LIP-STUDENTS-23-26

Anomaly Detection as a Valuable Tool for Uncovering Unexpected Phenomena at Colliders

Simão Silva Cardoso[1,a]

[1] *Universidade do Minho, Portugal*

Project supervisors: Nuno Castro, Rute Pedro, Miguel Peixoto

January 4, 2024

Abstract. In this paper, anomaly detection algorithms are studied, analysed, and compared as a tool for discovering new physics at colliders. Unsupervised methods utilized in semi-supervised regimes like Isolation Forest, Autoencoder, and Variational Autoencoder were trained only on simulated events from the Standard Model, representing the background, and then tested on simulated signals, new physics phenomena, to assess their ability to identify such events as anomalies. To benchmark the semi-supervised approach, fully supervised neural networks were also trained. The contamination of the background of the semi-supervised methods, the effect of the latent space on the performance of the Variational Autoencoder, and the correlation between these algorithms were also aspects of the study. The findings show AUC scores ranging from 0.9358 to 0.9994 depending on the type of signal and method used, which is a promising prospect for applying semi-supervised methods to detect collider data anomalies.

Keywords: Standard Model, ATLAS, LHC, Machine Learning, Neural Networks, Isolation Forest, Autoencoder, Variational Autoencoder, Latent Space

1 Introduction

1.1 The Standard Model and Beyond

The Standard Model (SM) of particle physics is the prevailing theory that accurately describes, considering the available experimental data, three out of the four fundamental forces in the Universe, namely electromagnetic, weak, and strong interactions. Moreover, it classifies all elementary particles that are currently known. Although it is the most successful theory of particle physics to date, the SM could be better as there are many open questions, such as the nature of dark matter, the possibility of unification of all four fundamental forces, and the matter-antimatter asymmetry. These problems may be solved with physics beyond the Standard Model (BSM). With this said the research on this topic can be assisted by using anomaly detection methods on colliders' data, finding possible signals that might hint at new physics.

1.2 ATLAS/LHC Experiment

ATLAS is the largest Large Hadron Collider (LHC) detector for particle colliders. It is built with many layers of detection instruments wrapped concentrically around the collision point to record highly electrically charged energetic particles, allowing them to be individually identified and measured. A compact and highly sensitive innermost detector measures their direction, momentum, and charge in each proton-proton (pp) collision. It consists of three different systems of sensors, all immersed in a magnetic field parallel to the beam axis that bends the paths of the charged particles so that their momenta is measured as precisely as possible. Beams of particles travel at LHC with energies up to seven TeV or speeds up to 99.999999% that of light and collide at the centre of the ATLAS detector, producing new particles that fly out in all directions. Only one in a million collisions are labelled as potentially exciting and recorded for further analysis. The ATLAS is vital to investigate a wide range of physics phenomena that might one day establish particle masses' origin, have good prospects for discovering dark matter, and cast light on unification and even quantum gravity [1].

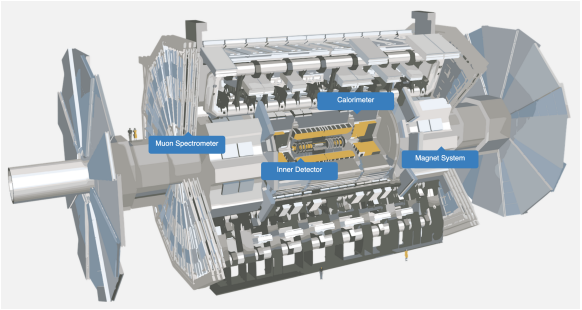

Figure 1. The ATLAS detector and its subdetectors [2].

1.3 Simulated Dataset

To study the anomaly detection (AD) methods, the datasets used for training and testing were based on pp collision data recorded at a centre-of-mass energy of $\sqrt{s}$ = 13 TeV with the ATLAS detector during 2015 and 2016. Monte Carlo events were simulated for all processes of interest [4], both the background and new physics type of signals datasets. (Fig. 2). This paper aims to assess how AD methods differentiate signals from the background, potentially leading to new discoveries in physics.

In order to use these datasets in the study, some common variables between them were selected as features to train the AD learning algorithms, such as the missing transverse momentum (MET), which is the negative vectorial sum of the transverse momenta of calibrated electrons,

[a] e-mail: simaocardoso23@ieee.pt

muons, small-R jets, and unassociated tracks [7], represented in Fig. 3, 4-momentum (p_x, p_y, p_z, e) of the jets and large-jets, scalar momentum sum of all objects (HT), and Delta represents higher-level features which can be derived from more basic ones. Thirty-one features were selected in total.

Background

This dataset simulates events predicted by the SM on pp collisions. These events include W and Z boson production in association with jets, top-quark production (both top-quark pair, $t\bar{t}$, and single-top-quark), non-resonant diboson production (WW, WZ and ZZ), and multijet production [7].

Resonant Dark Matter particles (S1)

The first type of signal is based on the production of resonant Dark Matter (DM) particles. An effective BSM model in which new mediators connect the SM particles and the DM candidates is usually considered. One production mechanism of new mediators is resonant. It produces a new scalar mediator ϕ decaying into a top quark and a DM candidate $\bar{\chi}$ [3].

Dark Matter production by two Higgs doublets (S2)

The second type of signal is based on the production of a heavy particle by the two Higgs doublets model (2HDM). There are two scenarios, but the one simulated is the 2HDM+a. This scenario is the most straightforward renormalizable and gauge-invariant extension of a simplified pseudoscalar mediator model. It adds a new pseudoscalar singlet that mediates the interactions between the SM and a singlet fermion $\bar{\chi}$ identified as the DM candidate [5].

Gluino pair (S3)

The third type of signal is based on the supersymmetric partners of quarks and gluons (squarks and gluinos). Squarks and the fermionic partners of the gluons, the gluinos ($\tilde{g}$), could be produced in strong-interaction processes at the LHC and decay via cascades ending with the stable lightest supersymmetric particle (DM candidate), which escapes the detector unseen, producing substantial missing transverse momentum [6].

Heavy Vectorial Triplet (S4)

Finally, the fourth type of signal is based on one kind of diboson resonance. It's the heavier version of the SM W and Z bosons, W' and Z' bosons, as parameterized in the Heavy Vector Triplet (HVT) framework, which can decay through $W' \longrightarrow WZ$ and $Z' \longrightarrow WW$ [7].

