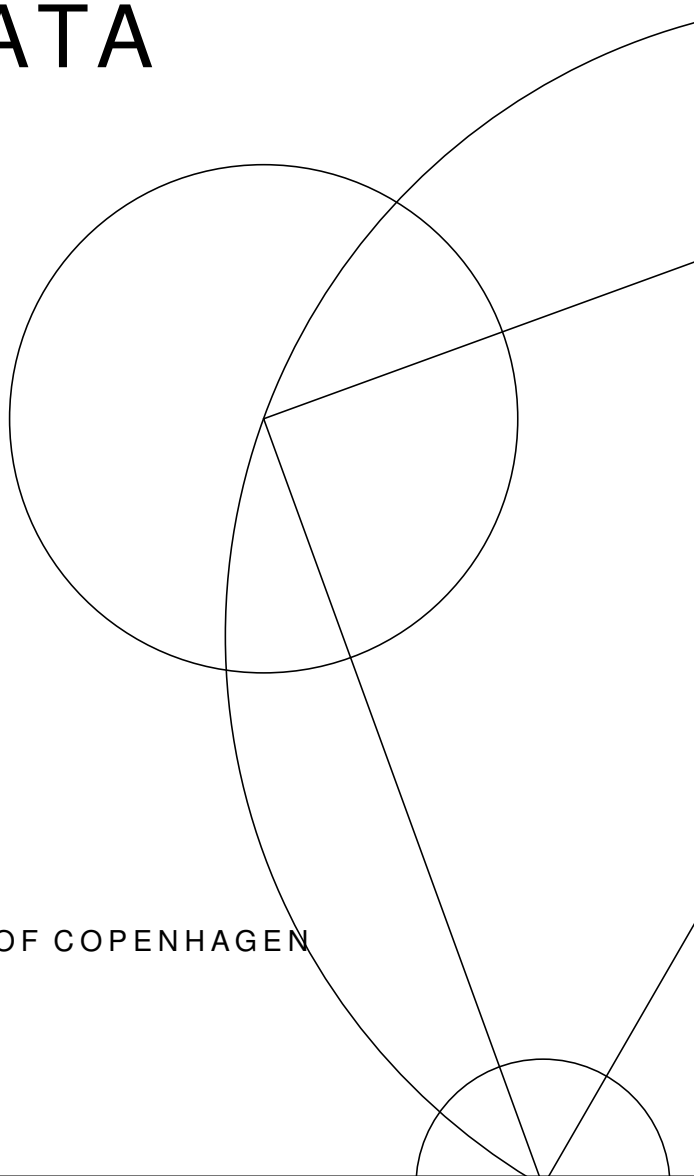
NICOLAS PALM PEREZ

ELECTRON IDENTIFICATION USING MACHINE LEARNING IN THE ATLAS EXPERIMENT WITH 2016 DATA



MSC THESIS
NIELS BOHR INSTITUTE, UNIVERSITY OF COPENHAGEN

ELECTRON IDENTIFICATION USING MACHINE LEARNING IN THE ATLAS EXPERIMENT WITH 2016 DATA

NICOLAS PALM PEREZ

NIELS BOHR INSTITUTE,
UNIVERSITY OF COPENHAGEN

Supervisor	Associate Professor Troels C. Petersen
Censor	Professor David Rousseau (Laboratoire de l'accélérateur linéaire)

Electron identification using machine learning in the ATLAS experiment using 2016 data
MSc Thesis, University of Copenhagen
viii + 78 pages; illustrated, with bibliographic references

Set in 10/14 pt Palatino Linotype using pdfLaTeX with the Tufte-LaTeX package
Cover art: University of Copenhagen logo grid

Thesis submitted on the 4th of September 2017 and defended on the 27th of September 2017 for the completion of the degree of Master of Science (MSc) in Physics at the Niels Bohr Institute, University of Copenhagen.

