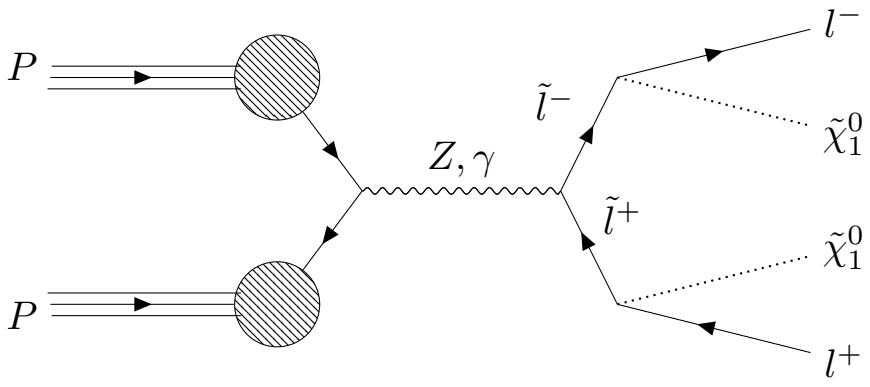
FYS5555 Project 3

Searching for two direct sleptons in a dilepton final states using supervised machine learning

William Hirst

Spring 2022



University of Oslo

In this rapport I compare two different approaches in the search for direct sleptons in the ATLAS 13TeV $10fb^{-1}$ data. The first approach uses a rectangle cut only approach, the second uses a mixture of rectangle cuts and machine learning (XGBoost). I found no excess of events beyond the prediction of the standard model using the first or second approach. Using the rectangle approach no combinations of slepton and neutralino masses were excluded. Using the rectangle/machine learning approach I was able to exclude the slepton mass of 100.5GeV and neutralino mass of 1GeV up to a 95% confidence, though with imprecise uncertainty. In the comparison I found that although the rectangle only approach cut more than 90% of all background, it was not able to preserve enough signal to produce any conclusive results. By applying the XGBoost classifier to the analysis I was able to seperate signal from background, therby creating a far more complex and accurate signal region.

1 Introduction

The standard model (SM) is perhaps one of the most successful scientific theories ever created. It accurately explains the interactions of leptons and quarks as well as the force carrying particles which mediate said interactions. In 2012 the SM achieved one of its crowning achievements when we discovered the Higgs boson. Much of the accolade was rightfully given to the theoretical work on the SM, but another aspect of the discovery was equally important. Data analysis was and is a crucial part of any new discovery in physics. One of the most important and exiting tools is machine learning.

When on the search for beyond standard model (BSM)-physics, one aims to search in isolated regions where one suspects to find it. The traditional approach is to define regions by preforming rectangle cuts on event variables. In the case that the BSM is significantly different from the SM, this can be quite successfully. But, in the case that the two are relatively similar, rectangle cuts will not suffice. Machine learning aims to apply cuts that in practice are not (or at least do not seem) rectangle, but are curved. One often speaks about the machine learning creating a complex, curved region in the phase space spanned by the events.

In this rapport I will use a mixed approach of rectangle cuts and machine learning and compare it to a strict rectangle cut approach in their ability to search for new physics in a dilepton final state using the ATLAS 13TeV $10fb^{-1}$ data [1]. The new-physics which will be the focus of this rapport will be Super Symmetry[1]. The methods compared in this raport will be tested on their ability to provide new information about the real data, either it be discovery or exclusion.

This rapport is structured in 4 sections; theory, implementation, results & discussion and conclusion. In the theory section I will discuss the basis strategy of the search, Super symmetry and why it is worth the search, the different background-channels which will contribute to the analysis, the machine learning algorithm of choice and the statistical framework used in analysing the results. In the implementation section I will further explain the different methods used in the analysis, how I have chosen to handle the data as well as the steps involved in applying the machine learning framework. The results and discussion section will present all cuts used in defining different regions, the results from training

[1] More on the motivation behind super symmetry in section 2.2

the classifier, the results from applying each region, the final results from each search and discuss possible ways forward. Finally all results and discussion, will be summarised in the conclusion section.

2 Theory

In this section I will discuss the necessary theory for both the physical analysis (supersymmetry, statistical tools ect.) as well as the tools need for the ML analysis. Note that the sections on XGBoost-classifier and boosted decision trees is inspired from my previous work in the articles [2] and [3].

2.1 Basic search strategy

In most attempts to search for BSM-physics the basic idea is to choose a new model of interest and compare Monte Carlo data (MC) with with real data in designated regions we call signal regions (SR). To define these SR we use MC. As will be made clearer in later section, we want to define these signal regions as a region in the feature space with the most amount of predicted MC of the BSM physics (signal) and the least amount of SM MC (background). Defining these regions can sometimes be done through classical rectangle cuts on certain features. This means manually finding thresholds for different variables and excluding all event either below or above this threshold. Using machine learning allows us to define more sophisticated cuts based on trends in the features space that might not visible to the human eye.

2.2 Super symmetry

A supersymmetric theory has for many years been an interesting candidate for a new physics beyond the standard model. Supersymmetry (SUSY) aims to alter the standard model such that mass and force is treated equally. Supersymmetry suggests that each SM particle has an additional superpartner(SP) which we call a sparticle. The sparticles all differ with half a spin from the original SM particle. Because SUSY is a broken symmetry, sparticles are expected to be much higher than the corresponding SM masses, often in the range of 100-1000GeV. The difference in spin means that the SP of a fermion is a boson and the SP of a boson is a fermion. SUSY is a candidate to fix many problems in physics, some of which are; the hierarchy problem, fixing the mass of the higgs and possibly the mystery of dark matter.
In this analysis I have chosen to search for supersymmetry through the signal of two direct sleptons decaying to a lepton and a neutralino. The process closely resembles the Drell-Yan production, but with additional neutralinos. The direct slepton-diagram is shown in figure 1. In the rapport only the sleptons $\tilde{e}$ and $\tilde{\mu}$ (SP of e and μ) will be considered.

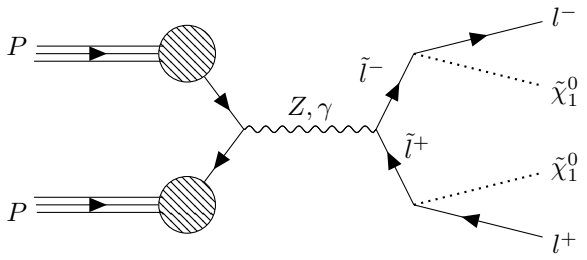


Figure 1: The Feynman diagram for the direct slepton signal channel.

The mass of the slepton and neutralino are one of the many unknowns surrounding the sparticles. Therefore, to simulate the existence of sleptons and neutrlainos in the MC, we must make educated guesses. In this rapport we will study many possibilities, each possibilities having a different combination of slepton and neutralino mass. The masses of the slepton ranges from 100GeV to 700GeV and for the neutralinos the masses range from 1GeV to 300GeV.

2.3 Standard model background

In the analysis of direct slepton signal there are several irreducible background channels. These channels are; diboson-, top- and boson+jets-channels. There are two additional channels included in this analysis but will not contribute an equal amount. These channels will be referred to as the "others" category. The category consist of the higgs- and topX-channels.

2.3.1 Dibson (WW, WZ, ZZ)

Based on different articles published by ATLAS (for example the ATLAS paper from 2020, [4]) the diboson channels seem to be the largest source of background events. This is partly due to the variation of diagrams, but most importantly the resemblance to the signal diagram. Three of the most important diboson channels included in the analysis are; WW, WZ and ZZ.
In figure 2 I have drawn a Feynman diagram of the WW channel. The figure shows a two quark-annihilation forming a neutral boson, which decays into a pair of W-bosons. The W-boson later decay into a pair of charged and neutral leptons.

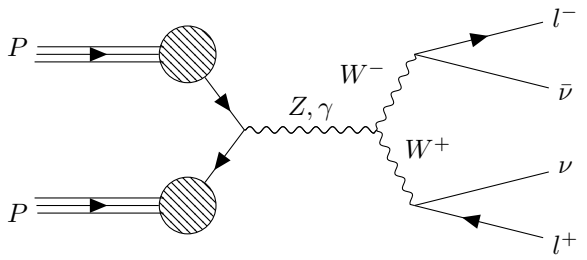


Figure 2: The Feynman diagram for the background diboson channel WW.