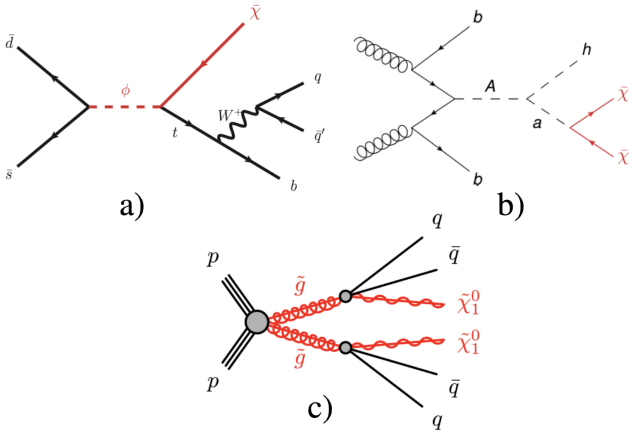

Figure 2. Feynamn diagrams of the signals: a) Resonant Dark Matter particles (S1), b) Dark Matter production by two Higgs doublets (S2), and c) Gluino pair (S3).

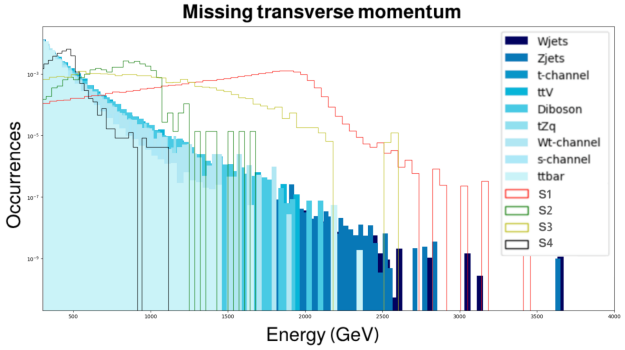

Figure 3. Missing transverse momentum (MET) of all background components and signals.

2 Anomaly Detection Methods

2.1 Supervised Neural Networks

In supervised learning, each event in the datasets is labelled 0 for background and 1 for each signal type. The goal consists of implementing an algorithm capable of high accuracy in predicting the label of an event when given its features. The algorithm in question is a deep neural network (NN). It's an algorithm whose architecture comprises layers with many neurons, leading to an output [8].

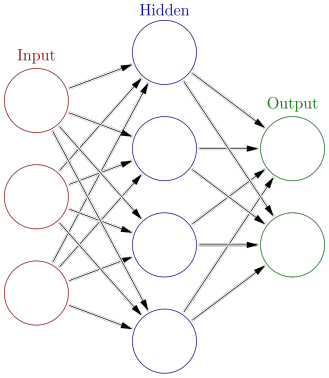

Figure 4. Schematic of a deep neural network.

Each layer is a mathematical function, where the output of one layer is the input of the next, which takes the following form:

$$\mathbf{f}_l(\mathbf{z}) = \mathbf{g}_l(\mathbf{W}_l\mathbf{z} + \mathbf{b}_l)$$

Where $\mathbf{z}$ is the input from the previous layer, l is called the layer index and can span from 1 to any number of layers, the function $\mathbf{g}_l$ is called an activation function. The parameters $\mathbf{W}_l$ (a matrix) and $\mathbf{b}_l$ (a vector) for each layer are learned via gradient descent. `ReLu` activation functions were used in the hidden layers except on the output layer, where a `sigmoid` was mandatory to truncate the values between 0 and 1. The loss function used was Binary Cross-Entropy (BCE), typically used in binary classifications. The goal is to find a set of $\mathbf{W}$ which minimises it via gradient descent. The BCE takes the following form:

$$\min_{\mathbf{W},\mathbf{b}} \frac{1}{N}\sum_{i}^{N}[y_i \log_2[\mathrm{NN}(\mathbf{x}_i, \mathbf{W}, \mathbf{b})]+$$

$$(1 - y_i)\log_2[1 - \mathrm{NN}(\mathbf{x}_i, \mathbf{W}, \mathbf{b})]]$$

Where $\mathbf{W}$ is the weight and $\mathbf{b}$ the bias learned by the NN, $\mathbf{x}_i$ the feature vector of the ith event, y_i is the correspondent true label and N the total number of events.

2.2 Semi-Supervised Methods

In semi-supervised learning, the datasets are not labelled. The algorithms are only trained on the background dataset and then tested on both the background and the signals. Theoretically, this allows the algorithms better to understand a background event than an unseen signal and then convert this process into an anomaly score.

Isolation Forest

The Isolation Forest (IF) algorithm [9] randomly selects a feature and then selects a split value between the maximum and minimum values of the selected feature. Since a tree structure can represent recursive partitioning, the number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node. This path length averaged over a forest of such random trees is used to measure the anomaly score. The shorter the path, the bigger the anomaly score.

Autoencoder

The Autoencoder (AE) [10] is a deep architecture that learns to compress (encode) and then decompress (decode) data through a bottleneck intermediate layer with a smaller dimensionality than the data, also called latent space. In this paper, the AE is trained by minimising the reconstruction error between the decoded dataset and the original through Mean Squared Error (MSE) that takes the following form:

$$\mathrm{loss} = \min_{\mathbf{W},\mathbf{b}} \frac{1}{N}\sum_{i}^{N} \|\mathrm{AE}(\mathbf{x}_i, \mathbf{W}, \mathbf{b}) - \mathbf{x}_i\|^2$$

Where $\mathrm{AE}(\mathbf{x}_i, \mathbf{W}, \mathbf{b}) = \hat{\mathbf{x}}_i$ is the decoded dataset. These reconstruction errors can be used as a measure for the anomaly score since, in theory, this algorithm better reconstructs the background and, therefore, has smaller anomaly scores than the signals.

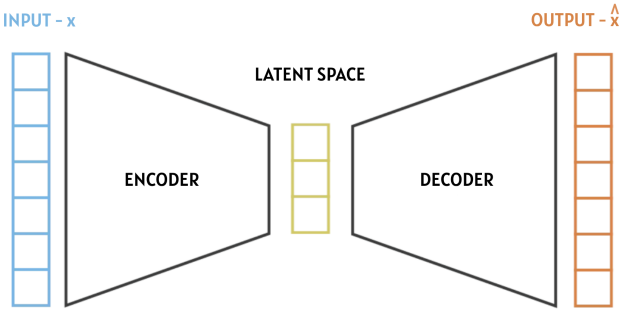

Figure 5. Schematic of a deep Autoencoder architecture.

