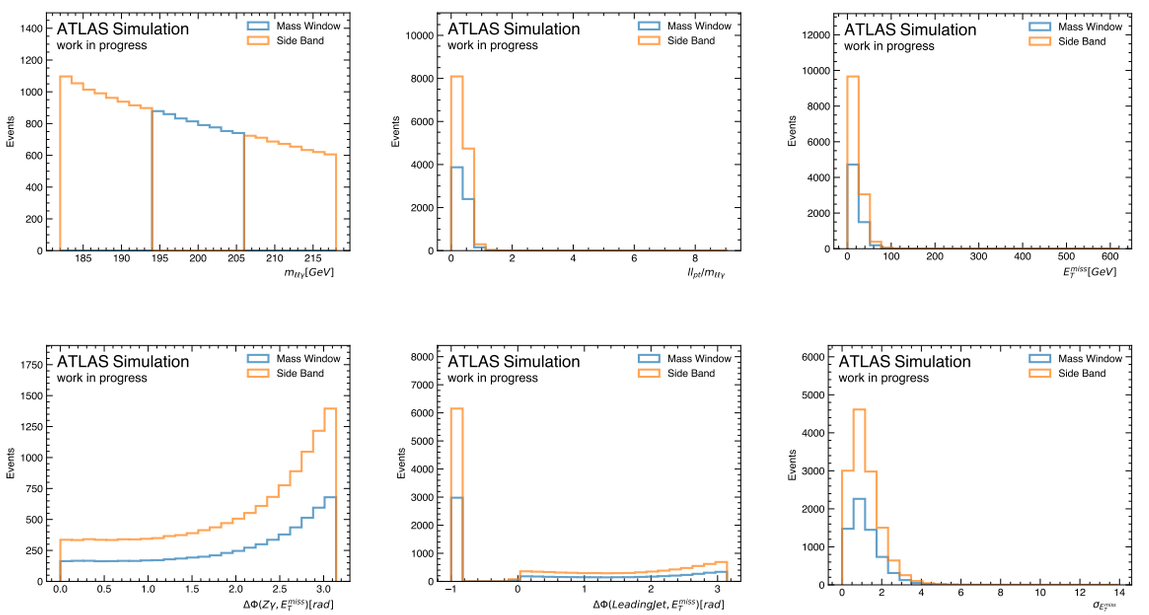

Figure 1: $Z\gamma$ dataset important feature distributions.

1.2. Semi-Supervised Machine Learning

In high energy physics, fully-supervised ML methods are frequently used as binary classifiers. The model is trained on labelled data where each event, $\vec{x_i}$, has a corresponding target/label,

 ISBN: 978-0-620-97693-0

$y_i \in \{0, 1\}$. The model can therefore be understood to be trained on two sample types, a signal sample, with label 1, and background sample, with label 0. For each given event, $\vec{x_i}$, the model generates a an output response, $\hat{y_i} \in (0, 1)$. Model training therefore aims to minimise the difference between the targets and the responses by using a loss function, usually in the form of binary cross-entropy:

$$\ell(y, \hat{y}) = -y \cdot \log \hat{y} + (1 - y) \cdot \log(1 - \hat{y}), \tag{1}$$

The minimising of the loss function to find the optimum solution in full-supervision ML can therefore be described using the following equation:

$$f_{full} = argmin_{f:\mathbb{R}^n \to [0,1]} \sum_{i=1}^{N} \ell(y_i, \hat{y_i}), \tag{2}$$

In semi-supervised models Ref. [17], the model is trained on partially labelled datasets. This means that the model is trained on one sample of pure background, labelled 0, and one unlabeled sample made up of a mixture of signal and background events, labelled 1. Therefore as apposed to full-supervised methods, Eq. 2, the semi-supervised method can be described using the following equation:

$$f_{semi} = argmin_{f:\mathbb{R}^n \to [0,1]} \sum_{K} \ell \left(\frac{1}{|K|} \sum_{i \in K} \hat{y_i}, y_K \right), \tag{3}$$

where K denotes the batches of training data and y_K is the signal ratio in each batch.

The quantification of uncertainties propagated within ML methods is vital in sub-atomic physics analysis as it allows both an understanding of the accuracy of any predictions made and exposes the level of validity of any ML based discoveries. The uncertainties in fully supervised techniques used in particle physics are well defined and extensively researched [18], however the uncertainties propagated in semi-supervised techniques have not been quantified to the same extent. This research therefore focuses on measuring the uncertainties, represented as fake signals, produced in the training of semi-supervised models within a given phase space.