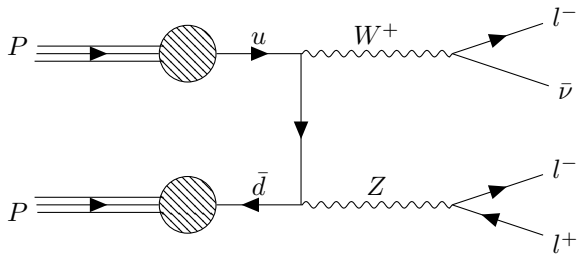


Figure 3: Feynman diagram for the background diboson channel WZ.

In figure 3 I have drawn a Feynman diagram of WZ. The figure shows an up quark coupled to a W-boson and a down quark, and an anti-down quark coupled to the aforementioned down quark and a Z-boson. The W-boson decays into a charged and neutral lepton pair and the Z can decay into a pair of quarks or leptons. In figure 4 I have drawn the Feynman diagram of the ZZ-channel. The figure shows a diagram resembling that of the WZ-channel but with two Z-bosons. In this case the most relevant process for this analysis is the on where one of the Z-bosons decays into a pair of charged leptons and the other decays into a pair of neutral leptons.

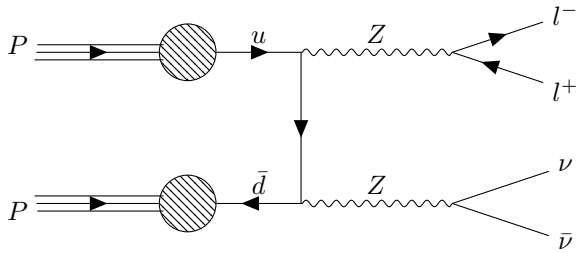


Figure 4: Feynman diagram for the background diboson channel ZZ.

2.3.2 $t\bar{t}$

In figure 5 I have drawn the Feynman diagram of the $t\bar{t}$-channel. Contrary to the diboson-channels, the proton-proton collision of the $t\bar{t}$-channel is initially mediated through gluons. The pair of gluons together form a $t\bar{t}$-pair. The $t\bar{t}$ decays into a jet-pair and two W-bosons. The W-bosons will each decay into a charged-neutral lepton-pair.

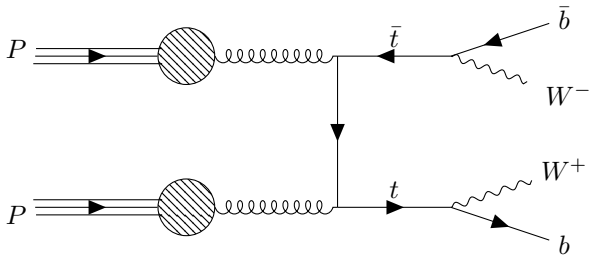


Figure 5: The Feynman diagram for $t\bar{t}$-channel.

2.3.3 Single-top (Wt)

In figure 6 I have drawn the Feynman diagram of a single-top-channel. The figure shows a bottom quark and a gluon that form a W-boson and a top-quark through another quark. The W-boson decays into a charged-neutral lepton pair and the top-quark will decay similarly to the top quarks discussed in the diagram for $t\bar{t}$.

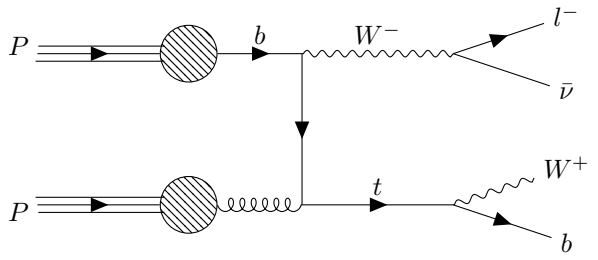


Figure 6: The Feynman diagram for singletop-channel.

2.3.4 W/Z+Jets

In figure 7 I have drawn one example of Feynman diagrams for both the W- and Z+jets channel. The diagram shows two fermions that create a gluon and a Z- or W-boson through the t-channel. The gluon decays into a jet and the W- or Z-boson decays into a pair of leptons, either one or two charged. As is evident from the work made by ATLAS in articles like their work in 2020 [4], normally in searches for direct sleptons, these channels can be easily cut. But, as will be discussed in later sections, the lack of certain features in the data will make the channels contribute an abnormal amount.

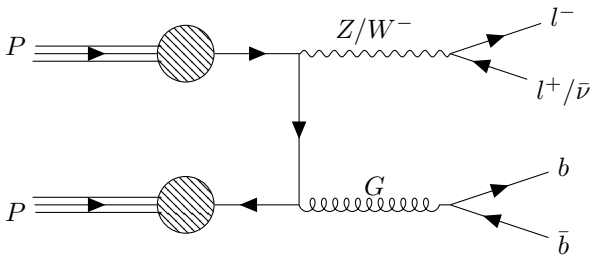


Figure 7: Feynman diagram for the background channels W+jets and Z+jets.

2.4 Gradient-boosted trees [3]

In this rapport I will use the XGBoost-classifier which uses gradient-boosted trees. Gradient-boosting is a machine learning algorithm which uses a collective of "weak" classifiers in order to create one strong classifier. In the case of gradient-boosted trees the weak classifiers are a collective of shallow trees, which combine to form a classifiers that allows for deeper learning. As is the case for most gradient-boosting techniques, the collecting of weak classifiers is an iterative process.
We define an imperfect model $\mathcal{F}_m$, which is a collective of m number of weak classifiers, estimators. A prediction for the model on a given data-points, x_i is defined as $\mathcal{F}_m(x_i)$, and the observed value for the aforementioned data is defined as y_i. The goal of the iterative process is to minimize some cost-function $\mathcal{C}$ by introducing a new estimator h_m to compensate for any error, $\mathcal{C}(\mathcal{F}_m(x_i), y_i)$. In other words we define the new estimator as:

$$\tilde{\mathcal{C}}(\mathcal{F}_m(x_i), y_i) = h_m(x_i), \tag{1}$$

where we define $\tilde{\mathcal{C}}$ as some relation defined between the observed and predicted values such that when added to the initial prediction we minimize $\mathcal{C}$.
Using our new estimator h_m, we can now define a new model as

$$\mathcal{F}_{m+1}(x_i) = \mathcal{F}_m + h_m(x_i). \tag{2}$$

The XGBoost [5] framework used in this analysis enables a gradient-boosted algorithm, and was initially created for the Higgs ML challenge. Since the challenge, XGBoost has become a favorite for many in the ML community and has later won many other ML challenges. XGBoost often outperforms ordinary decision trees, but what is gains in results it looses in interpretability. A single tree can easily be analysed and dissected, but when the number of trees increases this becomes harder.

2.5 Discovery and exclusion

As mentioned in previous sections, the main idea of the search is to define a region where we will compare the MC-background to the data and analyse any differences. Any analysis in the signal region hopes to either discovery or exclude certain BSM physics. The usual way of doing so is through statistics. In this rapport the statistical analysis will not be the focus and will therefore not be explained in heavy detail. Nonetheless there are certain

expressions and terms that must be defined.
For the sake of a discovery one would hope to see an excess in the data, measure its significance and find the model best fitting to the excess. A significance of a discovery is measured by the relation between the background and the apparent signal. Given a signal region with n_{obs} events of real data and b events of MC background, we can define the signal as $s = n_{obs} - b$. Then a significance of any signal, Z is defined as

$$Z = \frac{n_{obs} - b}{\sqrt{b}}, \tag{3}$$

or

$$Z = \frac{s}{\sqrt{b}}. \tag{4}$$

Introducing the possibility of uncertainty in the amount of background, δb we can further define

$$Z = \frac{s}{\sqrt{b + (\delta b)^2}}. \tag{5}$$

In physics one often deem the results a discovery if Z is larger than 5. Given a significance with no indication of discovery, we move over to the intent of exclusion.
In exclusion we are interested in studying the probability of a model which predicts a number of signal events s, given n_{obs} real data events. To do so we will use Bayesian statistics.
The probability of model predicting s signal events in the signal region, given n_{obs} event of real data is in bayesian statistics written as

$$P(s|n_{obs}) = \frac{P(n_{obs}|s)P(s)}{P(n_{obs})} \tag{6}$$

Approximating the probability distribution with a Poisson distribution[2] we can define this probability as

$$P(n_{obs}|s) = L(n_{obs}, s) = (s + b)^{n_{obs}} \frac{e^{-(s+b)}}{n_{obs}!}, \tag{7}$$

where b is the predicted background, and L is simply the likelihood.
From this point we can define s_{up} as being the upper limit of any signal, where any model that predicts $s > s_{up}$ is excluded. We define s_{up} as

$$\int_0^{s_{up}} P(s|n_{obs}) = 0.95, \tag{8}$$

meaning that any $s > s_{up}$ is excluded within a 95% confidence. If we assume flat prior both $P(n_{obs})$ and $P(s)$ are independent of s, and will act as normalization constants. Therefore we can calculate s_{up} using equation 6.
To further extend this approach for HEP purposes we make the probability dependant of 4 parameters, the signal cross section σ, the background cross section σ_{bkg}, the integrated

