The use of Semi-supervision in the search for heavy resonances with the $Z\gamma$ final state at the LHC

Joshua Choma[1,*], Salah-Eddine Dahbi[1], Bruce Mellado[1,2], 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: [*]`nalamotse.joshua.choma@cern.ch`

Abstract. Unlike supervised learning which is known to assume a full knowledge of the underlying model, semi-supervised learning, weak supervision in particular allows with partial knowledge to extract new information from data. The objective of this study is to set up the search for heavy resonances at the electroweak scale with topological requirements. These resonances could be produced with different production mechanisms. In this case we will be focusing on the searches for new resonances in the $Z\gamma$ final state using the Monte Carlo simulated signal samples for 139 fb^{-1} of integrated luminosity for Run 2, collected at the LHC. The weak supervised learning approach will be implemented, which will then be compared to the performance of the full supervision approach.

1. Introduction

Large Hadron Collider (LHC) at CERN produces a very large amount of data, which is computationally intensive and requires super-computing abilities to process. This data is generated through proton-proton collisions, pp, at the ATLAS detector at high energies. The collisions at the LHC produce particles which physicists have been studying in search for a new physics beyond the Standard Model (BSM). For the probability of the BSM events to be produced, the particles have to be accelerated at extremely high energy and high luminosity.

Machine learning (ML), deep learning, in particular, comes across as one of the tools to use in this type of analysis because of its ability of handling complex and high dimensionality data. The application of ML to high energy physics started in the 1990s, used for analysis and which later developed into event identification and reconstruction in the 2010s [1]. There are quite a number of ML algorithms which have been used in physics which include support vector machines, boosted decision trees, kernel density estimation and artificial neural networks.

For the longest time ML has been following two most commonly known learning paradigms, namely, supervised and unsupervised learning. Supervised learning is known to assume full knowledge of the model since it is trained on labelled data. On the other hand, in unsupervised learning, the learning takes place without labels. The algorithm is expected to learn from itself by finding similarities in the data and assigning them to the same output unit. Weak supervision
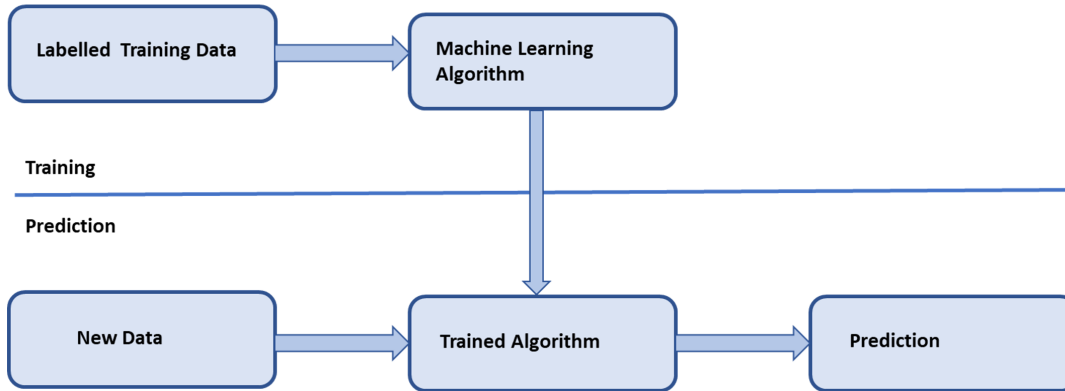
Figure 1: Schematic representation of supervised learning.

in addition to these two, it is a new paradigm that allows extraction of information with partial knowledge of the data.

The focus on this research is to search for resonances using $H \rightarrow Z\gamma$ in the final states in predefined categories shown in Table 1 where H is a Higgs-like scalar. The purpose of this study is to scan from 200 - 900 GeV to verify the ability of the proposed methodology, which will further be used to prepare a search for new phenomena in high mass in $Z\gamma$ final states at the LHC. As this is an ongoing study, this paper will present and focus more on the mass point of 200 GeV.

The rest of this paper is organised as follows: Section 2 discusses the ML techniques, Section 3 presents data selection and data preprocessing, while Sections 4 and 5 concludes this work and gives a brief discussion, respectively.

2. Machine Learning

This section gives a brief description of the methodology implemented in this study. Two ML techniques in the form of full supervised learning and weak supervised learning have been implemented and evaluated for the purpose of events classification. This is done to train the algorithm to learn what the signal and background events look like. These techniques have been implemented in conjunction with deep neural networks (DNNs) [2].