2. Methodology

A benchmark centre of mass of 200 GeV is selected and each data sample is divided into a mass-window of 194 to 206 GeV and side-bands from 194 to 182 GeV and from 206 to 218 GeV. The model is trained on $Z\gamma$ events, with sample 0 and sample 1 consisting of events within the side-band and mass-window region respectively. As neither sample contains signal, there is no significant separation expected, in the model output, between the mass-window and side-band samples.

2.1. Deep Neural Network Model

The Binary Decision Tree (BDT), Multi-Layer Perceptron (MLP) and Deep Neural Network (DNN) classification models were compared. The area under the Receiver Operating Characteristic (ROC) curve is used to evaluate the classifier performance and the Kolmogorov-Smirnov test is used to measure over-training of the models. The DNN is selected as the optimum classifier as it showed the best classification score and lowest over-training. As the model is being used as a binary classifier, the cross-entropy loss (Eq. 1) is used as the loss function during training.

The DNN architecture implemented in this study consists of an input layer (360 neuron), four hidden layers (180, 180, 90, 180 neuron respectively) and an output layer with a single

neuron. The input and hidden layers use the rectified linear unit (ReLu) activation function and the output layer uses the sigmoid activation function. A learning rate of $1 \cdot 10^{-3}$ is used with a learning decay of $3 \cdot 10^{-4}$. The model is run for 8 epochs using a batch size of 1.

2.2. Toy Monte Carlo Sample Generation

In order to evaluate the over-training of the DNN model, the model must be run on a large number of statistically unique samples. To this end a toy Monte Carlo (MC) generator is used to extract random batches of events from the $Z\gamma$ simulated dataset. Each batch of events is normalised, using the corresponding event weights, to mimic data produced at the ATLAS experiment. Each sample contains approximately 45500 side-band events and 23000 mass-window events, labelled 0 and 1 respectively. Therefore the toy MC generator is used to produce a single random normalised sample for each given run of the model.

2.3. Evaluating Over-training on Invariant Mass

In order to calculate the significance of false signals being generated, the following steps are applied to each toy MC sample generated.

(i) The DNN model is trained on the given sample using events within the side-band and mass-window regions as sample 0 and sample 1 respectively. Once trained the DNN output, in the form of a response distribution (example in Figure 2), is generated.

(ii) Batches of 50, 60, 70, 80 and 90% of the total events are taken from the response distribution. Each batch is extracted by starting at the response distribution's maximum, 1, and moving towards the minimum, 0, until the required percentage of events are collected. The events of each batch are then mapped to their corresponding invariant mass.

(iii) The invariant mass, $m_{\ell\ell\gamma}$, distribution of each batch is than analysed in terms of the mass-window and side-band. This is done by fitting an exponential function and an exponential + Gaussian function to each batch's invariant mass distribution:

$$f(x) = n_0 \cdot e^{ax+bx^2}, \tag{4}$$

$$g(x) = n_0 \cdot e^{ax+bx^2} + n_1 \cdot e^{\frac{(x-\mu)^2}{2\sigma}}, \tag{5}$$

where n_0, a, b and n_1 are constants produced in the fit; μ is the mean (fixed at centre of mass) and σ is the standard deviation (as calculated by the fit). The exponential function, Eq. 4, is therefore used to describe the background, in the side-band and mass-window, and the Gaussian function, Eq. 5, is used to define signal, within the mass window. An example of the invariant mass distribution fits is shown in Figure 3.

2.4. Significance Calculation

The significance of fake signals generated, in the mass-window, due to over training can be quantified as the difference between the log-likelihoods of the two functions. The following steps are implemented:

(i) The log-likelihood can be calculated using a Poisson probability mass function, p_X, on the first n terms of the invariant mass distribution $\{X_n\}$. The probability mass function of a term x_i is:

$$p_X(x_i) = e^{-\lambda \frac{\lambda^{x_i}}{x_i!}}, \tag{6}$$

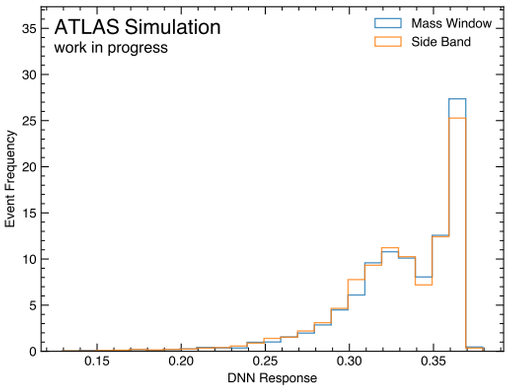

Figure 2: Example DNN response distribution output from a single toy MC sample.

where λ is the parameter of interest. The likelihood function, L, and log-likelihood, $\ln(L)$, can therefore be calculated as follows:

$$L(\lambda; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}, \tag{7}$$

$$\ln L(\lambda; x_1, x_2, ..., x_n) = -n\lambda - \sum_{i=1}^{n} \ln(x_i!) + \ln(\lambda) \sum_{i=1}^{n} x_i. \tag{8}$$

(ii) The log-likelihood of the two functions can then be used to calculate the model's uncertainty significance for the given run:

$$S_k = \sqrt{2 \cdot (\ln L_{eg} - \ln L_e)}, \tag{9}$$

where S_k is the Significance for the k^{th} run and L_{eg} and L_e are the log-likelihoods of the exponential + Gaussian function and the Exponential function, respectively.