[2]The significance is defined in the limit where the Poisson approximates a Gaussian distribution.

luminosity L_{int} (which for the sake of this rapport is $10fb^{-1}$) and the signal efficiency ϵ. The efficiency in this case is equal to the fraction of signal before and after the different region requirements were applied.
We can define

$$s + b = L_{int}(\epsilon\sigma + \sigma_{bkg}). \tag{9}$$

Using equation 6, 7, 9 and marginalizing for all parameters but the signal cross section, we can find a new expression for our original probability as

$$P(\sigma|n_{obs}) = N \int L\left(\sigma, L_{\text{int}}, \varepsilon, \sigma_{\text{bkg}}\right) p\left(L_{\text{int}}\right) p\left(\varepsilon\right) p\left(\sigma_{\text{bkg}}\right) dL_{\text{int}} d\varepsilon_{\text{sig}} d\sigma_{\text{bkg}}, \tag{10}$$

where N is a normalization constant. The probability is now dependant on the signal cross section. To calculate expressions as the ones above and to find s_{up} Monte carlo integration must be used. In this rapport I used the statistical framework created by Magnar Kopangen Bugge in the statisticsNotebook to calculate all limits.

3 Implementation

The cuts in the analysis are all chosen by me, although some are inspired by the work of Eirik Gramstad in his thesis from 2013 [6].

3.1 Cuts and signal regions

In this rapport I will define 2 signal regions. Both signal regions will have a common base, which I will call the analysis region (AR). The AR will be defined making relatively loose cuts on the variables of the MC. This region will aim to remove unnecessary background processes and at the same time keep as much of the signal as possible, around $20 - 60\%$. The hope of these cuts is to leave only signal and irreducible background events left.
The first signal region (SR1), will use AR as a base and add an additional cut from the ML output. This region will be used in the statistical analysis for any results produced from our ML-model. The second signal region (SR2) will also use AR as a base, but will add additional cuts on the variables. These cuts will be much stricter to remove as much irreducible background as possible, and will as a consequence remove a larger amount of signal than AR. Finally we will compare the two signal regions.

3.2 Data handling

For training and testing of the XGBoost classifier I created three data sets; one training, one validation and one test. All the data sets are based on the AR (1). The training set consisted of 80% of the MC, saving the remaining 20% for validation. Given the large amount of background, the data (both training and validation) was scaled such that the sum of the weights were equal for signal and background. This was done to ensure a realistic representation of the background data as well as making sure that the classifier is given a sufficient amount of signal. The final data set (testing) consisted of all the background data and is used to visualize the classifiers ability to act on all the SM MC.

Additional to the imbalanced data set, there was a problem with the weights produced by the MC-simulations, some were negative. The negative weight problem is a well known problem in the world of ML-analysis for HEP, and is a consequence of higher perturbative accuracy. For the purpose of visualising distributions, negative weights are not a problem. But, when using the weights in the training of the classifier, the XGBoost framework will not work.
How one chooses to deal with this problem can heavily effect results and many solution are suggested as a consequence. Given the time limitations of this rapport, the simplest solution was chosen. Namely, I chose to normalize all the weights as a whole, conserving the total sum of the weights and at the same time changing all negative signs. The simple procedure is shown bellow in equation 11,

$$w_i = P \mid w_i \mid \tag{11}$$

$$P = \sum_{j=0}^{N-1} \frac{w_j}{\mid w_j \mid}, \tag{12}$$

where w_i is the weight of an event i, $i \in [0, N-1]$ and N equals the total number of events.

3.3 Features

The data set used in training the classifier consisted of 34 features. Most of the features were taken directly from the data and are presented in table 4 page 34 in the ATLAS review of the open data from 2020 [1]. All features surrounding the leptons (branch names beginning with lep) are included in the data set as well as missin energy (branch names beginning with met). Additionally the dilepton mass was included as well as the stransverse mass. The stransverse mass was calculated using a MT2 calculater develpoed by Christopher G. Lester and Benjamin Nachman [7].
Features surrounding the jets were kept to a minimum. The only feautres regarding the jets inform of the the total number of jets above a certain transverse momentum threshold (20 and 60GeV), as well as the number of b-jets as classified by a ML-algorithm up to different accuracy's.

3.4 The XGBoost-classifier

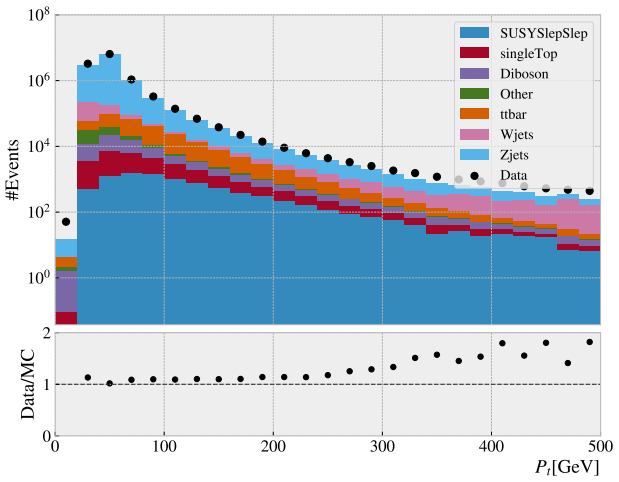
For the case of this rapport, no gridsearch was used in finding hyper-parameters for the XGBoost-classifier. This was a conscious decision based on two reason. Firstly the data set used in this analysis is very large. Any gridsearch that would have any hope of finding optimal parameters would be very time consuming. Secondly the algorithm behind boosting, and especially in the XGBoost framework often means that the parameters in the classifier are not as crucial as in other methods such as deep networks. It is an unproven phenomenon that the XGBoost-classifier often prefers its default settings (see [5] for full description). Therefore most of these parameters were used in the rapport with a couple exceptions.
Some of the parameters were changed to either make the classifier faster, or based on previous experience i similar analysis. Firstly the learning rate was changed to 0.1. Secondly the *max_depth* of the classifier was changed to 3. Finally the weights of the classifier was set to the sum of the weights of signal divided by the sum of the weights of background. This

was done as a suggestion by the XGBoost-contributors. Given the scaling of the weights this factor will be 1.