First printing, September 2017

Contents

Abstract v

Acknowledgement vii

Introduction viii

1 Theory 1

1.1 The ATLAS detector and LHC 2

1.1.1 Overview of the ATLAS detector 3

1.1.2 Inner Detector 4

1.1.3 The Electromagnetic Calorimeter 6

1.1.4 The Hadronic Calorimeter 7

1.1.5 Electron reconstruction 8

1.2 Electrons and other particles in ATLAS 10

1.3 Machine learning 13

1.3.1 Toy data example 13

1.3.2 General concepts in machine learning 13

1.3.3 Boosted Decisions Trees 15

1.3.4 Fisher's discriminant 17

1.3.5 Neural networks 17

1.4 The ATLAS likelihood 20

1.4.1 Construction of the likelihood 20

1.4.2 Binning of the likelihood 20

1.5 Event selection 22

1.5.1 Tag & Probe 22

1.5.2 Background selection 23

1.5.3 Data sets 25

2 Analysis 27

2.1 Introduction to the analysis 28

2.2 Setup 29

2.2.1 Phase-space binning 29

2.2.2 Input variables 33

2.3 MC signal and MC background 38

2.3.1 Method 38

2.3.2 BDT configuration in TMVA 38

2.3.3 Calorimeter BDT 39

2.3.4 Isolation BDT 41

2.3.5 Track BDT 42

2.3.6 Combining calorimeter and track BDTs 43

2.3.7 Results 44

2.4 Data-driven 46
2.4.1 Method 46
2.4.2 Purification of data 47
2.4.3 BDT configuration in TMVA for data training 52
2.4.4 Calorimeter BDT 54
2.4.5 Isolation BDT 54
2.4.6 Track BDT 54
2.4.7 Combining calorimeter and track BDTs 56
2.4.8 Results 57
2.4.9 Effects of correlations between sub-classifiers 62
2.5 Additional variables 63
2.5.1 Re-weighting 63
2.5.2 Performance as a variable of eta, mu and Et 64
2.6 Isolation 66
2.7 Neural networks 67
2.8 Concluding remarks 70
2.8.1 Summary 70
2.8.2 Outlook 70
Appendices 73
A.1 Correlation between variables in data 74
References 77

Abstract

The identification of electrons in the ATLAS experiment is done using a likelihood (LH) based method, which is constructed based on Monte Carlo simulations. In this work machine learning algorithms have been employed for electron identification, as these are expected to be more performant.

The first results in this thesis are from implementation of Boosted Decision trees (BDTs) based on the same variables and the same MC samples as the LH. This yields an increase in background rejection compared to the LH. The improvements decrease when testing the classifier in data. Therefore, a data-driven training method has been developed. Data from 2016 at $\sqrt{s} = 13$ TeV has been used. The data-driven method includes a removal of mis-labeled events providing 99% pure samples for training. The removal is done by separating the discriminating variables into a calorimeter sub-classifier and an inner detector sub-classifier. They can be used to remove mis-labeled events for each other. An isolation classifier has also been constructed to aid in the cleaning process. The data-driven method was implemented for two different boosting algorithms for the BDTs and for a neural network. For adaptive boosting, the results at 92% signal efficiency corresponding to medium LH, gives an improvement in background rejection of 94%. For gradient boosting, the improvement is 104%. Using additional variables yielded an improvement of 109%. The isolation classifier yields improvements of 100-600% compared to an often used isolation variable. The implementation of the neural network results in improvements up to 20% in the calorimeter, while for the inner detector it performed poorly compared to the BDTs.

Acknowledgement

I would like to thank my supervisor, Troels C. Petersen. It has been a very educative year not only in relation to physics, and it has by far been my most joyful year during my education.

I would also like to thank Daniel Nielsen for his endless hours of discussions related to code, physics and everything in between. During this time, a friendship also developed which I am grateful for.

I would also like to thank Alejandro Alonso for advising me on physics and machine learning.

And last, but not least, thanks to my office mates, Rosanna Ignazzi and Christian Michelsen for an awesome atmosphere and Friday wine.

Introduction

This thesis is not a thesis specifically in particle physics or in computer science but in the field between the two fields. The work during the project has been focusing on the application of machine learning (ML) in experimental particle physics. The amount of data in experimental particle, especially from the LHC, is enormous. With the development of GPUs and ML algorithms that are highly parallelized, the training time of the algorithms have decreased drastically. This has increased the complexity of the algorithm that is within reach in reasonable training time, and therefore more tasks are solvable using ML.

There are two chapters in this thesis, a theory chapter and a analysis chapter.

The theory chapter is consisting of an introduction to the ATLAS experiment, a brief overview of the physics detected by ATLAS, a introduction to the basics of ML, a description of the identification of electrons at present and, finally, the method used to obtain data.

In the analysis chapter, the results achieved from implementing BDTs and neural networks are presented together with all the steps made to get there. Furthermore, a new approach on how to make training data-driven is presented.

I Theory

1.1 The ATLAS detector and LHC

The Large Hadron Collider (LHC) was built for proton-proton, proton-lead or lead-lead collisions with a high center-of-mass (c.o.m.) energy in order to study new physics, and with a high collision frequency to study rare processes. Along LHC, four different experiments are conducted, ATLAS, CMS, ALICE and LHCb. The four experiments have individual and overlapping scientific interests.

A map of the four experiments and the different accelerators are shown in Figure 1.1.

