Applications of the XGBoost Machine Learning Algorithm in Particle Physics

Shizhe Liu, Sasan Hapuarachchi, Pruthvi Shrikaanth, William Shi, Joel Regi

King Edward VI Camp Hill School for Boys

Abstract

The rise in technological developments in artificial intelligence has unlocked new avenues of exploration in the intersection of machine learning (ML) and particle physics. We evaluated the potential of the XGBoost (ML) algorithm (a powerful gradient-boosted decision tree classification algorithm) to streamline the process of identifying rare particle decays. We achieved this by comparing the XGBoost's performance in four different classifications in particle physics with the performance of existing classification methods, such as restrictive cuts. We found that whilst in some cases the algorithm provided near-perfect prediction results, the algorithm was overly rigorous in other cases, leading to large numbers of signal events being dismissed as background events by the algorithm.

1. Introduction

1.1 Motivation

Currently, the process of searching for rare particle decays presents a significant challenge for particle physicists, as these decays can only be found in a tiny proportion of the millions of events picked up by the sensors in particle detectors. The emergence of the XGBoost, a powerful classification algorithm, holds potential in aiding particle physicists to streamline this challenge, as it may provide a better alternative to existing methods in identifying the events that contain elusive particle decays.

However, there are limited studies on the performances of the XGBoost algorithm in particle physics classifications. Therefore, we aim to address this gap by identifying the specific areas within particle physics where XGBoost excels and to compare its performance with existing methods, such as applying restrictive cuts, through evaluating its performance in carrying out four different classifications to look for:

1. Higgs boson events
2. Supersymmetry events
3. Beyond standard model Z' events
4. Kaluza klein graviton events

1.2 XGBoost Classification Algorithm

As advancements in Artificial Intelligence continue to be made, increasingly powerful machine learning algorithms are being developed at a rapid pace, and the XGBoost algorithm is a powerful classification algorithm that has arisen as a result of this technological revolution. It uses gradient boosted decision trees to provide accurate prediction results. Boosting is a technique where new models correct errors made by existing ones and are added one by one until no further improvements occur, and gradient boosting allows models to predict the errors made by the previous models to help give an accurate prediction result. The main advantages of the XGBoost are its exceptional speed and accuracy due to its ability to discern subtle patterns in the training data to provide accurate predictions, which may prove to be invaluable when carrying out particle physics classifications. (1)

1.3 AMS Metric

We evaluated the performances of the XGBoost classifier algorithm at carrying out the different classifications using the Approximate Median Significance (AMS) metric. When providing the true positive and false positive rates of the classification, the AMS metric uses the Wilks theorem to compare the probabilities of observing the signal background hypothesis and the background only hypothesis, to provide an overall figure representing the performance of the classifier. Therefore, the AMS provided us with a standardised way of assessing the XGBoost algorithm across the different applications. (2)

2. Results

2.1 Higgs Boson Events Classification

Prior research advancements in particle physics have revealed that it is possible to represent every particle as a wave in a quantum field (the Higgs field), which suggests that there would be a particle associated with this field (the Higgs boson) (3). When we wrote a program that trained the XGBoost algorithm to classify Higgs Boson events using over 400,000 samples of data (4) (containing a mixture of signal and background events), and tested it with another set of 400,000 events (ATLAS Collaboration), all of the test events that the XGBoost classifier had classified as signal events were indeed a real signal event. This yields an AMS score of *infinity*, which suggests that the XGBoost algorithm performed well in classifying Higgs Boson events.

This can be reinforced by observing the ROC curve, which compares the true positive rate against the false positive rate. As evident in figure 2.1.1, the TPR initially increases significantly, further suggesting that the XGBoost is accurate.