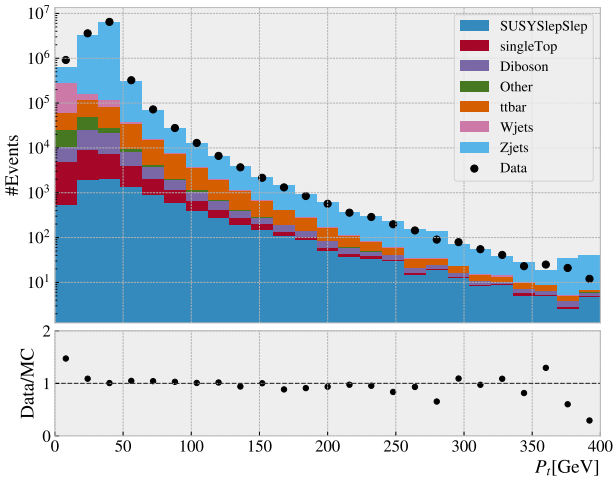
4 Results & Discussion

4.1 Monte Carlo and Data

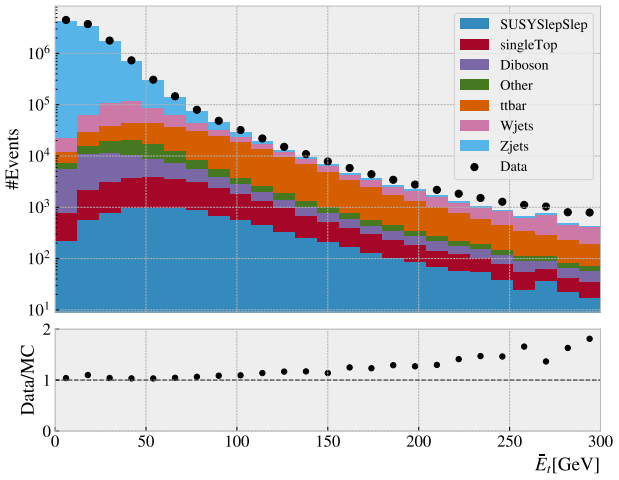
Given the main idea of the search is to compare MC to real data, one must first make sure that the MC in-fact resembles the real data. To do so I have plotted the distribution of 9 variables. In addition I plot the ratio of real data divided by MC for each bin. For the purpose of justifying the cuts made for AR and SR1 the contribution of SUSY signal is included in the distribution plot, but given the small amount of signal it will not effect the comparison. However the SUSY signal is not included in the ratio plot.
In figure 8 I have plotted the transverse momentum for the first 8a and second lepton 8b, missing transverse energy 8c, dilepton mass 8d and the stransverse momentum 8e. In all the aforementioned figures we observe a clear similarity between the MC and real data. This is especially evident by studying the ratio-plots under each plot. Most of the ratios are located around 1, more so even for the bins with the highest number of events. From the ratio plots we also observe that the largest difference between MC and data happens for events with larger energy's/momentum.
In figure 9 we plot the distribution of the nr of B-jets 9a, the number of jets with transverse momentum larger than 20GeV 9b and 60GeV 9c and the number of events with same or opposite sign 9d. From all the plots in figure 9, we observe the same as for figure 8. In other words we observe a clear similarity between the MC and real data. From our study of the distribution of different variables we can conclude that the MC adequately resembles the real data.



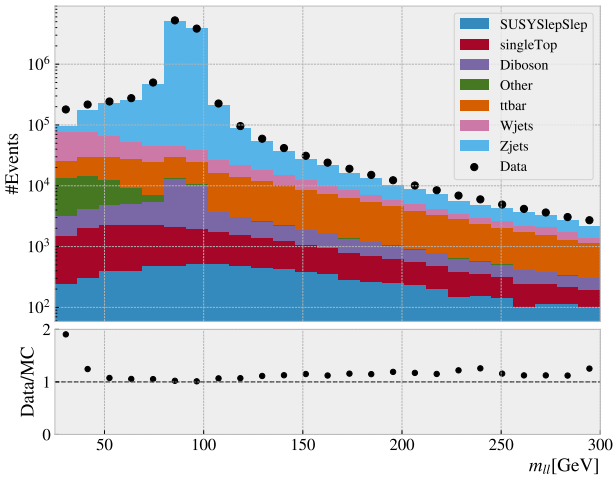
(a)



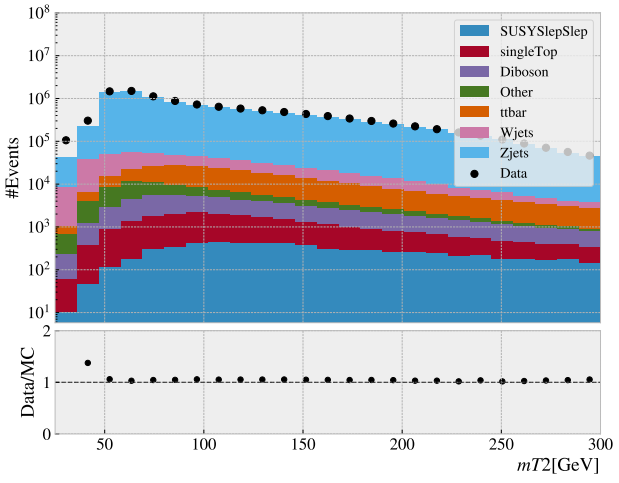
(b)



(c)

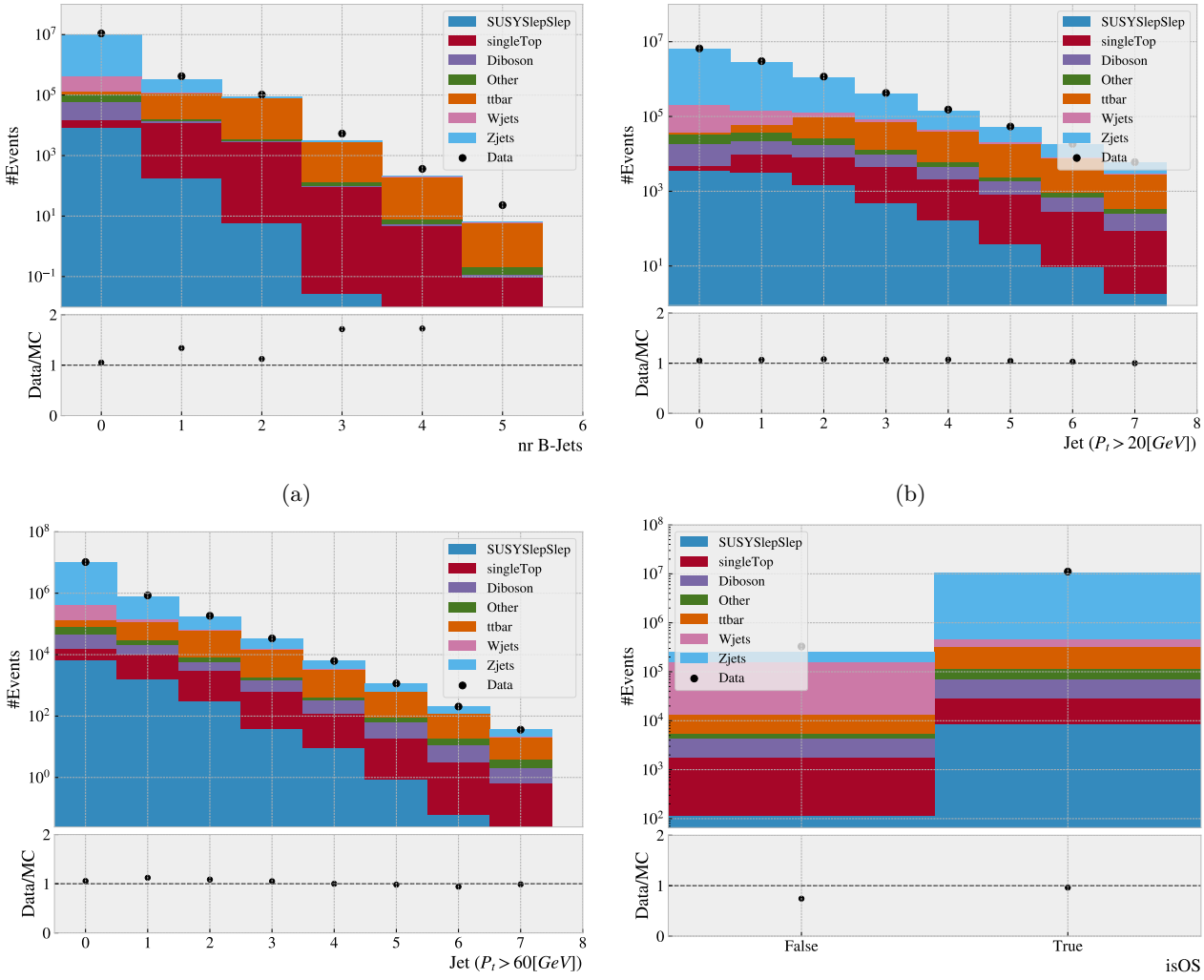


(d)



(e)

Figure 8: The distribution of transverse momentum of first 8a and second 8b lepton, missing transverse energy 8c, dilepton mass 8d and the stransverse mass 8e for the MC and real data. Additionally the ratio of data over MC is plotted under each distribution histogram.