Variational Autoencoder

The Variational Autoencoder (VAE) architecture also comprises an encoder and a decoder trained to minimise the reconstruction error between the decoded and initial data. However, a slight modification of the encoding-decoding process is applied to introduce some regularisation of the latent space. The algorithm is trained as follows: The input is encoded as a distribution over the latent space, and a vector (z) is sampled from that distribution:

$$z = \mu_x + \sigma_x \odot \zeta$$

Where μ_x is the mean vector, and σ_x is the deviation vector. A reparametrisation trick ($\zeta \sim N(0, 1)$) is applied to permit backpropagation of error through the network. Then, the sampled vector is decoded, and the reconstruction error can be computed. Another difference to the AE is that the loss is a sum of the MSE and the Kullback-Leibler Divergence (KL). $P(z|X)$ is the probability distribution that projects the data into latent space. But since we do not have that distribution, we estimate it using its simpler estimation $Q(z|X)$. Now while training, the encoder should try to learn the simpler distribution Q(z|X) such that it is as close as possible to the actual distribution. This is where KL divergence is used as a measure of the difference between two probability distributions. The loss function then takes the following form:

$$\mathrm{loss} = \mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}}) + \beta\mathrm{KL}[N(\mu_x, \sigma_x), N(0, 1)]$$

Where MSE and KL functions are used, and β is a constant to make sure both outputs from these functions are in the same order of magnitude.

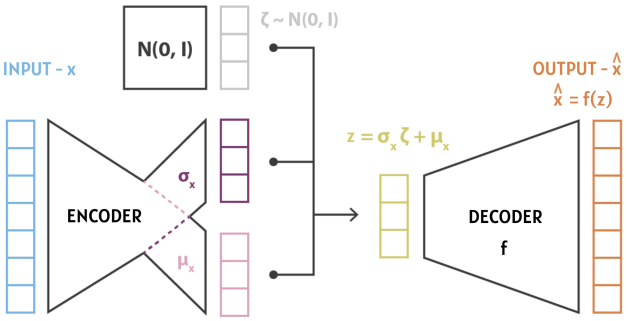

Figure 6. Schematic of a Variational Autoencoder architecture.

3 Implementation and Analysis of the Methods

The background and signals datasets were split into train, validation, and test datasets with equal statistical weights to ensure equal representativity. A standard scaler then transformed the datasets to ensure the individual features look more or less like standard normally distributed data (Gaussian with 0 mean and unit variance), as this is a common requirement for many machine learning estimators.

Neural Networks

Four similar algorithms were created for the NN, one for each signal type. They were all trained with the background and the respective signal train datasets and labels as a supervised binary classification problem. During this phase, the respective validation datasets were also introduced to provide an unbiased evaluation of the model fit on the training dataset and to be used as early stopping. After the algorithms were trained, the respective test datasets were used to make predictions. Since the outputs of the algorithms were a continuous float between 0 and 1 to each event on the test datasets, thanks to the `sigmoid` activation function of the output layer, they served as a measure of the anomaly score. These anomaly scores, in this case, four arrays, one for each algorithm, were then plotted in histograms, also called model scores, to study if the algorithms could distinguish the background from the signals successfully. To assess the performance, the anomaly scores were also used to plot a receiver operating characteristic curve (ROC curve). The area under the curve (AUC) for each ROC curve was also recorded—the closer to the unit, the better the performance.

The performance of NN does not significantly degrade when they are applied to another signal type than the one used for training, as long as these signals are similar from a topological point of view [10]. The algorithm trained with the signal S1 was then used to predict the other signals and repeat the analysis process to verify this statement.

Semi-Supervised Methods

For semi-supervised learning, all algorithms were trained only on the background and its validation datasets and then tested on the background mixed with each signal type individually. To analyse the results from the IF, since its anomaly scores go from -1 to 0, they were renormalized to go from 0 to 1. After that, they were used to plot the model scores and ROC curves for each signal.

Since AE and VAE are similar, they were studied similarly. Firstly, the algorithms, after training, encoded and decoded the background, originating the reconstructions. These were then plotted in histograms for each feature in the dataset against the original background features. Secondly, the algorithms also encoded and decoded each signal type. The MSE function was then applied between the reconstructions and the original datasets, obtaining the reconstruction losses. The logarithm of base 10 of the reconstruction losses was calculated, to avoid huge intervals between reconstruction losses and then renormalized from 0 to 1 as a metric for the anomaly scores. Finally, the usual model scores and ROC curves were plotted for each signal.

Contamination

In a real-life scenario, the datasets produced at the ATLAS detector might contain a new physics phenomenon hidden. To simulate this aspect, an algorithm must be trained with the background and a signal type but with the respective normalised combined weight, as this new physics phenomenon has very little statistical representativity. This approach was applied to the IF and VAE algorithms and tested like before. Degraded performances compared to uncontaminated algorithms were expected.

Latent space dimensionality

VAE being the most complex AD method, more profound research was put into it. Its latent space's smaller dimensionality than the other layers directly affects how well the algorithm reconstructs the original data. A study was conducted where the number of neurons of the latent space was varied, and the AUC score and the reconstruction loss were recorded to determine the most optimal latent space dimensionality.

4 Results

This section presents the results from the testing phase of these algorithms and further analysis. It is important to emphasise that the algorithms were optimised to produce better results, such as distinguishing accurately between background and signals.

4.1 Neural Networks

The four algorithms trained, one for each signal type, had similar architectures. The best architecture was shallow, with just one hidden layer and 128 neurons. The activation function that got the best results was the regular `ReLu` function, which filters only positive values to the next layer. This architecture also benefitted from a learning rate scheduler `ReduceLROnPlateau` whose callback monitored a quantity. If no improvement is seen for a `patience = 25` number of epochs, the learning rate is reduced by a factor of 10. The quantity monitored in this case was the AUC, which needed to be maximised.

Table 1. Gridsearch for the NN.

Variable HP	Possible Values
Number of Layers	[1,2,3]
Number of Neurons	[128,256]
Fixed HP	**Fixed Values**
Max Epochs	[200]
Batch Size	[4096]
Learning Rate (LR)	[0.001]

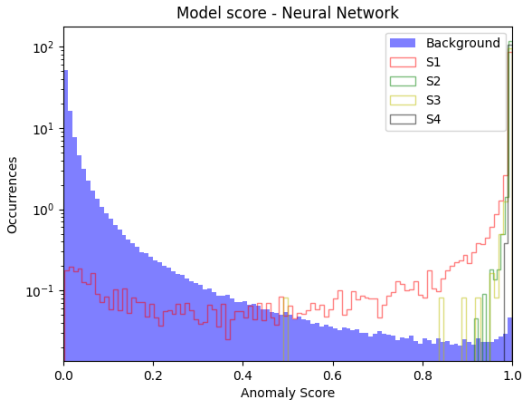

Figure 7. Model score for the NN trained on each signal. The occurrences are on a logarithmic scale. The filled colour blue represents the background predicted by the algorithm trained with S1, and the coloured lines are the signal types.