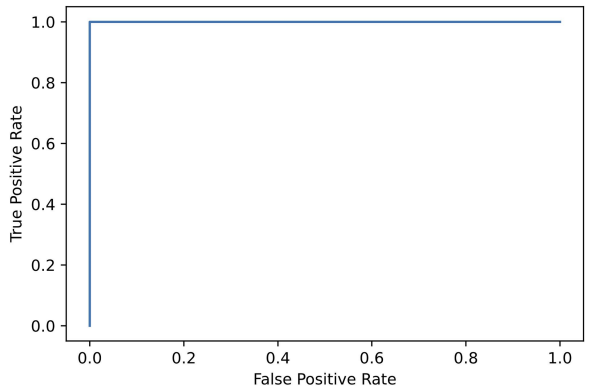


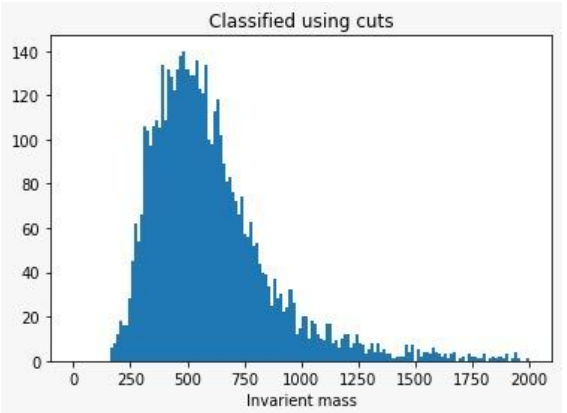
Figure 2.1.1: Higgs Boson Classification ROC Curve

2.2 Beyond Standard Model Z' Events Classification

A type of particle we looked at is the Z' boson, hypothesised with the Topcolor model - a model for electroweak symmetry breaking, in which a top anti-top pair forms a composite Higgs boson. In this model, a Z' boson was predicted to exist, which is the particle that we decided to investigate to train the XGBoost classifier with.

To find these particles, we can look at the decay channel of it. In this case, it decays into a top anti-top pair in events with a single charged lepton, large-R jets and missing transverse momentum. The lepton must have a transverse momentum > 30 GeV, missing transverse energy > 20 GeV, a small-R jet close to the lepton, a large-R jet passing the top tagging requirements(mass > 100 GeV, N-subjettiness ratio < 0.75) etc.

Finally, we plot the histograms of the invariant masses from the original (Fig 2.2.1) and the results from the XGBoost classifier (Fig 2.2.2), shown below. The AMS value obtained from the graph is .



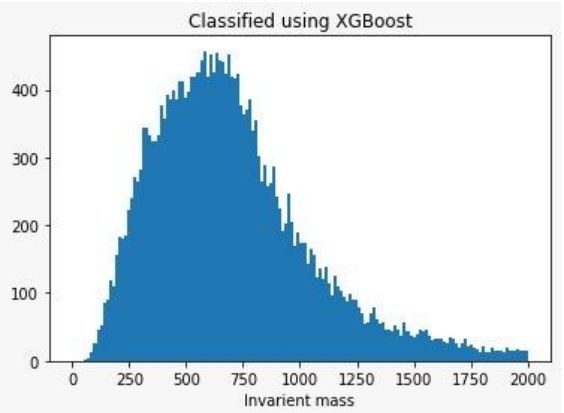


Figure 2.2.1: Z' events classified using restrictive cuts
Figure 2.2.2: Z' events classified using the XGBoost algorithm

2.3 Kaluza Klein Graviton Events Classification

One type of particle we shall look at is a Kaluza Klein graviton hypothesised using the Randall-Sundrum model (5) (a model for gravity where gravity propagates through warped

extra dimensions). In a similar way to atoms having excited states or low energy states, particles can have corresponding Kaluza Klein states where the particle has extra mass (instead of energy) in other dimensions (6).

In order to find the Kaluza Klein graviton we can search for the particles it decays into, in this case the particle decays into a gamma-gamma pair. So to find this particle we need to bump hunt gamma ray photons with transverse energies over 20 GeV (5). To do this we made the following cuts to the ATLAS data set: the event must have two photons, it must activate the photon trigger, both photons must have a transverse energy greater than 20GeV. Once we obtained our data points we subtracted any data points that could have been formed by the Higgs -> GammaGamma decay channel and plotted a graph of transverse energy against frequency using a fitting function.