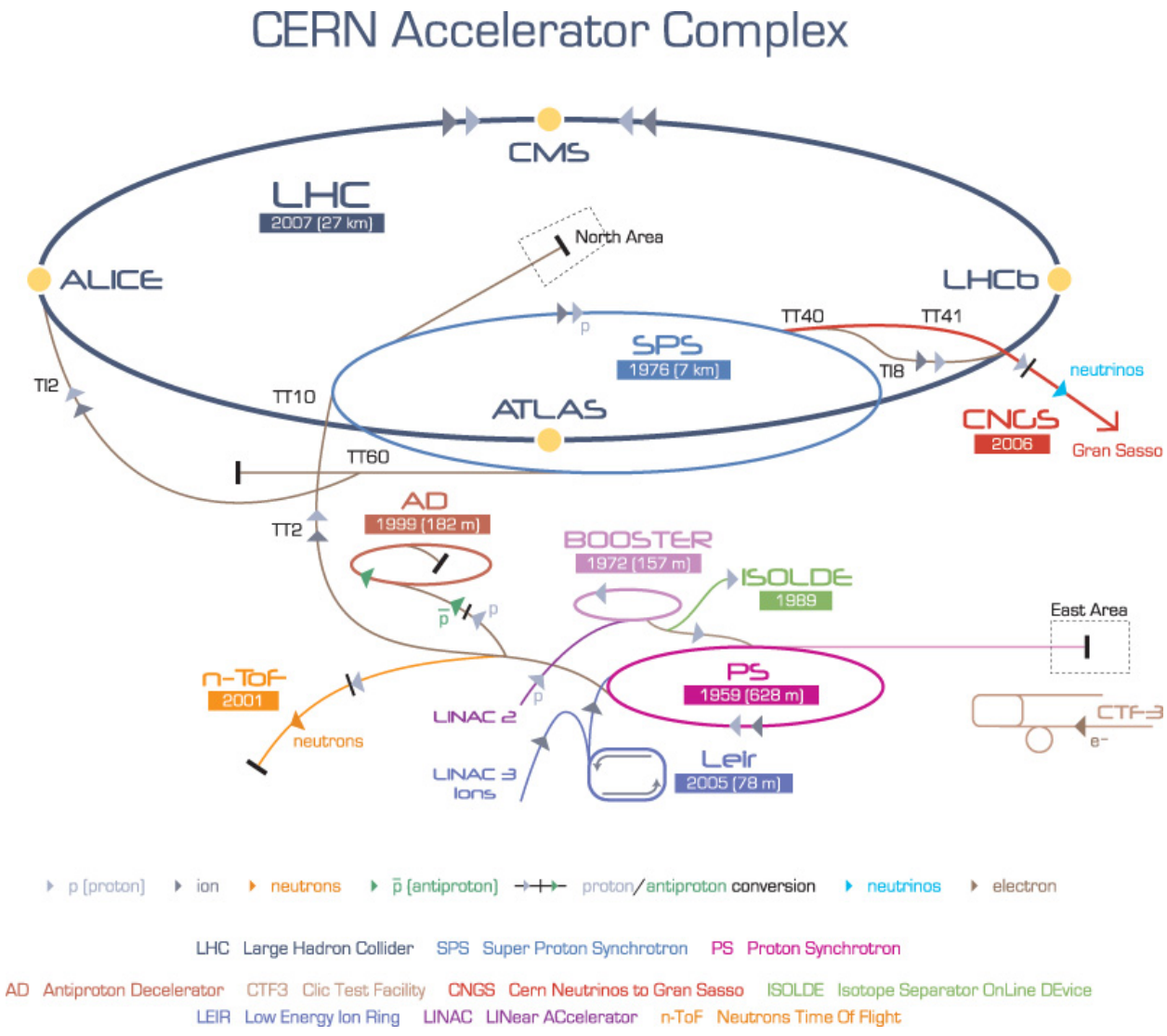


Figure 1.1: The CERN Accelerator Complex. From [1].

Protons are accelerated through a series of accelerators to increase the energy of the protons before entering LHC. A proton beam is injected into the LHC in both directions. When the protons enter the LHC, they have an energy of 450 GeV. The protons are then further accelerated to obtain a c.o.m. energy of 13 TeV [2]. After reaching the required energy, the two opposite directed beams are forced into

collisions at the four experimental sites along the ring. A beam consists of up to 2808 bunches of protons with each bunch consisting of around 10^{11} protons. The bunches are at nominal running separated by approximately 25 ns. It is only very few protons that collide when passing a collision site. After a collision, new particles may be created and either pass through the detector, or decay to something less exotic which then passes through the detector. In both cases the identification of the particles are crucial for searching and studying new physics.

1.1.1 Overview of the ATLAS detector

The ATLAS detector consist of several nested, cylindrical sub-detectors, all of them with different purposes to measure different quantities. There are several sub-detectors where the major and essential parts are the inner detector (ID), the electromagnetic calorimeter (ECAL), the hadronic calorimeter and the muon spectrometer. The muon spectrometer detects muons. For electron identification, ID and ECAL are the most important ones, and therefore only those two will be presented in details. The hadronic calorimeter plays a minor role in electron identification and a major role in the measurement of energy for hadrons. An overview of the detector is shown in Figure 1.2. In the following diagrams of the detector, cylindrical coordinates are used due to the geometry of the detector (R, ϕ) with center in the interaction point. z is used as the coordinate along the beam-line. For describing the trajectory of a particle, transformed polar coordinates are used, ϕ and η (pseudorapidity) where $\eta = -\ln\tan(\theta/2)$ [3].