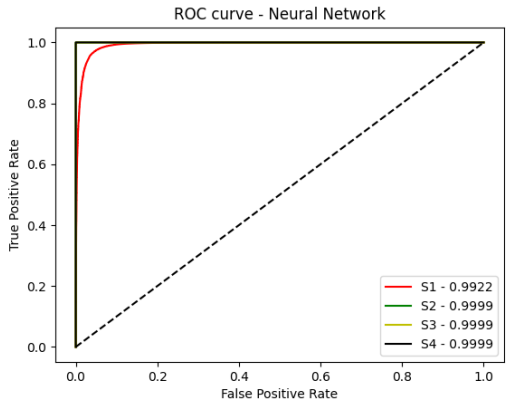

Figure 8. ROC curve for the NN trained on each signal.

The histogram from Fig. 7 shows that the NN could precisely predict the anomaly scores. The signals S2, S3, and S4 were almost entirely classified closely to the unit's score, while S1 spanned from 0 to 1. As for the background, all NN algorithms predicted scores that also spanned the entire spectrum. However, the number of occurrences above a threshold calculated by the average score of a background predicted by its algorithm plus one standard deviation would be statistically insignificant. This can be verified by Fig. 8 with perfect ROC curves and AUCs of 1.0 to the least troublesome signals S2, S3, and S4, while S1 got an AUC of 0.9922.

The algorithm trained with the S1 was then tested on the other signals, reproducing the results from Fig. 9 and Fig. 10. The performance of the NN was degraded as the anomaly scores it gave spanned the entire spectrum for the other signals, resulting in a lower distinguishing ability from the background. This also led to lower AUC values, suggesting that the signals are different from a topological point of view, mainly the signal S4.

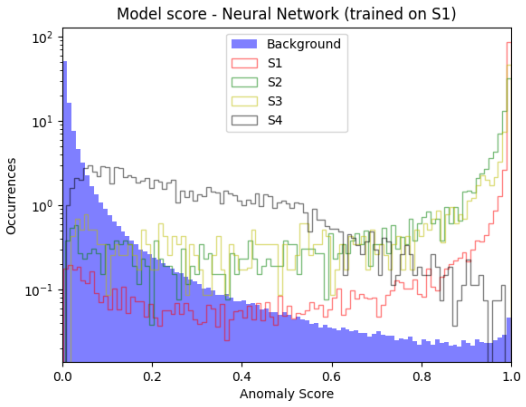

Figure 9. Model score for the NN trained on signal S1. The occurrences are on a logarithmic scale. The filled colour blue represents the background, and the coloured lines are the signal types predicted by the algorithm trained with S1.

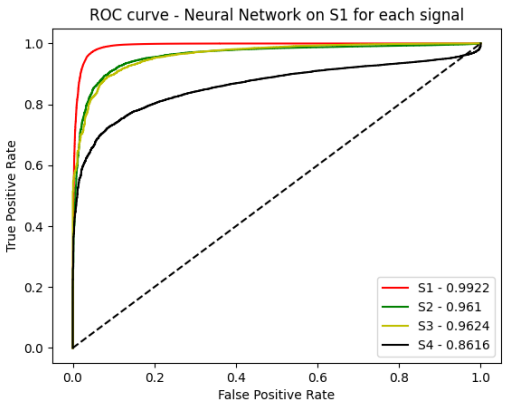

Figure 10. ROC curve for the NN trained on the signal S1.

4.2 Isolation Forest

The IF algorithm lacks customisation ability like the other algorithms studied. The only hyperparameter optimised was `n-estimators`, which represented the number of trees in the classification and was set to a maximum (100). Another hyperparameter mandatory to its use, called `contamination`, was left to default since its only purpose is to act as a threshold for classification on the anomaly scores.

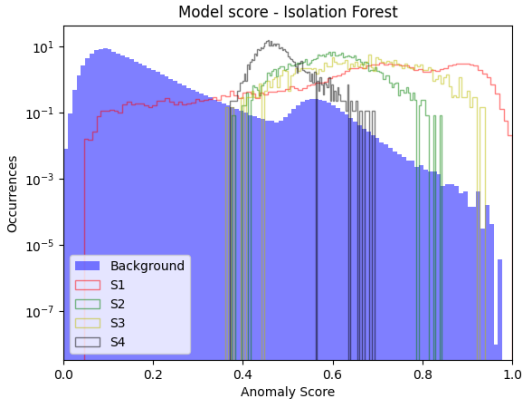

Figure 11. Model score for the IF. The occurrences are on a logarithmic scale. The filled colour blue represents the background, and the coloured lines the signal types.

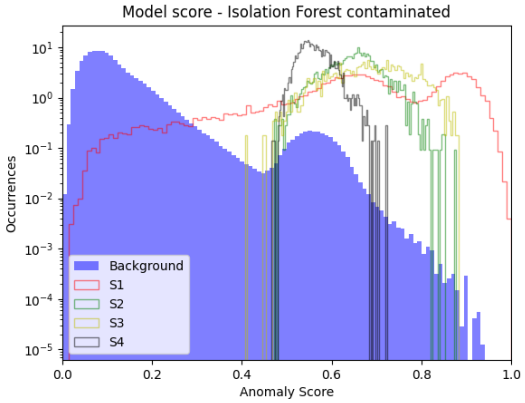

Figure 13. Model score for the contaminated IF. The occurrences are on a logarithmic scale. The filled colour blue represents the background, and the coloured lines represent the signal types.

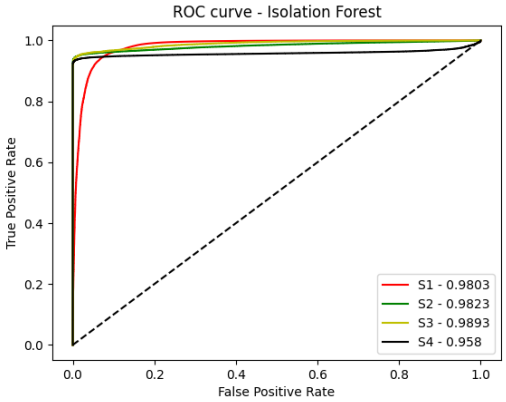

Figure 12. ROC curve for the IF.