2.1. Full Supervised Learning

Full supervised classification is one of the most popular learning paradigms of ML. The name dictates that the dataset should come with labels. Each example $\vec{x_i}$ comes with a label $y_i \in \{0, 1\}$ in a case of binary classification task. Figure 1 shows a schematic representation of full steps involved in supervised learning. The purpose of this approach is to learn a mapping from x to y while minimising the loss function (see equation 2) which can be in a form of binary cross-entropy:

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

where $\hat{y}$ is the model output and y is the target output. The loss function is given by [3]:

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

where f is the predictor function, $\hat{y}_i$ is the ith model output, y_i is the corresponding target output and ℓ is the loss for a single example [4]. For this task, the training data contains labelled signal

Table 1: Yield for the considered processes normalised to the expected events yields and signal injection rates for VBF, ggH, ZH and WH signal samples for 139 fb^{-1} of integrated luminosity for Run 2.

Process	Selection	Sideband Region	Signal Region	Signal
ggH	Inclusive	13175	6444	160
ZH & WH	$MET_{sig} > 2.5$ GeV	808	394	40
ZH & WH	$N_{jets} \geq 2$ GeV; $60 < m_{jj} < 120$ GeV	839	414	40
VBF	$N_{jets} \geq 2$ GeV; $\Delta\eta_{jj} > 2$ GeV; $m_{jj} > 300$ GeV	498	245	30

and background events, with 1 and 0 used as the labels, respectively. This is done to ensure that the algorithm is able to generalize on an unseen data to make predictions. This learning paradigm requires a large amount of labelled data, especially when applied on a deep learning algorithm.

2.2. Weak Supervised Learning
Unlike supervised learning, semi-supervised learning, weak supervision in particular allows with partial knowledge to extract new information from the data. This is different from supervised learning which is known to assume a full knowledge of the underlying model. Weak supervised learning is less expensive compared to supervised learning since it takes less time manually labelling the data. Weak supervised learning enables the model to learn from data with imprecise labels [5], it is for this reason that it is regarded as cheap form of supervision [6, 7]. These benefits sparked a lot of interest from researchers in high energy physics [3]. The most commonly known types of weak supervision come in three ways, namely, *incomplete*, *inexact*, and *inaccurate* supervision [8]. The names are self-explanatory, *i.e. incomplete* supervision comes with incomplete labels, *inaccurate* supervision with inaccurate labels and *inexact* supervision for coarse-grained labels [8].

3. Data Selection and Preparation
This work explores the separation power of weak supervision technique in comparison with full supervision. The performance of these two techniques will be tested on ATLAS Monte Carlo samples. This corresponds to simulated non-resonant $Z\gamma$ dataset, as it is the dominant background, representing more than 90% of the total background. The signal in this research represents the simulated Higgs-like to $Z\gamma$ final state [9]. Data preprocessing plays a fundamental role in ML and has a significant influence on the performance of ML methods [10, 11, 12]. The data is normally scaled to the intervals of $[0, 1]$ and $[-1, 1]$ to ensure that features have the same degree of influence [13]. This ensures that the values use a common scale however, the difference in the ranges is not distorted. For this study MinMax scaler was used to normalize the data. MinMax scaler is defined by:

$$x'_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{3}$$

where x_i is the ith entry/record for the variable x, x'_i is the rescaled entry, whereas $min(x)$ and $max(x)$ represent the minimum and maximum entries, respectively. Table 1 shows the number of expected events yields for both sideband and signal region for all predefined categories for this mass point. The width of the $Z\gamma$ invariant mass for the sideband region is defined as 12% of center mass of the resonance Higgs-like signal while the signal region is 6%. Different signal production mechanisms (WH, ZH, ggH and VBF) were injected, these injected numbers are

Table 2: Maximum significance for different background rejections from the weak supervised learning DNN response distribution.

Category	Max Significance	Background Rejection (%)
Inclusive	1.97	0
$MET_{sig} > 2.5$ GeV; ZH	3.65	99
$MET_{sig} > 2.5$ GeV; WH	2.87	99
$N_{jets} \geq 2$ GeV; $60 < mjj < 120$ GeV; ZH	1.91	30
$N_{jets} \geq 2$; $60 < mjj < 120$ GeV; WH	1.88	0
$N_{jets} \geq 2$ GeV; $\Delta\eta_{jj} > 2$ GeV; $m_{jj} > 300$ GeV	2.31	90

defined by 2σ, where σ is the statistical uncertainty of the background in the signal region, given by:

$$\sigma = \sqrt{B_{MW}} \tag{4}$$

where B_{MW} represents the number of background events in the mass window region. The signal region for this mass is defined as $194 < m_{\ell\ell\gamma} < 206$ GeV whereas the sideband is between $182 < m_{\ell\ell\gamma} < 194$ GeV and $206 < m_{\ell\ell\gamma} < 218$ GeV.

Throughout this research, background events in the sideband region will be represented by sample 1. Sample 2 is made up of the background and the signal in the mass window region.