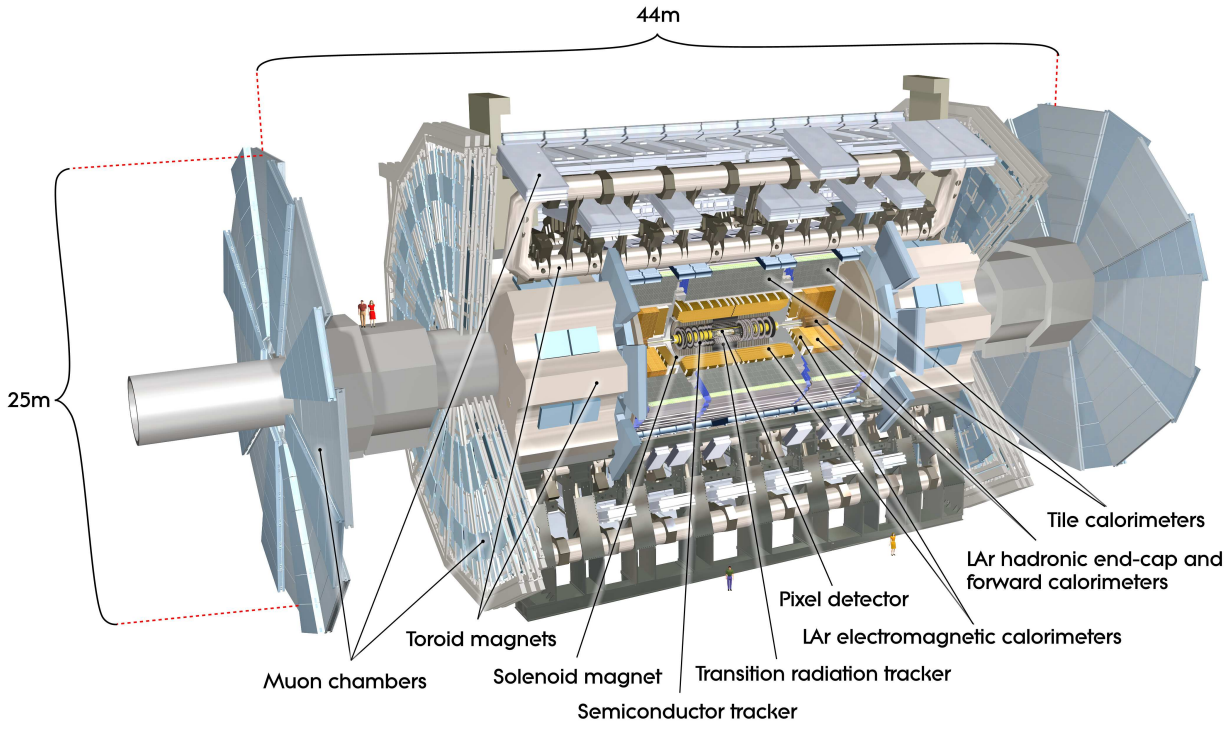


Figure 1.2: The ATLAS detector. From [3].

1.1.2 Inner Detector

The inner detector's main function is to track charged particles. An axial magnetic field of 2 T originating from a solenoid coil is surrounding the ID allowing measurements of momentum for charged particles. The ID is composed of three different sub-detectors, the Pixel detector which is placed closest to the beamline providing the best hit resolution, the Silicon Microstrip Tracker (SCT) which has a lower resolution per hit and the Transition Radiation Tracker (TRT) which again has lower resolution but also provides identification information for electrons. The first two detectors covers $|\eta| < 2.5$ and the TRT $|\eta| < 2.0$ [4]. All three sub-detectors consists of barrel and end-cap parts. A schematic figure of the ID barrel is shown in Figure 1.3 and the end-cap is shown in Figure 1.4.