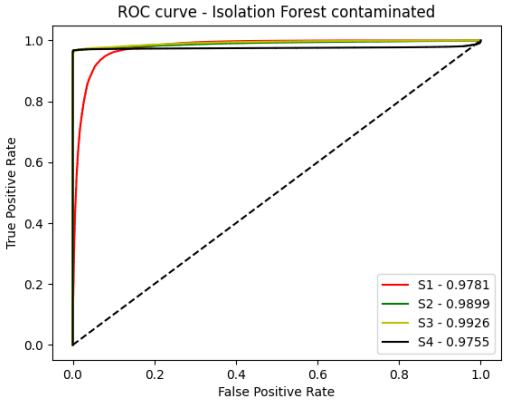

Figure 14. ROC curve for the contaminated IF.

As the NN, from Fig. 11, IF also gave the background and S1 both low and high anomaly scores, while the other signals got a smaller range but with higher average anomaly scores. The overlap between the background and the signals' distributions is relatively small, which led to high AUC scores, shown in Fig. 12. However, there was still a degradation compared to the NN AUC scores, as was expected when shifting from fully supervised learning to semi-supervised.

4.3 Contaminated Isolation Forest

Four IF algorithms were trained with the background and its respective signal type to simulate the contamination. The algorithms had the same hyperparameters as the original IF algorithm tested above. During the training phase, the normalised combined weight of the background and signals was used to get the right statistical representativity.

Compared to IF, the model score of contaminated IF in Fig. 13 showed a slightly better anomaly score range for the background and smaller intervals with higher averages of anomaly scores for the signals S2, S3, and S4, while the signal S1 remained practically the same. These changes can be noticed in the AUC scores represented in Fig. 14 14, which increased slightly, except for S1, compared to the scores given by IF.

4.4 Autoencoder

The latent space dimension most suitable was 8 neurons. To sustain this hourglass-shaped architecture and use 2^n number of neurons, the encoder was built with an input layer of 32 neurons and a hidden layer with 16 and, on the other side, the decoder was built symmetrically but with its output layer being made of 31 neurons, which is the number of features. Similarly to the NN, every layer had the `ReLu` function, except for the output layer, which had none.

Table 2. Gridsearch for AE.

Variable HP	Possible Values
Number of layers (Encoder=Decoder)	[2,3,4]
Number of Neurons	[16,32,64]
Fixed HP	**Fixed Values**
Max Epochs	[2000]
Batch Size	[4096]
Learning Rate (LR)	[0.001]

As the AE was trained only on the background training dataset, it learned to reconstruct its input accurately. Fig. 15 shows the background test dataset reconstructions of MET, one of the most critical features since its background and signal distributions are significantly different Fig. 3, and a 4-momenta large jet (p_x, p_y, p_z, e).

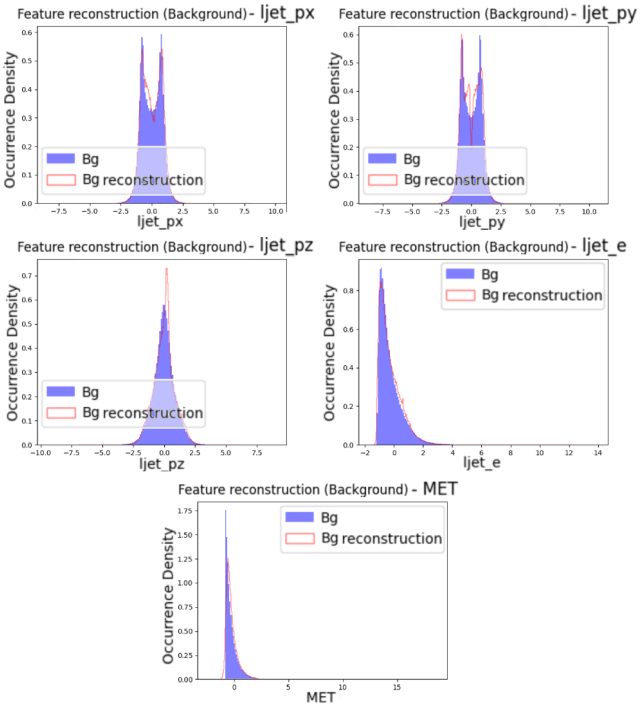

Figure 15. Original background and its reconstruction of the features MET and a large jet. The filled colour blue represents the original background test dataset, while the thick blue line represents the reconstructed background.

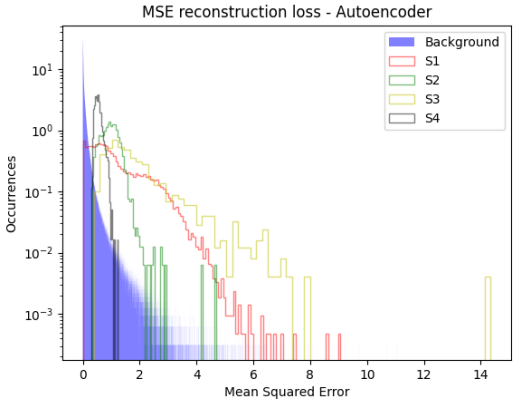

Figure 16. Reconstruction loss for background and all signals in logarithmic scale.

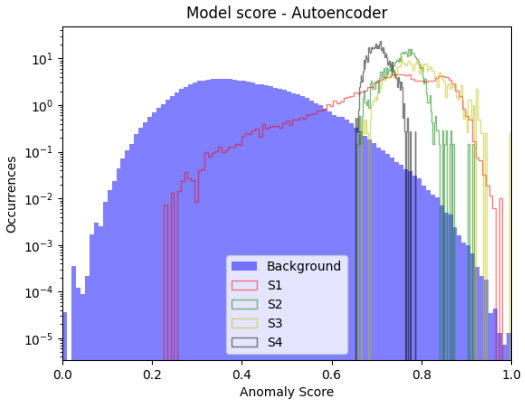

Figure 17. Model score for the AE. The occurrences are on a logarithmic scale. The filled colour blue represents the background, and the coloured lines the signal types.

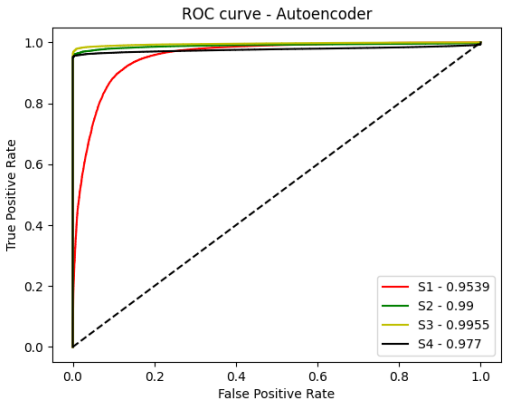

Figure 18. ROC curve for the AE.