(iii) Repeating the process with statistically random toy MC samples a number of times (initially 500 times) will produce the statistical deviations in significance of fake signals being generated. The uncertainty generated, within the semi-supervised model, can therefore be quantified. As the samples are limited by the MC statistics, the number of runs is limited to 500.

3. Results

3.1. Invariant Mass Distribution with Cuts

In order to analyse false signals generated in the training of the model, the DNN is trained and the output response distribution analysed, for each toy MC sample. An example of the response distribution produced in a single run can be seen below in Figure 2.

Background rejection batches, containing 50, 60, 70, 80 and 90% of the total events, are extracted from the DNN response distribution and mapped to their corresponding invariant masses. An example of the number of events extracted for each background rejection, in a single run, is shown in Table 1. The fit functions, Eq. 4 and 5, are applied to each invariant mass distribution in order to expose the extent of fake signals generated. Examples of the 60 and 80% background rejections, for a single run, are shown in Figure 3.

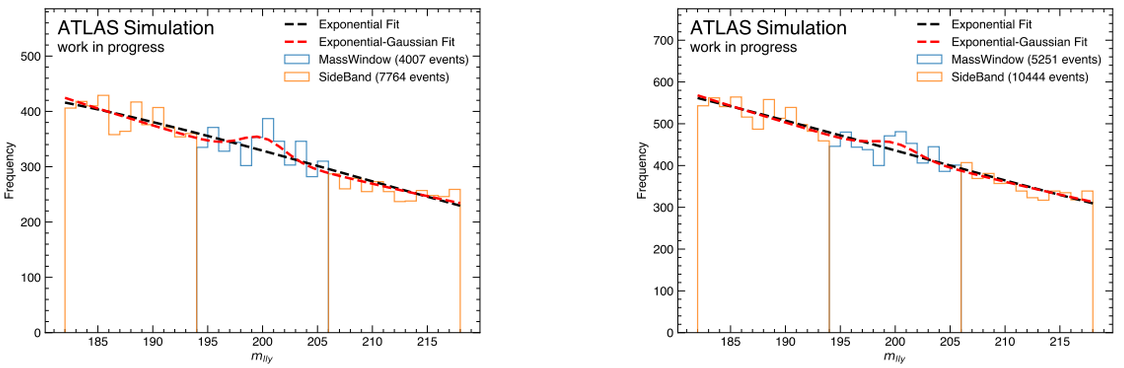

Figure 3: Example $m_{\ell\ell\gamma}$ distributions of 60% and 80% background rejections, for a single run.

3.2. Significance Distributions

For each background rejection, of a given run, the significance is calculated using the difference in the log-likelihoods of the fit functions, Eq. 9. Therefore for each background rejection, of a given run, the significance is calculated. The breakdown of background rejections and corresponding significance for an example run is shown in Table 1.

Table 1: Example of the number of events and corresponding significance, related to each background rejection, for a single run.

% Events	Mass-window events	Side-band events	Significance (σ)
50	3347	6462	1.80
60	4007	7764	2.85
70	4642	9091	2.15
80	5251	10444	1.87
90	5841	11816	1.36

Repeating this methodology on multiple toy MC samples, produces significance distributions for each background rejection. These significance distributions can therefore be used to quantify the extent of false signals produced in the model. The results below, in Figure 4, demonstrate examples of the significance distributions produced when the model is run on 500 toy MC samples.

4. Conclusions

The investigation into quantifying the uncertainty generated, through the over-training of semi-supervised techniques, using $Z\gamma$ resonances was performed using pure background toy MC generated samples and a semi-supervised DNN model. The invariant mass distributions for various background rejections was used to measure the fake signals produced by the model. This in turn was quantified through the calculated significance for the background rejections of each run. The significance distributions produced on 500 samples, Figure 4, form the positive side of a normal distribution for all background rejections. The significance distributions therefore verify