XGBoost classifiers was trained with data samples containing the relevant features (e.g. electron/muon number, transverse mass, R-jet data ... etc) to identify the events, which had good photons and the events which had photon isolation (this problem required two classifiers) . The models were then used in parallel to identify the events, which contained the decay of the Kaluza Klein graviton (both good photons and photon isolation). The events were then plotted on a histogram (Fig. 2.3.1) and compared with the original histogram (Fig. 2.3.2) to see the performance of the classification. The AMS values were 866.8 for the good photon classification and 866.7 for the photon isolation classification.

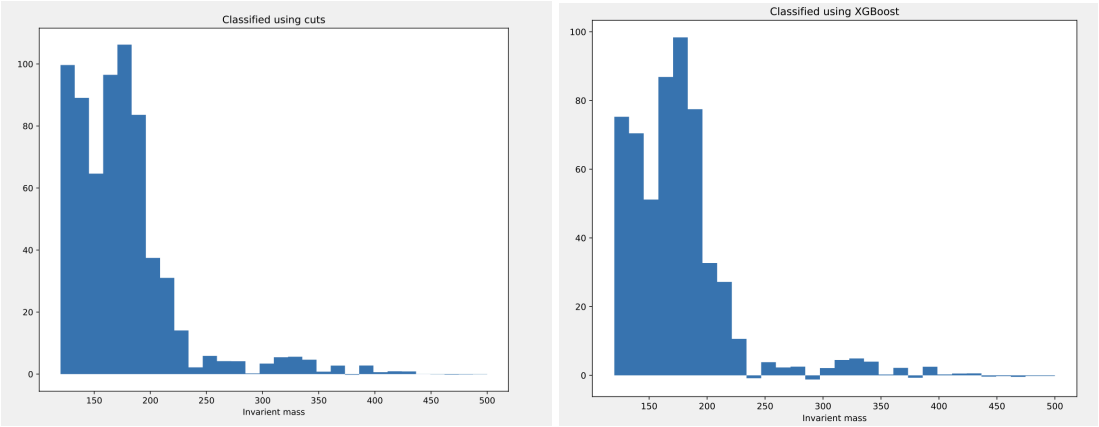


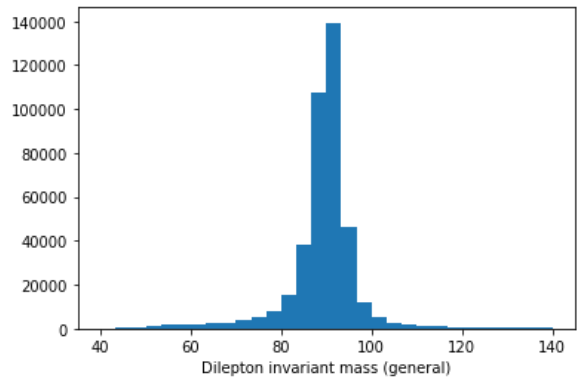
Figure 2.3.1 (left): Kaluza Klein graviton events classified using restrictive cuts
Figure 2.3.2 (right): Kaluza Klein graviton events classified using the XGBoost algorithm

2.4 Supersymmetry Events Classification

Supersymmetry is the idea that every fermion has a partner boson with different spin properties, where fermions have half-integer spin values and bosons have integer spin values. (7) We can search for supersymmetric particles by examining pairs of particles created from collisions in CERN. To do this, we can use Python to examine ATLAS open data and make ‘cuts’ on it. These cuts essentially filter out data that we do not need, leaving us with a subset of data that is much more useful to examine supersymmetric events. We made these cuts by first using uproot to load the ATLAS open data. We then created the framework for our histograms to be plotted later. Following this, we selected all of the information we needed to make cuts from the data and stored it. Then, we set up variables

to store the momentum of particles by creating four-vectors using their Lorentz factor and velocity. A series of cuts were made, starting with including only collisions between electrons and muons in pairs of the same type and opposite charge. We then selected events where each particle had a minimum momentum, which we decided by following advice from the ATLAS experiment at CERN. Following this, we calculated the momentum of leading and trailing leptons, and summed them by their vectors to find the dilepton invariant mass. We then made some more cuts based on this mass and the accuracy of detected jets. Now, we categorised these leptons into categories based on the magnitude of their invariant mass and MT2 variables (stransverse mass). Finally, we plotted the occurrence of these leptons as they varied at different dilepton invariant mass values for each histogram, with general (least strict), loose and tight requirements depending on dilepton invariant mass and MT2 values.