This algorithm didn't reconstruct all the signals' features with the same accuracy as the background. These expected

results can be seen in Fig. 16, which shows the reconstruction loss for all datasets using MSE as a metric. Comparing these AUC scores to both IF algorithms, the overlap between the background and the signals' distributions is considerably smaller. AE performed better on the signals S2, S3, and S4, while S1 was harder to distinguish from the background, giving it a smaller AUC score.

4.5 Variational Autoencoder

A difference between VAE and AE is the optimization function, now Adadelta instead of Adam used on the AE, which provided slightly better reconstructions. For the latent space study, the architecture remained the same throughout the cases, only varying the number of neurons of the mean and deviation vector layers. As usual, the algorithm was trained on the background training dataset and tested on the background and all signals testing datasets, producing their reconstructions. Its effect on the AUC scores for each signal type and the background reconstruction loss was recorded in Fig. 19 and Fig. 20. From these results, it is possible to conclude that the 11 number of neurons seemed the best choice as it got the second-highest average AUC score and the lowest reconstruction loss.

Table 3. Gridsearch for VAE.

Variable HP	Possible Values
Number of layers (Encoder=Decoder)	[2,3,4]
Number of Neurons	[16,32,64]
Latent Space Dime.	[5 to 12]
Learning Rate (LR)	[0.01,0.1,1.0]
Fixed HP	**Fixed Values**
Max Epochs	[5000]
Batch Size	[4096]

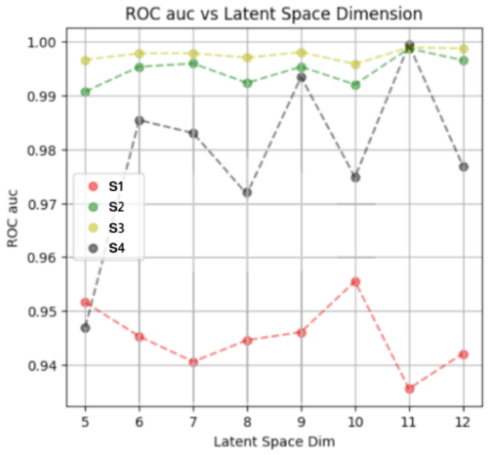

Figure 19. AUC scores vs number of neurons of the latent space layers for each signal (on the left) and their average (on the right).

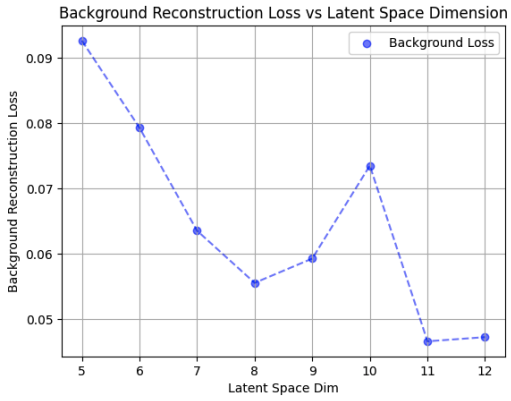

Figure 20. Reconstruction loss vs number of neurons of the latent space layers for the background.

The same study was conducted with the new optimised VAE to obtain its reconstructions of the background in Fig. 21. VAE's reconstructions of the same features compared to AE's were slightly more accurate on the large jet, while AE still prevailed on the MET. These better reconstructions led to better distinctions between the background and the signals due to the lower reconstruction loss against the original background, shown in Fig. 22.

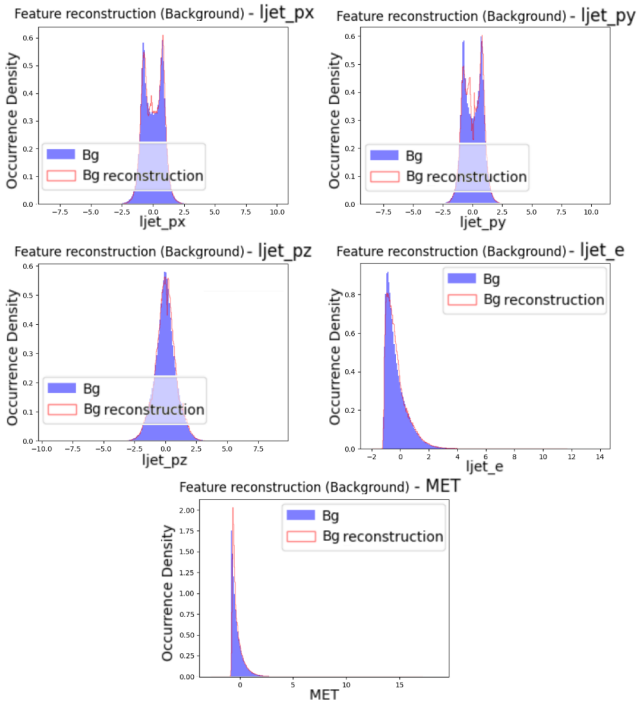

Figure 21. Original background and its reconstruction of the features MET and a large jet. The filled colour blue represents the original background test dataset, while the thick blue line represents the reconstructed background.

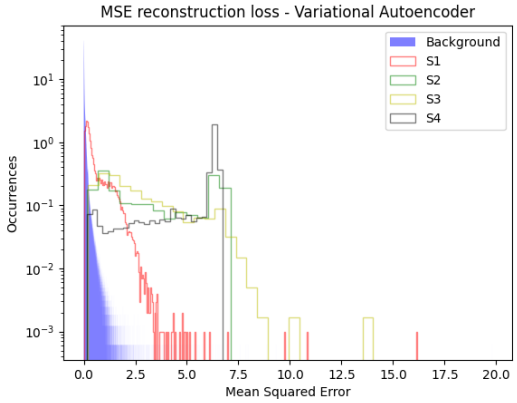

Figure 22. Reconstruction loss for background and all signals in logarithmic scale.

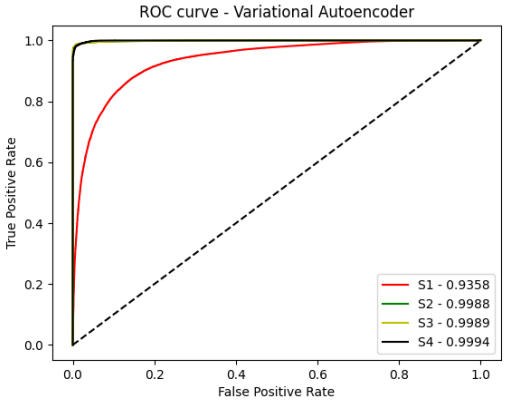

Figure 24. ROC curve for the VAE.