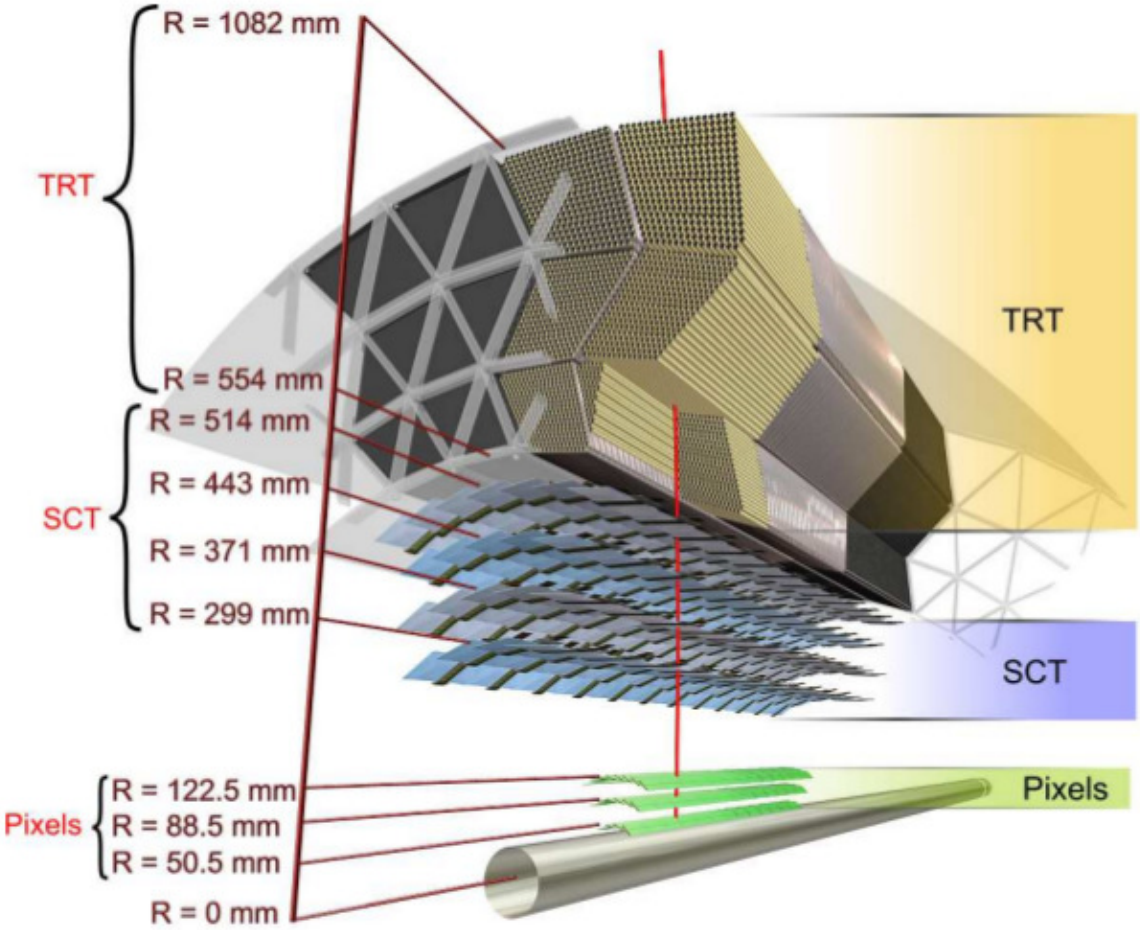


Figure 1.3: The barrel of the Inner Detector. The insertable B-Layer is not shown in this figure. From [3].

The Pixel Detector

The pixel detector consists of three cylindrical layers of pixel sensors in the barrel region, and three layers in the two end-caps. There are a total of 1744 pixel sensors providing approximately 80 million

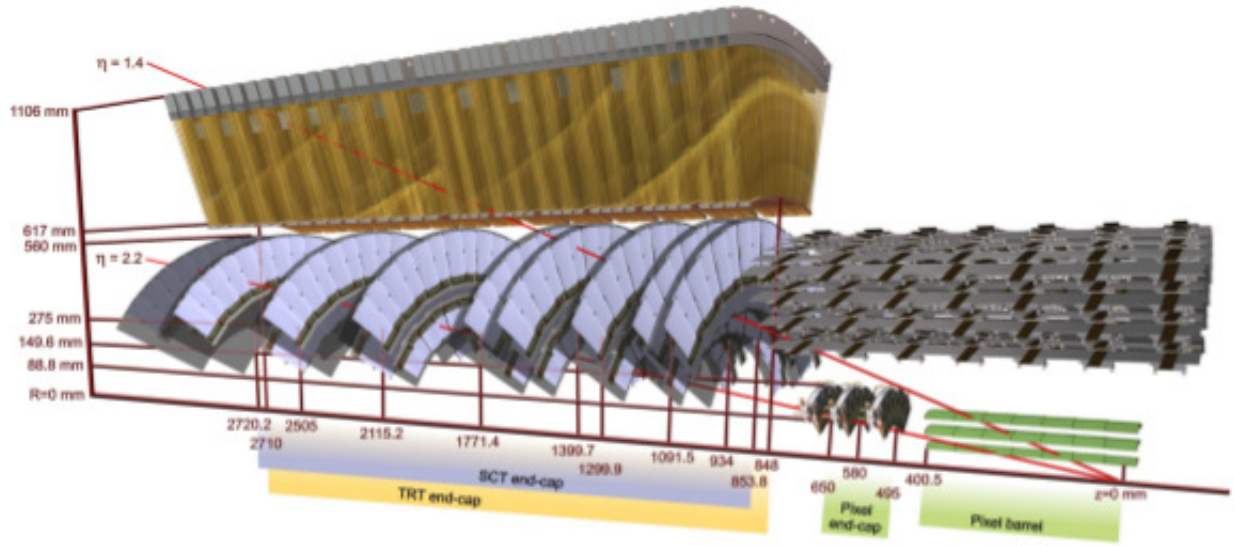


Figure 1.4: The endcap of the Inner Detector. From [3].

pixels. The resolution is $10 \times 115\,\mu m^2$ in $R - \phi \times z$ for the barrel and $R - \phi \times R$ for the end-caps. For Run 2, another layer was inserted before the Pixel detector, the Insertable B-layer (IBL), to overcome the increasing pile-up [5]. The pixel detector usually provides $3 - 4$ hits per track.

The Silicon Microstrip Tracker

The SCT consists of four double-barrel layers and nine end-cap layers. The SCT provides a resolution of $17 \times 580\,\mu m^2$ in $R - \phi \times z$ for the barrel and $R - \phi \times R$ for the end-caps. The SCT is the most powerful tracker in terms of the relative resolution of a track due to the larger radius. It is placed further away from the beamline compared to the pixel detector, and it has an average of 8 hits per charged track, and therefore it contributes more to the relative resolution.

The Transition Radiation Tracker

The TRT is placed furthest away from the beamline. It consists of tubes filled with either Xenon or Argon gas. It does not provide information on η for geometric reasons. When a charged particle traverses the tubes, the gas inside the tubes is ionized. On average a charged track will have $30 - 36$ TRT hits. The probability of an X-ray producing a transition radiation hit in the straws and thereby having a high threshold hit depends on the γ-factor of the particles, and since electrons are lighter than other charged particles, the γ-factor is higher for electrons. In Figure 1.5, the high threshold hit probability is shown as a function of the γ-factor. The figure shows data from Run 1. For higher pile-up, the function is shifted slightly upwards which results in a worse discrimination between electrons and non-electrons.