(a) (b)

(c) (d)

Figure 9: The distribution of nr of B-jets (efficiency 80%) 9a, jets with a transverse momentum above 20[GeV] 9b and 60[GeV] 9c and the number of lepton pairs with opposite or equal charge 9d for the MC and real data. Additionally the ratio of data over MC is plotted under each distribution histogram.

4.2 Cuts for AR

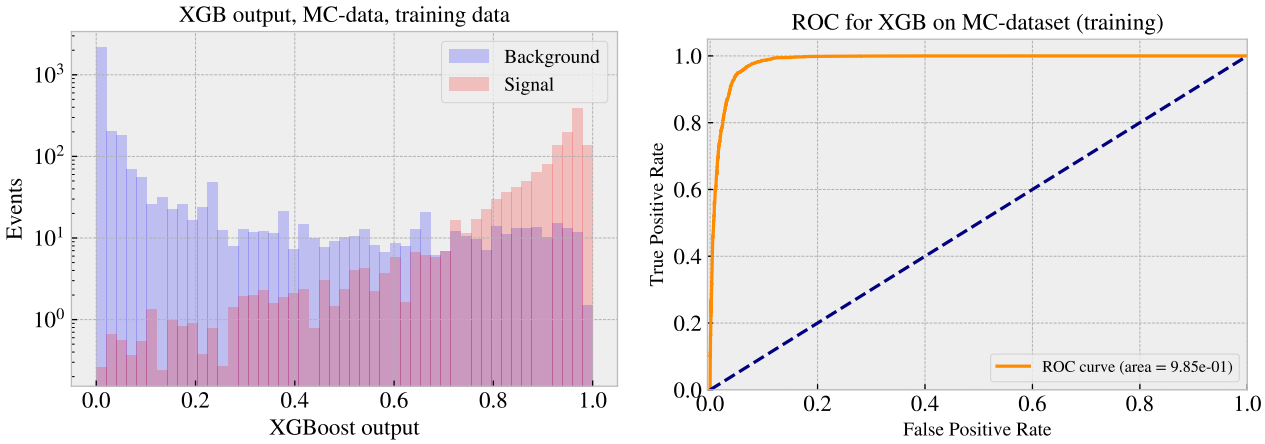
In this section I will present the cuts made to define the AR as well as the result from applying them to the full data set. By studying the distribution plots in figure 8 and 9, I aimed to remove as much of the background by making cuts before any of the peaks for the SUSY signal. By doing so I defined the region as presented by the cuts in table 1. Note that the cut applied on the missing transverse energy is the only case were a upper-limit is defined. This was done as an attempt to remove regions were the MC and real data differ in a unsystematic way.
By applying the cuts for the AR a great amount of the data was removed. The resulting reduction on each channel is presented in table 2. From the table we can see that Z+jets and W+jets was effected the most, removing more than 99.99% from both channels. Dibson and ttbar were reduced 96% and 94% respectively. As expected these channels are harder to remove given their similarity to the signal. The 'others'-channel was like the Z/W+jets channel almost completely removed. The SUSY-signal was however only reduced by 74% making it the second largest channel in the AR, only beaten by the W+jets channel.

Table 1: Feature cuts in AR

Features	Thresholds
$\bar{E}_t$	$> 50,\ < 250[GeV]$
P_t (1)	$> 80[GeV]$
P_t (2)	$> 25[GeV]$
m_{ll}	$> 120[GeV]$
m_{T2}	$> 150[GeV]$
$Nr\ B-Jet$	< 2
$isOS$	$True$
$isSS$	$False$

Table 2: Events analysis of MC in total and in AR

Channel	**Nr Event total**	**Nr Events AR**	**Reduction [%]**
$Diboson$	4.5×10^4	1.5×10^3	96%
$ttbar$	2.1×10^5	1.1×10^4	94%
$Z + jets$	1.0×10^7	1.9×10^3	99%
$W + jets$	1.0×10^7	2.7×10^3	99%
$Others$	3.7×10^5	6.2×10^1	99%
$SUSYSlepSlep$	8.8×10^3	2.3×10^3	74%

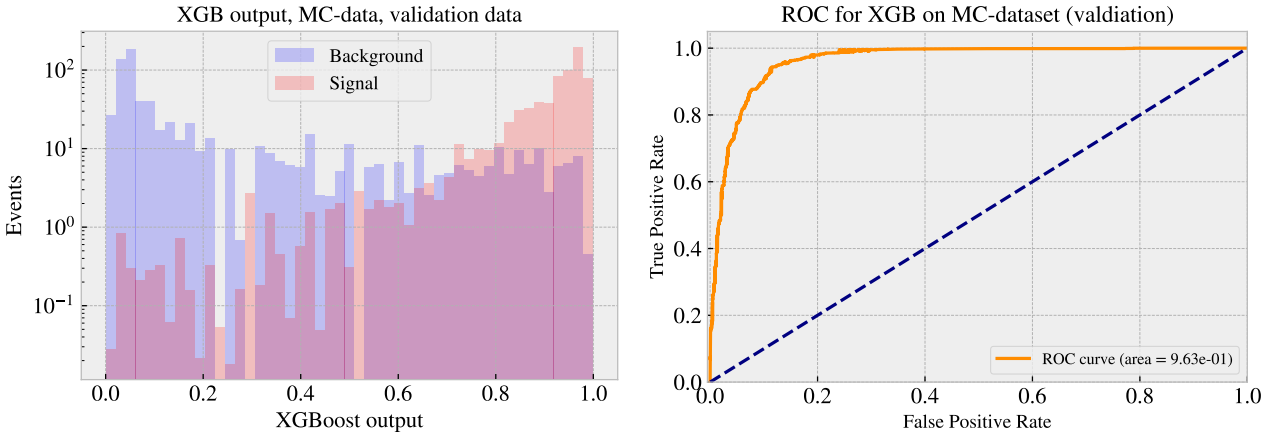


(a) (b)

Figure 10: The output distribution 10a and the ROC-curve 10b produced by the XGBoost-classifier on training data.

4.3 Training and validation

As mentioned in section 3.2, all training, validation and testing was done in in the AR. Firstly I preformed a fit of the classifier on the training set, containing 80% of the signal and background. After training I plotted the distribution as well as the ROC-curve of the output from the classifier from the training data. The results are shown in figure 10a and 10b. Figure 10a shows a clear separation between the background and signal. This is evident from the ROC-curve which creates an area of 0.985. The distribution plot shows that most of the background is located in the lowest output-bin. This again confirms the classifiers ability to accurately classify background and signal.
After training and analysing the reuslts on the training data, I create similar plots for the validation data. The distribution and ROC-curve is displayed in figures 11a and 11b. In figure 11a we see a clear separation between the signal and background, although not as prominent as in the training data. From figure 11b we see that the classifier achieves an excellent ROC-curve area of 0.963. The optimal threshold should be placed in the area where the ratio of true positive divided by false positive is highest (i.e closest to (0,1) in the plot). From figure 11b we can observe that this is in the region of [0.85, 0.95]. It is clear from the validation data that the XGBoost has picked up differences in trend for the signal and background and can be used as a tool in the hope to create an effective signal region.
Finally we plot the distribution of all the background data in figure 12. The figure displays the classifiers ability to separate more than 90% of the MC background to the lowest values of the output.



(a) (b)

Figure 11: The output distribution 11a and the ROC-curve 11b produced by the XGBoost-classifier on validation data.