After obtaining these results, we stored them in a CSV. We then used the XGBoost machine learning algorithm and trained it on half of our data, testing it on the other half to see how well it predicted our results. To find the accuracy of our algorithm, rather than simply calculate the proportion of predictions that were 'correct' or within an acceptable range, we used the AMS metric - we did this by defining a function that implements the AMS metric and then calling it for each category (general, loose, tight). The graphs plotted by the XGBoost algorithm are below, next to our resultant histograms. The algorithm produced graphs and AMS values ~1.689 (loose) and ~1.192 (tight), while the general category had an AMS value of ~ 603.226. This may have been a result of using a smaller dataset as this would've resulted in a weaker model. The comparison of the graphs shows us that our loose events classification predicted by the XGBoost algorithm (Fig. 2.4.4) is most similar to the loose events classification made using cuts (Fig. 2.4.3), sharing a shape with the tight classification graphs (Fig. 2.4.5, 2.4.6). Hence, loose cut requirements are the best for building an accurate model to detect supersymmetric particles although they may lead to more false positives than desirable when compared to tight requirements.



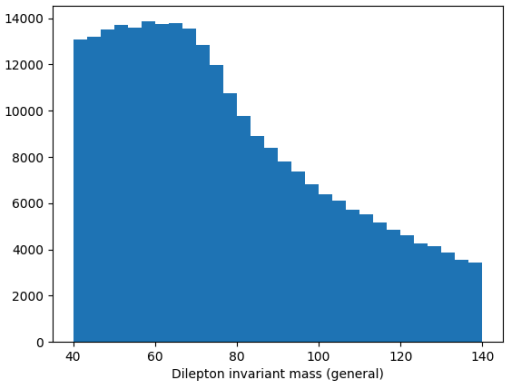
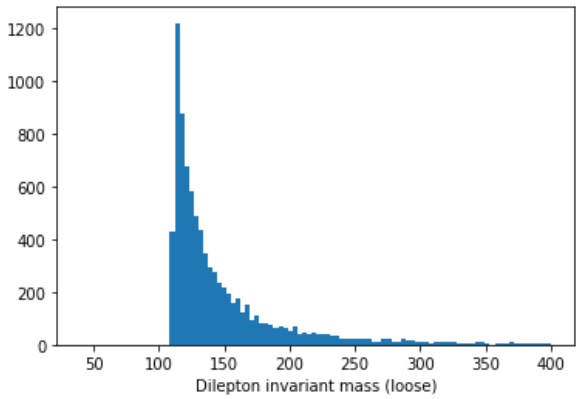


Figure 2.4.1 (left): SUSY events general classification using restrictive cuts
Figure 2.4.2 (right): SUSY events general classification using the XGBoost algorithm



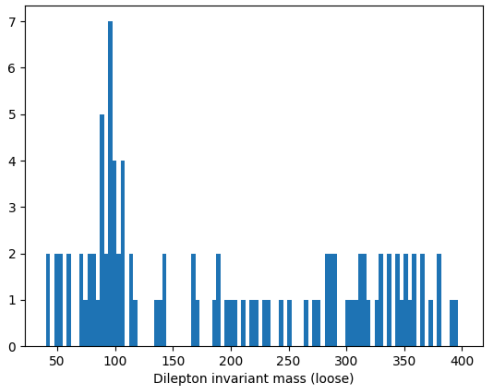
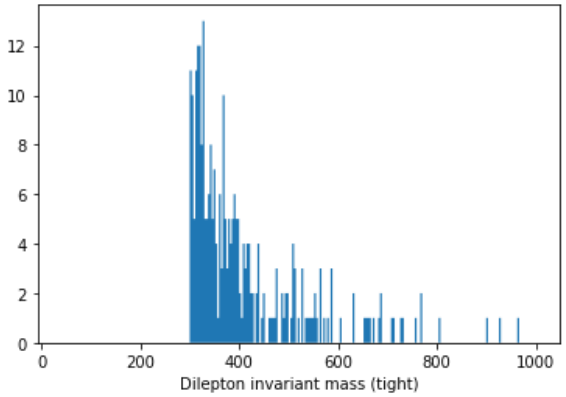