The amount of straws with Xenon has decreased due to leakages and Argon has replaced Xenon since it is cheaper. For straws filled with Argon, the difference in the onset is much smaller. Finally, the fraction of high threshold hits compared to the total number of hits are used in the TRT likelihood for electron identification.

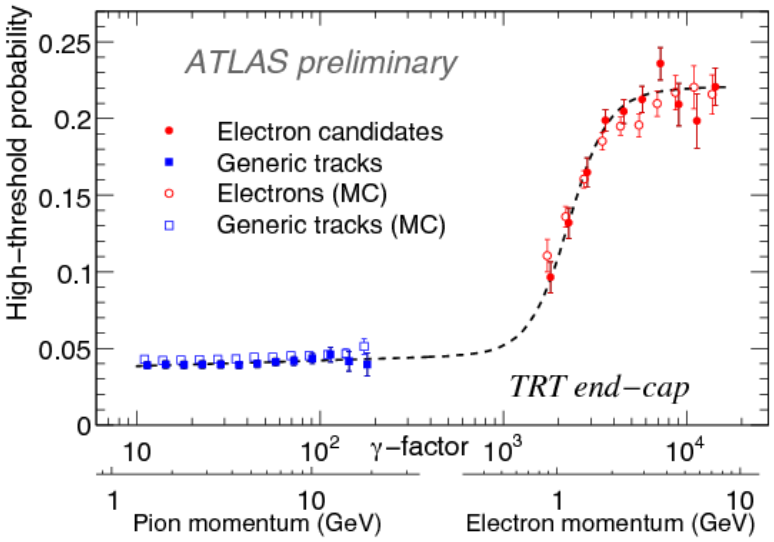


Figure 1.5: The onset curve for Xenon gas in the TRT. For an increasing γ-factor the probability of getting a high threshold hit increases. Electrons usually have a higher γ-factor than non-electrons like pions. The data is from Run 1.

1.1.3 The Electromagnetic Calorimeter

The ECAL is a liquid Argon calorimeter. The ECAL is accordion-shaped. This ensures that no transversing electrons or photons will pass through undetected, see Figure 1.7.

The barrel has four layers: a presampler and a first, second and third layer. Unlike ID, these are very different.

The presampler is correcting for energy loss upstream due to material in the solonoid magnet. For charged particles, the material in the magnet causes a loss of energy while passing through the magnet coil. The presampler is very thin in order to measure if a charged particle is passing through, and it is has a small radiation length such that the particle is depositing a minimal amount of energy in the layer. Furthermore, it is also made such that the probability of a photon converting is small. Depending on η, an additional energy term is added to the measured energy of the charged particle. It only covers up to $|\eta| < 1.8$.

The first layer provides accurate η measurements. It consists of strips with a granularity of 4.69 mm or $\Delta\eta = 0.0031$. This is approximately $1/8$ of the granularity from layer two. In Figure 1.7 it is shown in details.

The second layer measures the main energy deposit of electrons and photons. The depth of the layer is 16 radiation lengths. The granularity in $\Delta\eta \times \Delta\phi$ plane is 0.25×0.245.

The third layer contributes to the measurement of the shower

development from particles and the energy entering the hadronic calorimeter. For electrons, the amount of energy entering the hadronic calorimeter should be small due to the more than 20 radiation lengths they have passed through.

From $1.37 < |\eta| < 1.52$, the calorimeter changes from barrel to end-cap, and therefore the calorimeter is not performing as well in this crack area. In Figure 1.6, an illustration of the calorimeter is shown. The endcap covers from $1.52 < |\eta| < 3.2$. ECAL has an interaction length of approximately one. This means that many hadrons do not interact in the ECAL, which is where most photons and electrons deposit their energy. This feature provides a good discrimination between hadrons and electrons and photons. Furthermore, it is possible to discriminate between electron and photon shower shapes in the ECAL.