It is possible to verify that VAE reconstructed better the background and more poorly the signals S2 and S4 compared to AE, while S1 improved slightly (Table 4). The conversion of these results to the model score in Fig. 23 and the ROC curve in Fig. 24 show that VAE performed better than AE in all signals except on S1. This is possibly due to the MET feature, as S1's distribution is closer to the background's. VAE didn't reconstruct the MET as accurately as AE, leading to a lower anomaly score distinction between the signal and background.

Table 4. Reconstruction loss for both AE and VAE on the background and all signals.

	Rec. loss - AE	Rec. loss -VAE
Background	0.434 ± 0.127	0.349 ± 0.103
Sinal S1	0.738 ± 0.109	0.593 ± 0.113
Sinal S2	0.758 ± 0.033	0.797 ± 0.087
Sinal S3	0.790 ± 0.052	0.781 ± 0.076
Sinal S4	0.705 ± 0.021	0.860 ± 0.065

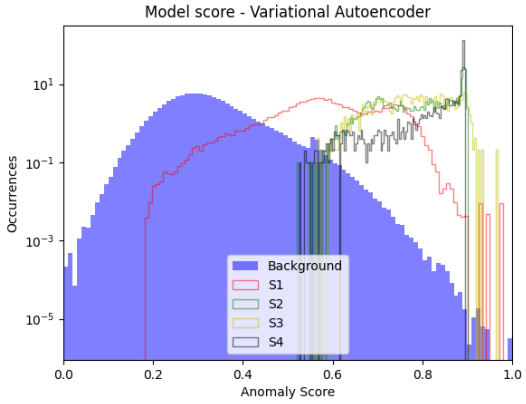

Figure 23. Model score for the VAE. The occurrences are on a logarithmic scale. The filled colour blue represents the background, and the coloured lines the signal types.

4.6 Contaminated Variational Autoencoder

Similarly to the contaminated IF, four VAE algorithms were trained with the background and a respective signal type to simulate the contamination. The algorithms had the same hyperparameters as the original VAE algorithm tested above. Once again, the normalised combined weight of the background and signals was taken into account to get the right statistical representativity.

Compared to VAE, the model score of contaminated VAEs in Fig. 25 showed a slightly worse anomaly score range for the background and more considerable intervals with lower averages of anomaly scores for the signals S2 and S3. The signal S4 also showed a lower average of the anomaly scores but with a similar interval between minimum and maximum scores, while the signal S1 boosted its average. These changes can be noticed in the AUC scores represented in Fig. 26, which decreased slightly, except for S1, compared to the scores given by VAE.

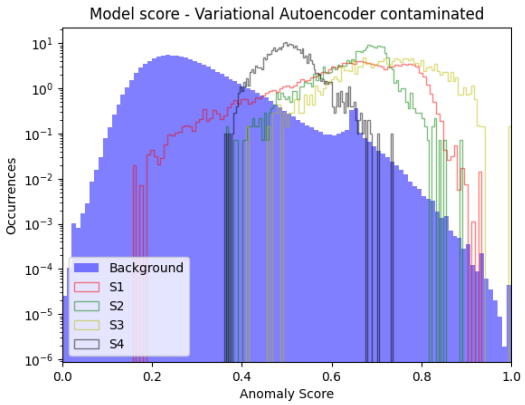

Figure 25. Model score for the contaminated VAE. The occurrences are on a logarithmic scale. The filled colour blue represents the background, and the coloured lines the signal types.

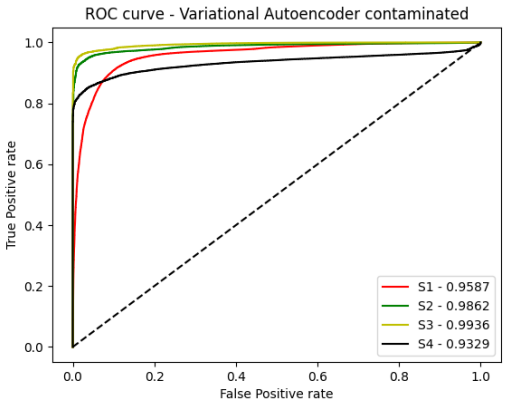

Figure 26. ROC curve for the contaminated VAE.

4.7 Correlation Between algorithms

Different AD algorithms might *see* various anomalies, leading to uncertain results. Scatter plots with the anomaly scores of each semi-supervised method were made for each signal type to see if this is the case. These scatter plots represent two-dimensional distributions of the anomaly scores for the different AD methods and, on the diagonal, the distribution of the model score per method.

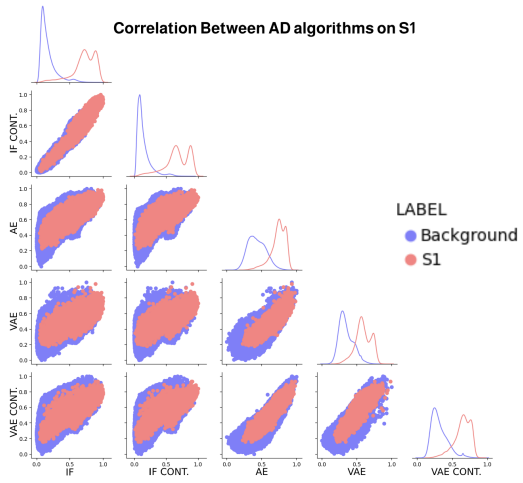

Figure 27. Scatter plots of anomaly scores on the signal S1.

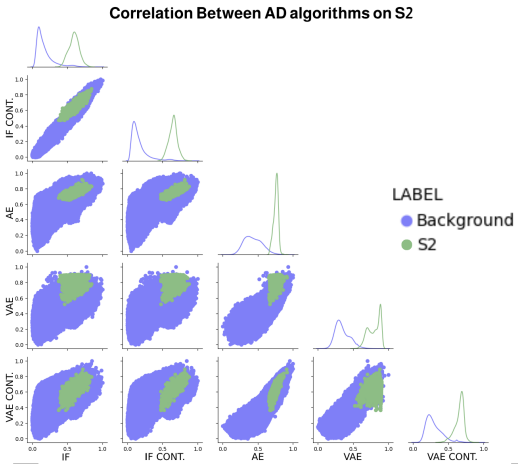

Figure 28. Scatter plots of anomaly scores on the signal S2.