Figure 2.4.3 (left): SUSY events loose classification using restrictive cuts
Figure 2.4.4 (right): SUSY events loose classification using the XGBoost algorithm



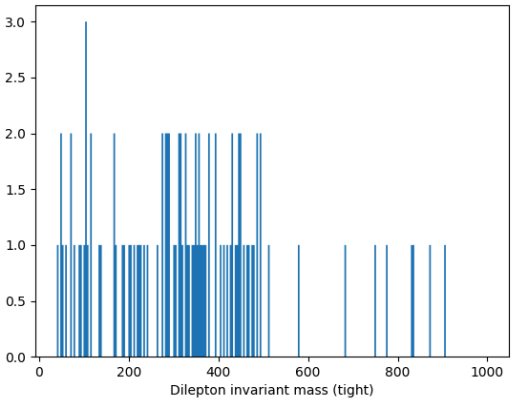


Figure 2.4.5 (left): SUSY events tight classification using restrictive cuts
Figure 2.4.6 (right): SUSY events tight classification using the XGBoost algorithm

3. Discussion, Conclusion and Future Work

From BSM particles to delving into supersymmetry we have explored the performances of the XGBoost machine learning algorithm across a wide range of groundbreaking classification problems in particle physics.

Our results revealed significant variations in the XGBoost algorithm's performance across our tested range of classifications in particle physics. In some areas, such as the Higgs boson classification, the XGBoost gave perfect or near-perfect prediction results. However, in other cases, it was clear that the XGBoost displayed excessive rigidity, leading substantial portions of the signal data being excluded and dismissed as background data, which resulted in thin meagre graphs, as seen in our exploration of SUSY.

The XGBoost is a supervised learning algorithm, and so relies on labelled datasets. Therefore, the algorithm works best in classifications where the selection criteria is well defined, as it would allow accurate labelled training datasets to be generated. In these cases, researchers may not feel the full benefit of the XGBoost, as the algorithm will only ever (at best) replicate the predetermined cuts. However, this is an issue that we hope to

address in our future work by exploring the potential of deep learning models to identify the selection criteria for particle decay classifications, which have very few selection criteria identified so far.

References

(1) Brownlee, J. A “Gentle Introduction to XGBoost for Applied Machine Learning” *Machine Learning Mastery.* https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/ (2021)

(2) Adam-Bourdarios, C. *et al*. “Learning to discover: the Higgs boson machine learning challenge”. pp. 9-10, https://higgsml.lal.in2p3.fr/files/2014/04/documentation_v1.8.pdf (2014)

(3) CERN. “What's so special about the Higgs boson?” https://home.cern/science/physics/higgs-boson/what (2018)

(4) ATLAS collaboration (2014). Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014. *CERN Open Data Portal.* http://opendata.cern.ch/record/328# (2014)

(5) Wulf, E. “Search for Randall-Sundrum Gravitons at the LHC: Recent Results from ATLAS” https://indico.cern.ch/event/129980/contributions/1350939/attachments/90398/129356/evan_wulf_RS_diphoton_dpf.pdf (2011)

(6) CERN. “Extra dimensions, gravitons, and tiny black holes”. https://home.cern/science/physics/extra-dimensions-gravitons-and-tiny-black-holes (2018)

(7) CERN. “Supersymmetry”. https://home.cern/science/physics/supersymmetry (2018)

Figures