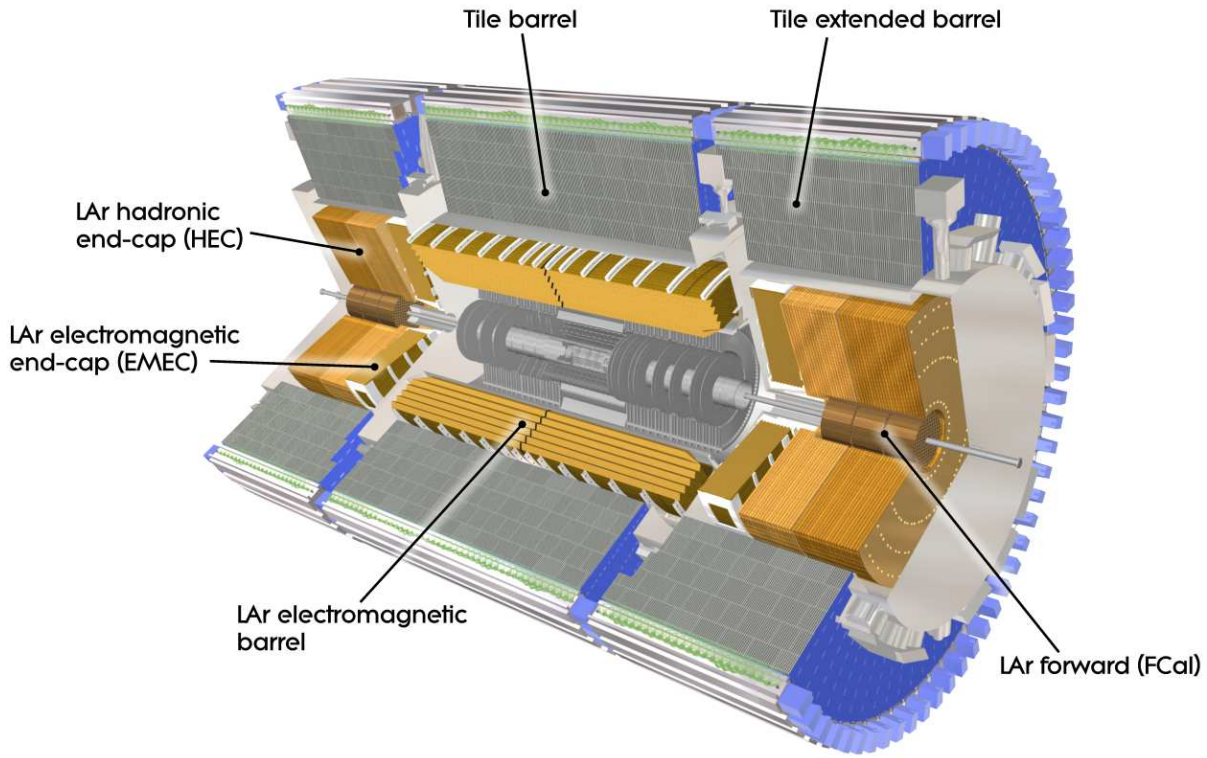


Figure 1.6: The calorimeters of ATLAS. From [3].

1.1.4 The Hadronic Calorimeter

The hadronic calorimeter (HCAL) is a tile calorimeter. It consists of steel as the passive material and plastic scintillators. It has $8 - 10$ of hadronic interaction lengths. Its main purpose is to measure energy from hadrons. For electron identification, it is used to measure the amount of energy entering the calorimeter. For electrons, all the energy is usually deposited in the ECAL but for high energy electrons a small fraction of the energy can enter the HCAL. For jets this fraction

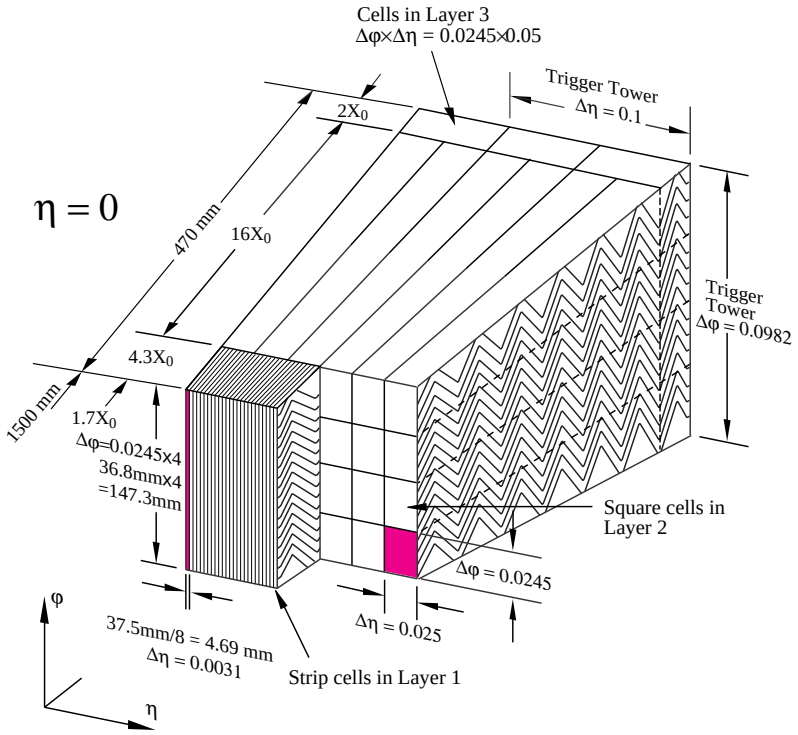


Figure 1.7: A section of the barrel electromagnetic calorimeter. From [3].

is big.

1.1.5 Electron reconstruction

Electron reconstruction for $|\eta| < 2.47$ happens in several steps. For $|\eta| > 2.47$, the ID does not provide tracking information and the procedure is different. The overview of the reconstruction is listed below with further details following:

- The calorimeter seed-cluster reconstruction is performed.
- Tracks are reconstructed in the ID.
- Tracks loosely associated to a cluster and with more than 3 SCT hits are refitted using the Gaussian Sum Filter (GSF) [6].
- Matching the track and cluster to an electron.