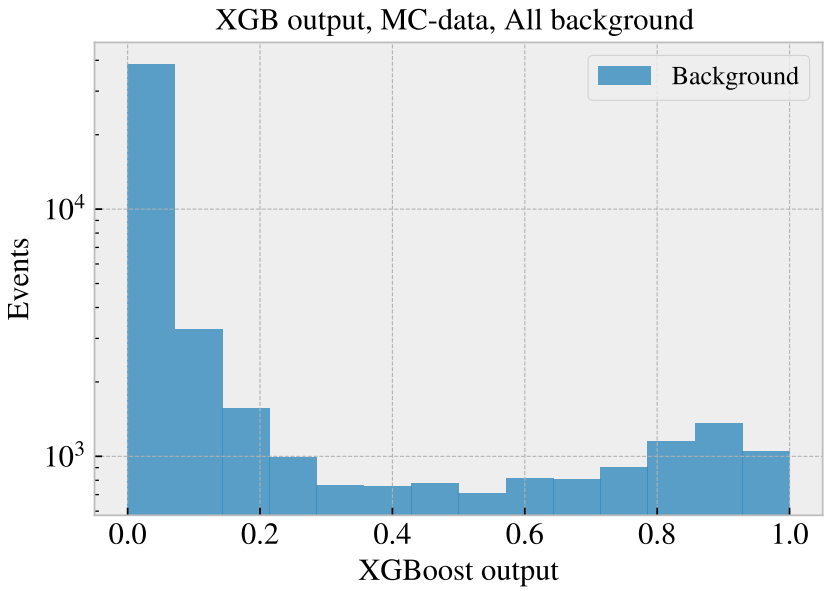
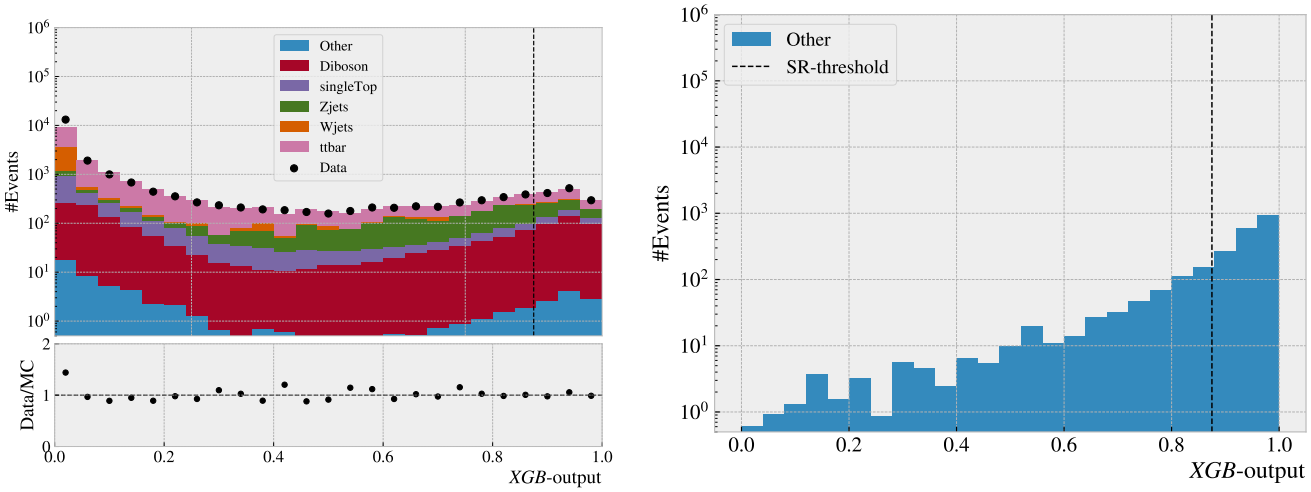


Figure 12: The output distribution of the XGBoost on all background MC data.



(a) (b)

Figure 13: XGBoost output distribution for MC and data 13a and the SUSYSlepSlep-MC 13b in SR1.

4.4 Cuts for SR1

After training the XGBoost-classifier on the MC we can apply it on the real data and compare. In figure 13a I have plotted the distribution of the XGBoost output for both the MC background and the real data. Under the figure we see the ratio-plot between the two in each bin. From the distribution and ratio plot we see a clear similarity between the output for the MC and real data. Given the goal of comparing the two in different region, it is vital that the XGboost treats both similarly.
In the distribution plot in figure 13a we also observe that most of the data is centered around the lower values of the output. Comparing to the distribution plot of the MC-signal in figure 13b, we can see that the XGBoost classier has accurately separated most of the background from the signal. Note also that the MC-background in these plots are from the AR, meaning the background data in the plots are the events most similar to the signal, making the results of the XGBoost that much more impressive.
Finally in figure 13a, I plotted a dotted line representing the cut of the ML-output, defining the SR1. After some testing I found that the optimal threshold was for output values around 0.875, which is inline with our observations made from the ROC-curve in section 4.3. This means that SR1 is defined as the data satisfying the AR requirements (see table 1), as well as having a XGBoost output above 0.875. Using these requirements for SR1, we get the results as displayed in table 3.

Table 3: Events analsysis of MC in AR and in SR1

Channel	Nr Event AR	Nr Events SR1	Reduction [%]
Diboson	1.5×10^3	3.3×10^2	78%
ttbar	1.1×10^4	4.4×10^2	96%
$Z + jets$	1.9×10^3	3.1×10^2	83%
$W + jets$	2.7×10^3	4.0×10^1	99%
Others	6.2×10^1	1.0×10^1	84%
SUSYSlepSlep	2.3×10^3	1.8×10^3	21%

In table 3, we observe that all background MC was reduced down to bellow 500 events, whereas the MC-signal still contains almost 2000 events, making the signal the largest channel in SR1. By comparison to the cuts made for the AR, the XGboost-classifier was able to cut far more efficiently, again keeping in mind that the data it was given is the background most similar to the signal. Although SR1 only cut 21% of the signal, it still cut more than 75% of all the background, even reducing channels like ttbar and W+jets by 96% and 99%. As was briefly mentioned in section 2.3.4, the W/Z+jets-channels are normally negligible in searches such as these. In articles like the ones from ATLAS in 2020 [4] or the work by Eirik Gramstad in 2013 [6], cuts are normally made on different variables for the jets to reduce W/Z+jets-channels, or even ttbar. This would make the diboson-channel the largest irreducible background. In the data set used for this analysis, there was not enough features for the jets to reduce W/Z+jets- or ttbar-channels to a similar degree. Therefore ttbar and Z+jets effect the signal region more than usual. On the other hand the classifier was able to reduce the W+jets-channel to an almost negligible amount, 40 events.

4.5 Cuts for SR2

Producing any result using the rectangle cut only approach, required a much stricter strategy than was the case for AR. The cuts made in this section were characterized by the attempt to reduce the amount of background in AR and thereby subsequently reducing any relative difference between the MC and real data. The resulting cuts for SR2 are displayed in table 4, which are made in addition to the AR cuts.

As discussed in the analysis of feature distribution 4.1, there seemed to be the most unsystematic deviation between MC and data in the areas with highest energies. This explains that some of the cuts added to the region are upper limits. Although these upper-limits reduce the possibility of fluctuation differences in the data as well as background, one also risk removing a large amount of the higher mass sleptons and neutralinos.

The results from defining the new region SR2 are presented in table 5. The table shows a larger reduction on all channels but the W+jets, compared to the SR1 region. All channels are reduced by more than 90%. The cost of the reduction of background, is the reduction of the signal which is now more than 80%.

Table 4: Feature cuts in SR2

Features	Thresholds
$\bar{E}_t$	$> 75,\ < 150[GeV]$
P_t (1)	$< 200[GeV]$
P_t (2)	$> 60[GeV]$
m_{ll}	$< 300[GeV]$
$Nr\ B-Jet\ (P_t > 20GeV)$	< 2

Table 5: Events analysis of MC in AR and in SR2

Channel	Nr Event AR	Nr Events SR2	Reduction [%]
$Diboson$	1.5×10^3	1.4×10^2	90%
$ttbar$	1.1×10^4	5.8×10^2	94%
$Z + jets$	1.9×10^3	1.0×10^2	94%
$W + jets$	2.7×10^3	1.2×10^2	95%
$Others$	6.2×10^1	5.0×10^0	92%
$SUSYSlepSlep$	2.3×10^3	4.2×10^2	81%

4.6 Discovery or exclusion

Given the structure of the data in the analysis, I was not able to preserve any information surrounding the uncertainty of the data. Instead, I have chosen to add a 5% uncertainty to any MC background data in the signal regions. This uncertainty was used both for the significance and the limits. Given the arbitrary nature of choosing 5%, the results in the analysis will not be absolute.