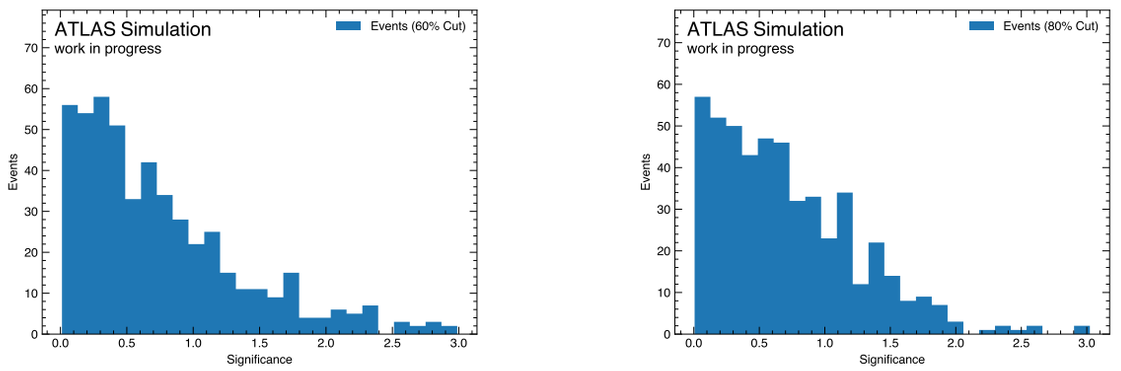

Figure 4: Significance distribution, for 60% and 80% background rejection, for 500 runs.

that the extent of fake signals generated, does not refute any scientific observations made using the semi-supervised technique. The study however is limited by the MC statistics produced using full simulation and future research should consider adopting generative models to increase the size and statistics of the dataset used for analysis.

References

[1] Aad G *et al.* (ATLAS) 2012 *Phys. Lett. B* **716** 1–29 (*Preprint* 1207.7214)
[2] Chatrchyan S *et al.* (CMS) 2012 *Phys. Lett. B* **716** 30–61 (*Preprint* 1207.7235)
[3] von Buddenbrock S, Chakrabarty N, Cornell A S, Kar D, Kumar M, Mandal T, Mellado B, Mukhopadhyaya B and Reed R G 2015 (*Preprint* 1506.00612)
[4] von Buddenbrock S, Chakrabarty N, Cornell A S, Kar D, Kumar M, Mandal T, Mellado B, Mukhopadhyaya B, Reed R G and Ruan X 2016 *Eur. Phys. J. C* **76** 580 (*Preprint* 1606.01674)
[5] von Buddenbrock S, Cornell A S, Fadol A, Kumar M, Mellado B and Ruan X 2018 *J. Phys. G* **45** 115003 (*Preprint* 1711.07874)
[6] Buddenbrock S, Cornell A S, Fang Y, Fadol Mohammed A, Kumar M, Mellado B and Tomiwa K G 2019 *JHEP* **10** 157 (*Preprint* 1901.05300)
[7] von Buddenbrock S, Ruiz R and Mellado B 2020 *Phys. Lett. B* **811** 135964 (*Preprint* 2009.00032)
[8] Hernandez Y, Kumar M, Cornell A S, Dahbi S E, Fang Y, Lieberman B, Mellado B, Monnakgotla K, Ruan X and Xin S 2021 *Eur. Phys. J. C* **81** 365 (*Preprint* 1912.00699)
[9] Crivellin A, Fang Y, Fischer O, Kumar A, Kumar M, Malwa E, Mellado B, Rapheeha N, Ruan X and Sha Q 2021 (*Preprint* 2109.02650)
[10] Beck G, Kumar M, Malwa E, Mellado B and Temo R 2021 (*Preprint* 2102.10596)
[11] Sabatta D, Cornell A S, Goyal A, Kumar M, Mellado B and Ruan X 2020 *Chin. Phys. C* **44** 063103 (*Preprint* 1909.03969)
[12] Abi B *et al.* (Muon g-2) 2021 *Phys. Rev. Lett.* **126** 141801 (*Preprint* 2104.03281)
[13] Aoyama T *et al.* 2020 *Phys. Rept.* **887** 1–166 (*Preprint* 2006.04822)
[14] Fischer O *et al.* 2021 *Unveiling hidden Physics Beyond the Standard Model at the LHC* (*Preprint* 2109.06065)
[15] Dahbi S E, Choma J, Mokgatitswane G, Ruan X, Lieberman B, Mellado B and Celik T 2021 *International Journal of Modern Physics A* ISSN 1793-656X URL http://dx.doi.org/10.1142/S0217751X21502419
[16] Aad G *et al.* (ATLAS) 2019 *ATLAS-CONF-2019-034*
[17] Dery L M, Nachman B, Rubbo F and Schwartzman A 2017 *Journal of High Energy Physics* **2017** 1–11
[18] Abdughani M, Ren J, Yang J M and Zhao J 2019 *Communications in Theoretical Physics* **71** 955