4. Results

A python API, Keras library with Tensorflow backend [14] was used for DNNs configurations. DNNs with four hidden layers of 200 nodes each and a single output node have been configured and implemented for this study. All of the hidden layers of the DNN used ReLu for an activation function and a sigmoid for the output layer. The input layer consists of 17 neurons, representing the kinematic features of the dataset. The separation power of the two techniques in conjunction with DNNs was evaluated using an ROC curve (receiver operating characteristic curve). Figure 2 shows the ROC curves for all the predefined categories with their respective signal injection. The performance in this case is measured by the area under the curve (AUC). Figure 3 (a) shows the DNN distribution plot from the weak supervised learning model when tested with pure signal and background. This is the selected plot for the VBF ($N_{jets} \geq 2$ GeV, $\Delta\eta_{jj} > 2$ GeV, $m_{jj} > 300$ GeV) category. This response distribution was further used to calculate the significance based on background rejection using the following equation:

$$\text{significance} = \frac{S}{\sqrt{S + B}} \tag{5}$$

where S is the number of signal events and B is the number of background events. Significance in this case can be regarded as the maximum ratio of signal to noise that is produced by the DNN classifier. The results of this significance are shown in Figure 3(b). The maximum significance together with the background rejection in terms of percentages are recorded in Table 2 for all categories. The two categories, inclusive and $N_{jets} \geq 2$ GeV; $60 < mjj < 120$ GeV, WH show to have a maximum significant at 0% background and this is due to having no clear separation between background and signal.

5. Discussion and Conclusion

In this study, we proposed the search for new resonances beyond the standard model using machine learning techniques, weak supervision and full supervision in particular. These learning
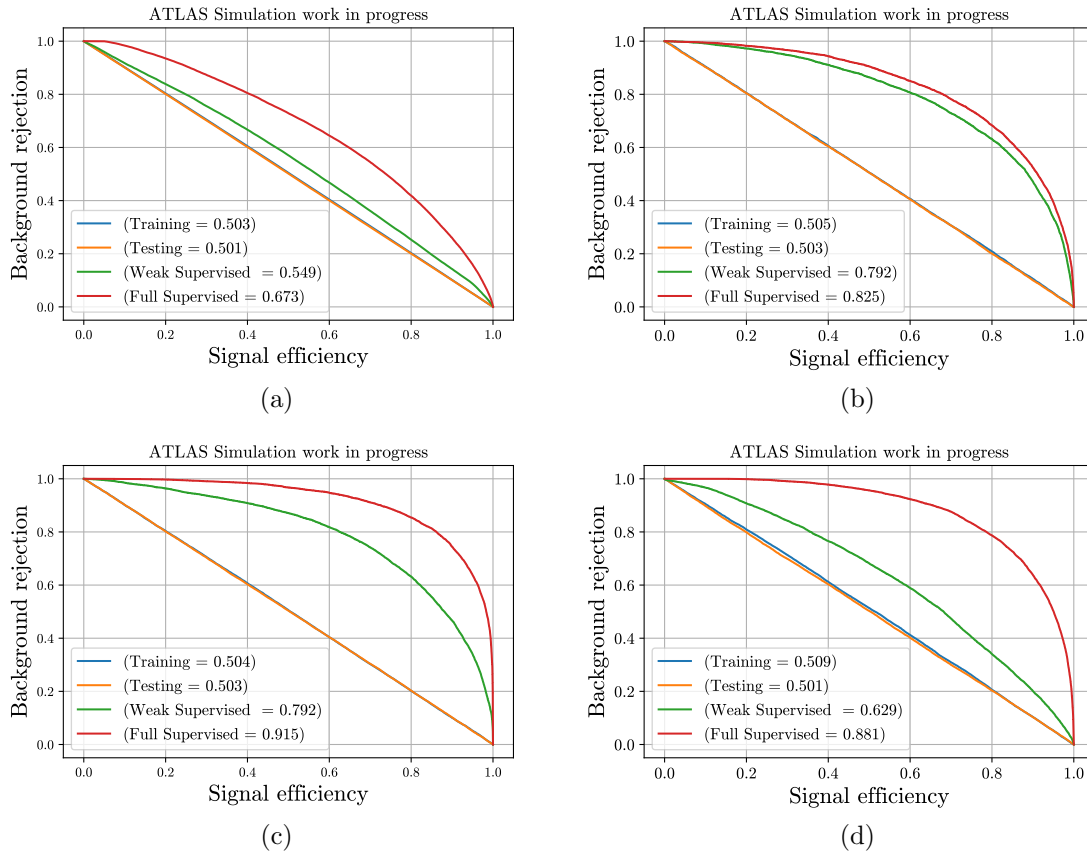
Figure 2: ROC curves showing weak supervising learning and full supervised learning results. Training and testing represent results for sample 1 and sample 2: (a) Inclusive, (b) $N_{jets} \geq 2$ GeV, $\Delta\eta_{jj} > 2$ GeV, $m_{jj} > 300$ GeV, (c) $MET_{sig} > 2.5$ GeV, WH and $N_{jets} \geq 2$ GeV, $60 < m_{jj} < 120$ GeV, ZH.
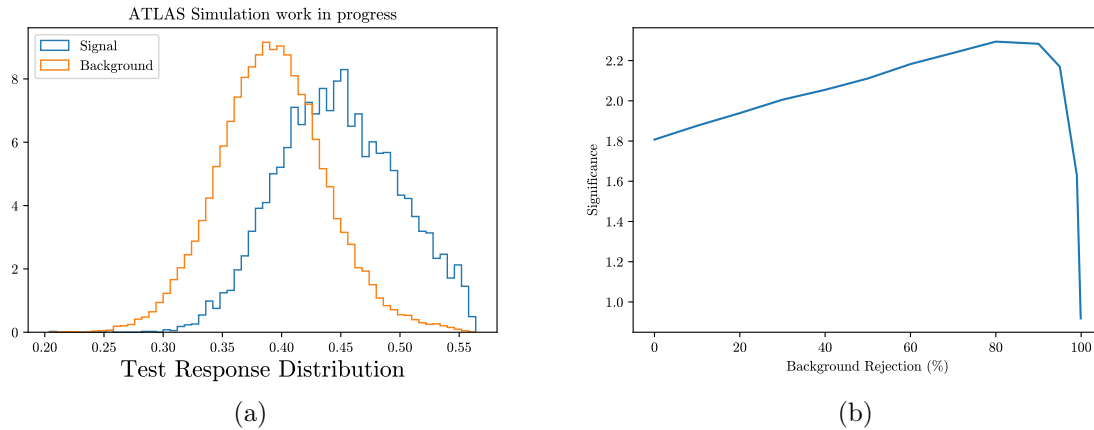