In table 6 I have written the resulting MC-background and data in both SR1 and SR2, along with the resulting significance. The Significance was calculated using equation 5. From the table we see that both SR1 and SR2 cut a relatively similar amount of events for both MC and data. SR1 achieved a significance of 0.25 whereas SR2 achieved a significance of 1.20, which are both far from 5. Neither of the regions found a deviation between the MC and data worth any further search as far as discovery is concerned.

Given the lack of discovery I aim to exclude certain masses for the SUSY model. I do this by studying each pair of different masses for the slepton and the neutralino, discussed in 2.2. Before I can calculate the limits and check for exclusion, I need to calculate the efficiency of each pair.

In table 7 I displayed the number of signal events in each region for each mass pair, along with the resulting efficiency. Already we see that SR1 has cut far less signal than SR2. In addition we observe that most of events left are in the smaller mass range.

Using the number of events and efficiency I calculated the limits for each off the mass pair.

Table 6: Final result of MC and data in SR1 an SR2

Region	**MC**	**Data**	**Significance (Z)**
$SR1$	1266	1284	0.25
$SR2$	1097	1174	1.20

Table 7: Exclusion of slepton and neutralino massses.

Masses[GeV]	**Events in SR1 [efficiency]**	**Events in SR2 [efficiency]**
$m_{\tilde{l}} = 100$, $m_{\tilde{\chi}} = 50$	554 [0.148]	105 [0.028]
$m_{\tilde{l}} = 100.5$, $m_{\tilde{\chi}} = 1$	1075 [0.260]	239 [0.057]
$m_{\tilde{l}} = 200$, $m_{\tilde{\chi}} = 100$	159 [0.431]	37 [0.100]
$m_{\tilde{l}} = 200.5$, $m_{\tilde{\chi}} = 1$	210 [0.529]	30 [0.075]
$m_{\tilde{l}} = 300$, $m_{\tilde{\chi}} = 200$	33 [0.455]	7 [0.102]
$m_{\tilde{l}} = 300.5$, $m_{\tilde{\chi}} = 1$	37 [0.486]	2 [0.032]
$m_{\tilde{l}} = 400$, $m_{\tilde{\chi}} = 300$	9 [0.450]	2 [0.093]
$m_{\tilde{l}} = 500$, $m_{\tilde{\chi}} = 100$	4 [0.454]	0 [0.000]
$m_{\tilde{l}} = 500$, $m_{\tilde{\chi}} = 100$	2 [0.287]	0 [0.000]
$m_{\tilde{l}} = 500.5$, $m_{\tilde{\chi}} = 1$	2 [0.491]	0 [0.000]
$m_{\tilde{l}} = 600$, $m_{\tilde{\chi}} = 1$	1 [0.322]	0 [0.000]
$m_{\tilde{l}} = 600$, $m_{\tilde{\chi}} = 300$	1 [0.309]	0 [0.000]
$m_{\tilde{l}} = 700$, $m_{\tilde{\chi}} = 300$	0 [0.00]	0 [0.000]
$m_{\tilde{l}} = 700.5$, $m_{\tilde{\chi}} = 1$	0 [0.00]	0 [0.000]

To visualize the reult I created two plots for each of the regions. The plots visualize the limits for each pair both for a 95% and 87% confidence. Any mass pair with a cross section above any of the aforementioned lines, is excluded up to that confidence.
In figure 14 I plotted the result for SR1. From the figure we observe that one of the mass pairs are excluded up to 95% confidence, namely the pair with a slepton mass of 100.5GeV and nautralino mass of 1GeV. The rest of the mass pairs were not excluded up to the same degree of confidence, although one was nearly excluded up to 87% confidence (200.5Gev and 1GeV). We observe that only the pairs with lower slepton mass are close to being excluded. The higher mass pairs (slepton mass $>$ 300GeV) are orders of magnitude off. This indicates the results are effected by low statistics. Given the analyse was made with an integrated luminosity of $10fb^{-1}$, a new analysis with higher luminosity's could lead to more conclusive results.
In figure 15 I plotted the results from SR2. The figure shows that none of the mass pairs

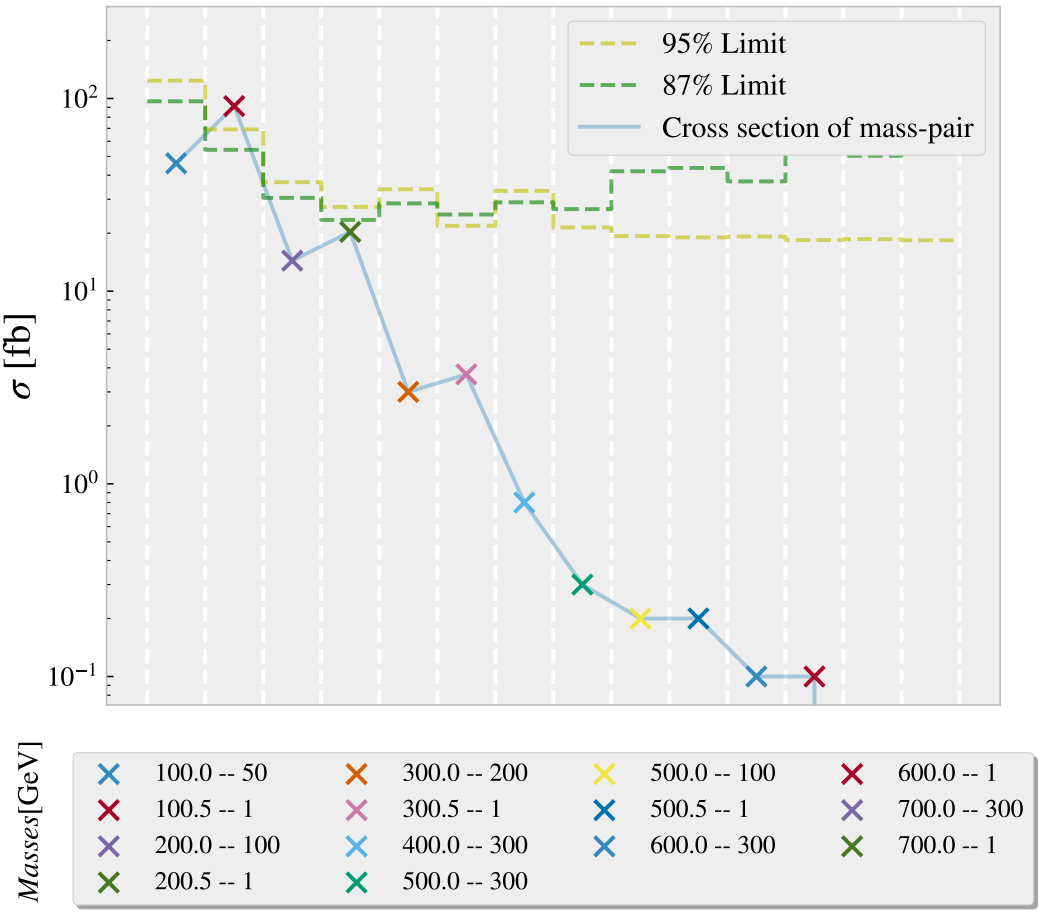


Figure 14: Visualization of crossection of each mass-pair in SR2 with each corresponding exclusion limit up to 95% and 87%.

were excluded. The closest mass pair to be excluded is the pair of 100.5GeV slepton mass and 1GeV neutralino mass, though the cross section of the pair is still an order of 10 away. By comparison with the results from SR1, the results from SR2 are far less conclusive. Additionally we can observe from the figure that the decrease in efficiency has increased the limits by a factor of 10, requiring far more statistics to make any conclusive statements about the data.

By comparing results from SR1 and SR2 we can observe that the rectangle-cut only approach in SR2 resulted in a signal efficiency that heavily effected any possibility of drawing any conclusion from the analysis. The combination of loose rectangle cuts and ML cut was able to create a signal region which efficiently separates background and signal and preserves enough events to conclusively exclude one mass pair.