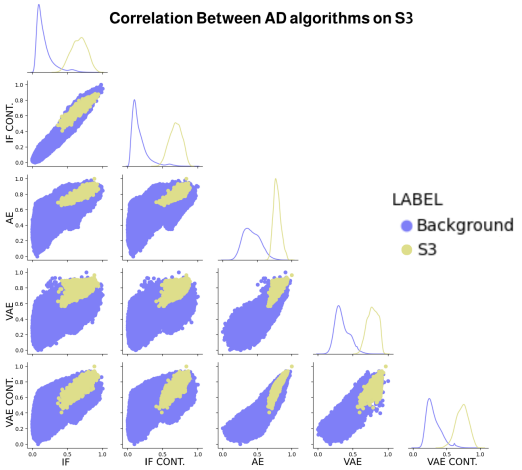

Figure 29. Scatter plots of anomaly scores on the signal S3.

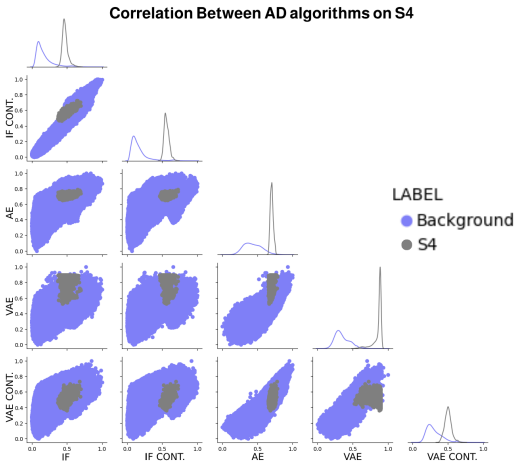

Figure 30. Scatter plots of anomaly scores on the signal S4.

The signal S1 led to lower correlations between algorithms as there appear to be no repeating scattered patterns of its anomaly scores. For the other signals, similarities in the shape and location of the two-dimensional plots could be seen. Also, these clusters are smaller than signal S1, which almost takes the same shape as the background. For the signals S2, S3, and S4, it's safe to assume that one AD algorithm is enough to uncover the correct outliers in the dataset; on the other hand, for the S1 signal, it is perhaps better to use more than one AD algorithm to ensure the proper detection.

4.8 AD Algorithms Score

To summarize the results from this study, the AUC scores of every semi-supervised method for each signal type were compared in the graph in Fig. 31.

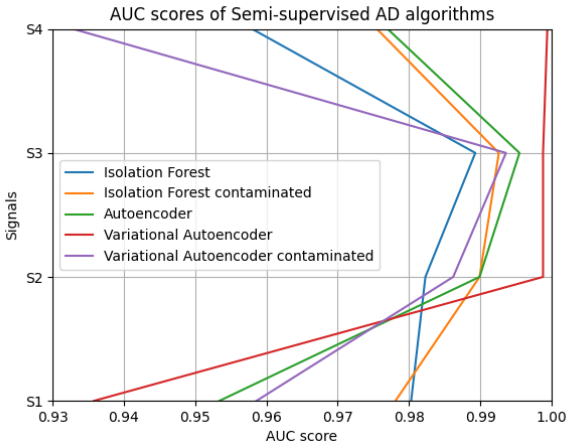

Figure 31. AUC scores by every semi-supervised method for each signal type.

All AD algorithms showed an overall excellent performance, some even at the same level as NN. Deeper algorithms like AE and VAE presented to be the best at distinguishing signals like S2, S3, and S4 but fell short on signal S1 to a shallow algorithm like IF, which got a more constant performance throughout the signals. While VAE got the highest AUC scores, the contaminated IF scored the highest average.

5 Conclusion

In this paper, three distinct semi-supervised AD algorithms were studied, one shallow and two deep, which were trained on simulated pp collision data at $\sqrt{s} = 13$ TeV. Compared to fully supervised algorithms, these AD algorithms performed well with different types of anomalous events achieving the top AUC scores: 0.9803 (IF), 0.9988 (VAE), 0.9989 (VAE), and 0.9994 (VAE) for the signals S1, S2, S3 and S4, respectively. However, there were some discrepancies: some types of signals are harder to distinguish from the background, leading these algorithms to have different notions of *outlyingness*. Contamination was also considered, which seemed to favour shallow algorithms and degrade deeper ones.

Distinct algorithms for AD are highly effective in isolating diverse types of BSM physics. Furthermore, these algorithms can complement each other in unsupervised searches for new physics, making them potential tools in particle physics research.

Acknowledgments

I am deeply grateful to Nuno Castro for the opportunity to conduct this research as a member of the LIP team. I also thank Rute Pedro for providing me with the necessary resources for this study. Finally, I thank Miguel Peixoto for assisting me with code debugging and generously sharing tips to improve the algorithms.

Code Source

GitHub link: https://github.com/sscardoso23/AD-ATLAS

References

[1] J. Ellis, *Physics Beyond the Standard Model*, Nucl.Phys.A827:187c-198c (2009), arXiv:0902.0357.

[2] *The ATLAS Detector*, ATLAS Experiment, (2023), Retrieved from Atlas Detector.

[3] The ATLAS Collaboration. *Search for invisible particles produced in association with single top quarks in proton-proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, ATLAS CONF Note, ATLAS-CONF-2022-036 (2022).

[4] The ATLAS Collaboration. *Search for large missing transverse momentum in association with one top-proton–proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Journal of High Energy Physics, 41 (2019), arXiv:1812.09743.

[5] The ATLAS Collaboration. *Search for squarks and gluinos in final states with jets and missing transverse momentum using 36 fb^{-1} of $\sqrt{s}$ = 13 TeV pp collision data with the ATLAS detector.* Phys. Rev. D, 97 112001, (2018), arXiv:1712.02332.

[6] The ATLAS Collaboration. *Search for dark matter produced in association with a Standard Model Higgs boson decaying into b-quarks using the full Run 2 dataset from the ATLAS detector*, Journal of High Energy Physics, 11 209 (2021), arXiv:2108.13391.

[7] The ATLAS Collaboration. *Search for heavy diboson resonances in semileptonic final states in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, Eur. Phys. J.C, 80 1165 (2020), arXiv:2004.14636.

[8] Andriy, Burkov, *The Hundred-page Machine Learning Book*. Andriy Burkov (2019).

[9] Liu, Fei Tony, Ting, Kai Ming, and Zhou, Zhi-Hua, *Isolation forest*, Data Mining, IEEE 8th International Conference on Computer and Communications (2022).

[10] M. Crispim Romao, N. F. Castro, R. Pedro. *Finding New Physics without learning about it: Anomaly Detection as a tool for Searches at Colliders*, Eur.Phys.J.C, 81 1, 27 (2021), arXiv:2006.05432.

[11] *Understanding Variational Autoencoders (VAEs)*, Towards Data Science (2019), Retrieved from Understanding Variational Autoencoders.