Figure 3: The two plots show the results for the VBF ($N_{jets} \geq 2$ GeV, $\Delta\eta_{jj} > 2$ GeV , $m_{jj} > 300$ GeV) category: (a) DNN response distribution for the weak supervised learning model, (b) significance calculated based on background rejection from the response distribution.

paradigms were used in conjunction with deep neural networks algorithm. The search was done in the $Z\gamma$ final state. The performance of the full supervision approach was compared to weak supervision. ROC curves were used as an evaluation metric to compare the two approaches. Based on this, it can be seen that the performance of weak supervision is reasonable and depends on the event configuration (see Figure 2) in comparison to full supervision for all categories. This is in agreement with the study carried out on the mass point of 105 GeV [2]. The study is used to setup for search for new phenomena in high-mass final states for the LHC.

References
[1] Albertsson K, Altoe P, Anderson D, Andrews M, Espinosa J P A, Aurisano A, Basara L, Bevan A, Bhimji W, Bonacorsi D *et al.* 2018 *Journal of Physics: Conference Series* vol 1085 (IOP Publishing) p 022008
[2] Dahbi S e, Choma J, Mellado B, Mokgatitswane G, Ruan X, Celik T and Lieberman B 2020 *arXiv preprint arXiv:2011.09863*
[3] Dery L M, Nachman B, Rubbo F and Schwartzman A 2017 *Journal of High Energy Physics* **2017** 1–11
[4] Cheong S Cs 229 project report: Weakly supervised classifiers with adversarial training in high-energy physics
[5] Dehghani M, Zamani H, Severyn A, Kamps J and Croft W B 2017 *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* pp 65–74
[6] Han J, Zhang D, Cheng G, Guo L and Ren J 2014 *IEEE Transactions on Geoscience and Remote Sensing* **53** 3325–3337
[7] Zhang F, Du B, Zhang L and Xu M 2016 *IEEE Transactions on Geoscience and Remote Sensing* **54** 5553–5563
[8] Zhou Z H 2018 *National Science Review* **5** 44–53
[9] Choi S, Muhlleitner M and Zerwas P 2013 *Physics Letters B* **718** 1031–1035
[10] Zhang S, Zhang C and Yang Q 2003 *Applied artificial intelligence* **17** 375–381
[11] Huang J, Li Y F and Xie M 2015 *Information and software Technology* **67** 108–127
[12] Keung J, Kocaguneli E and Menzies T 2013 *Automated Software Engineering* **20** 543–567
[13] Angelis L and Stamelos I 2000 *Empirical software engineering* **5** 35–68
[14] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado G S, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y and Zheng X 2015 TensorFlow: Large-scale machine learning on heterogeneous systems software available from tensorflow.org URL `https://www.tensorflow.org/`