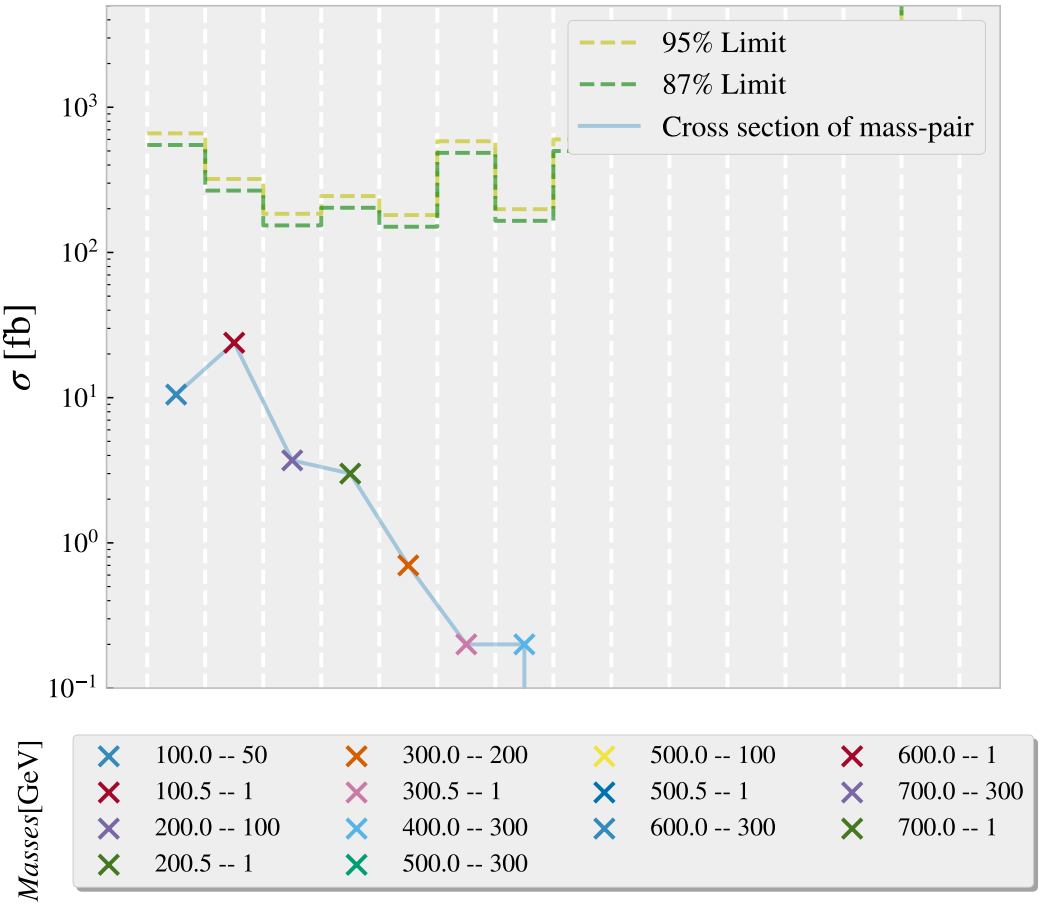


Figure 15: Visualization of crossection of each mass-pair in SR2 with each corresponding exclusion limit up to 95% and 87%.

4.7 The way forward

In this section I will discuss possible improvements to the analysis. The first and most obvious way of improving the search is by attaining more data. Most data sets today have a luminosity of more than 10-times the amount used in this analysis. By comparing the figures 14 and 15 we can observe that there is too little data to exclude or discover any of the lower mass sleptons. Additionally more data features regarding the jets would allow me to better reduce the jet-heavy channels like W/Z+jets or ttbar.
As mentioned in mentioned in section 4.6, the choice of 5% was rather arbitrary. While setting the uncertainty in the calculations of the limits, I noticed that the uncertainty heavily effected the results. Therefore, before any objective results can reached, the real uncertainty of the data and MC background must be added in the analysis.
As for the search itself, the way forward would be to define individual signal regions for each mass pair. For the sake of comparing between SR1 and SR2, one signal region was chosen for each method. I suspect that applying individual signal regions would drastically improve signal efficiency for each pair and lead to better analysis.
Finally there is the problem of the negative weights. For the sake of this rapport I chose the simplest solution, namely to preserve the total sum of the weights. In doing so I risk loosing information regarding the distribution of events in individual regions in the feature space. An alternative method called cell resampling (found in an article by Jeppe R. Andersen, Andreas Maier [8]), suggests a method which would aim to treat small pockets in the phase space ("cell") individually. This would better preserve information in the data and could lead to more accurate results.

5 Conclusion

In this rapport I searched for direct sleptons in the ATLAS open data $10fb^{-1}$. I did so in two different signal regions SR1 and SR2, were the regions were created using two different approaches. SR1 was created with a mixture of rectangle and curved machine learning cuts. SR2 was created using rectangle cuts only. Both signal regions used a common base called the analysis region (AR).
AR was created using loose rectangle cuts. I found that in creating AR I was able to remove more than 90% from every MC-background channels, even reducing 99% from some. At the same time I was able to preserve 26% of the signal.
When developing SR1 I used 80% of the MC data in AR and trained a XGBoost classifier to separate signal from background. When applying the trained classifier on the remaining data, I found that the classifier accurately separated signal from background, achieving a ROC-curve area of 0.963. Using this trained classifier on the remaining data in the AR, I created SR1. SR1 was defined as the data with a ML-score above 0.875. The new signal region was able to preserve more than 80% of the signal while removing more than 75% of all remaining MC-background. The main irreducible background channels in the regions were $Diboson$, $Z + jets$ and $ttbar$.
To create SR2 I was forced to add far stricter variable cuts. I doing so I cut more than 90% of all background, but subsequently 81% of the signal. The cuts totally reduced 7 mass-pairs compared to SR1 that only totally reduced 2. The main irriducible backgrounds in SR2 were found to be the same as for SR1 with the addition of W+jets.

Using the total number of MC-background and data in each signal region, I found that SR1 achieved a significance of 0.25 whereas SR2 achieved 1.20. Neither score was close to discovery, indicating that neither region found any direct slepton in the data.
For exclusion I used 14 different combination of slepton and neutralino masses. Calculating the limit for each pair using Monte Carlo integration, I was able to exclude 1 mass pair using SR1, whereas SR2 excluded to much of the signal data to give any conclusive results. By comparing the results from each region I found that using the XGboost classifier I was able to preserve far more signal data than only using rectangle cuts. The classifier was, as hoped able to detect 'hidden' patterns in the data which allowed for a curved signal region in the feature space. Although the rectangle cuts removed slightly more of the MC-background the subsequent decrease in signal efficiency resulted in a lack of statistics when analysing the results.

References

1. *Review of the 13 TeV ATLAS Open Data release* tech. rep. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-OREACH-PUB-2020-001 (CERN, Geneva, Jan. 2020), 34–35. https://cds.cern.ch/record/2707171.
2. Frette, S., Hirst, W. & Jensen, M. M. A computational analysis of a dense feed forward neural network for regression and classification type problems in comparison to regression methods, 5. https://github.com/Gadangadang/Fys-Stk4155/blob/main/Project%202/article/Project_2_current.pdf (2022).
3. Frette, S. & Hirst, W. Searching for the Higgs through supervised and unsupervised machine learning algorithms. https://github.com/WilliamHirst/Advanced-Machine-Learning/blob/main/article/Project_1.pdf (2022).
4. Search for electroweak production of charginos and sleptons decaying into final states with two leptons and missing transverse momentum in $\sqrt{s} = 13TeV$ pp-collisions using the ATLAS detector. *The European Physical Journal C* **80.** https://doi.org/10.1140%2Fepjc%2Fs10052-019-7594-6 (Feb. 2020).
5. The XGBoost Contributors. *XGBoost* version 1.5.2. https://xgboost.readthedocs.io/en/stable/#.
6. Gramstad, E. *Searches for Supersymmetry in Di-Lepton Final States with the ATLAS Detector at* $\sqrt{s} = 7$ *TeV* PhD thesis (Oslo U., June 2013).
7. Lester, C. G. & Nachman, B. Bisection-based asymmetric M T2 computation: a higher precision calculator than existing symmetric methods. *Journal of High Energy Physics* **2015.** https://doi.org/10.1007%2Fjhep03%282015%29100 (Mar. 2015).
8. Andersen, J. R. & Maier, A. Unbiased Elimination of Negative Weights in Monte Carlo Samples. https://arxiv.org/abs/2109.07851.