Seed-cluster reconstruction

A sliding window corresponding to 3×5 blocks with each block being the granularity of the EM calorimeter middle layer ($\Delta\eta \times \Delta\phi$) is used to search for seed clusters. A tower is defined as the sum of the energy of all the layers at an η and ϕ coordinate. A seed cluster will be a local maximum above 2.5 GeV. After finding a seed cluster, a clustering algorithm is used to create the cluster [7]. After completion, an object of cells called a cluster is formed. In Figure 1.16 top left, the green blocks represents clusters.

Track reconstruction

Track reconstruction is done in two steps. The first step is finding a track seed which consists of three hits in the SCT detector. Second, the seed is extended to a full track using a Kalman Filter [8]. The Kalman Filter associates all the likely hits to a track and afterwards the track is fitted with the global χ^2 fitter to find the most probable track from all the hits associated with the track [9].

Electron hypothesis track fit

The reconstructed tracks are matched with the EM clusters using the distance in η and ϕ after extrapolation to the middle layer of the ECAL. If a track is matched to an EM cluster and it has more than 3 precision hits, the track is refitted using a GFS. The GSF takes non-linear bremsstrahlung effects into account and it gives better tracking for electrons.

Electron candidate reconstruction

The refitted tracks are matched to the EM clusters again with more strict conditions. If several tracks matches a cluster, a primary track is picked based on algorithm using the distance R and track information.

After a successful reconstruction, the four momentum for the electron is calculated based on the best track associated to a cluster and the energy from that cluster. The energy is based on the cluster information and η and ϕ is taken from the track.

1.2 Electrons and other particles in ATLAS

The LHC collides protons, which are ordinarily found in atomic nuclei, at very high energies, and as a result it produces particles which can either be part of the Standard Model (SM), such as electrons and quarks, or could be new yet unobserved ones.

The production rate is approximately 10^8 events/s, or in particle physics terms, the cross section is 10^8 nb.

Not every event produced is of interest. Most events are results of scattering at low energies, which have already been studied in details in previous experiments, and are therefore discarded. The decision of whether to keep an event is done by the ATLAS trigger systems. If the passes the triggers the event is kept.

The discovery of the SM Higgs particle and the measurement of it's properties is arguably the most interesting result from the LHC so far. New searches are looking for beyond the SM model Higgs particles, dark matter candidates, supersymmetry, among others. For most cases, the new particles decay or interact with the known SM particles, which can be measured by the detectors.

As an example, the discovery of the Higgs particle was through different channels such as the $H \rightarrow ZZ^* \rightarrow llll$. Here, each l is a lepton. Therefore, the ZZ decay could be yielding a pair of muons and a pair of electrons, or four particles of the same kind. Due to conservation laws in particle physics, these leptons always come in particle-antiparticle pairs, such as electron-positron. Final states containing electrons are useful in many searches and were crucial in the Higgs boson discovery, therefore the ability to detect electrons is important. The cross section of the Higgs boson is 14 orders of magnitude smaller than the total collision cross section. With help from many different criteria distinguishing a Higgs boson from an ordinary event, one still needs to be able to reject 10^{14} events for each Higgs produced. This example is just to illustrate the task at hand when finding particles in ATLAS.

Regardless of the search, contributions from electrons in finding a particle will for almost all purposes come from a W or Z decay. Electrons originating from those two particle often come isolated, meaning the energy in the nearby area is low compared to the electron energy itself and can be used as a criteria when finding W or Z.

The cross section and production rate of different particles from the SM are shown in Figure 1.8. Around 90% of p-p collisions are pions, often produced in association with other hadrons (kaons, protons). One type of pion is the π^0, which almost instantaneously decay into two photons. Photons interact with the detector electromagnetically and therefore give rise to background when detecting electrons of interest. Photons do not have a track in ID, but can convert into an electron-positron pair. These will result in those photons being detected as electrons (the so-called converted photons), which are not of interest in an analysis looking for electrons. Fortunately, as mentioned earlier, π^0 are produced in association with other hadrons,

and by requiring the electrons to be isolated within certain criteria, most of the π^0 can be removed from the sample of electrons under study.

In this work, only electrons coming from a Z are of interest. The cross section for Z is roughly 10^2 nb. The decay of a Z to two electrons happens 3% of the time. That roughly gives a factor of 10^8 of other particles to electrons coming from a Z. This requires the electron identification to be very efficient at rejecting other particles from electrons. In order to have a pure sample of $Z \to ee$ candidates, one needs to be able to have at least a factor 10^8 background rejection. Overcoming this number is a crucial challenge in this thesis. It is desired to get above 10^{10} background rejection rate, resulting in a sample of $Z \to ee$ particles with 99% purity of electrons coming from Z decays. As with the Higgs, there are requirements like two electrons and that the invariant mass of the particle pair is within the mass of the Z that aids in the rejection of background.

Another worthy mention is the bottom quark. It can decay to a W^* and a c-quark, and with 10% chance the W will decay to an electron. This is typically not an interesting event and is considered background in this work, but since the cross section of b-quarks is large the contribution to mis-labeled electrons from this source is large. Often the electron coming from the W will not be isolated, and an isolation cut can remove most of the b-quark events.

Finally, photons are also present in ATLAS. As mentioned in the previous section, they have similarities in interaction with the ECAL but not in the ID compared to electrons. They can originate from many different processes and converted photons are one of the main contributors to misclassified electrons.

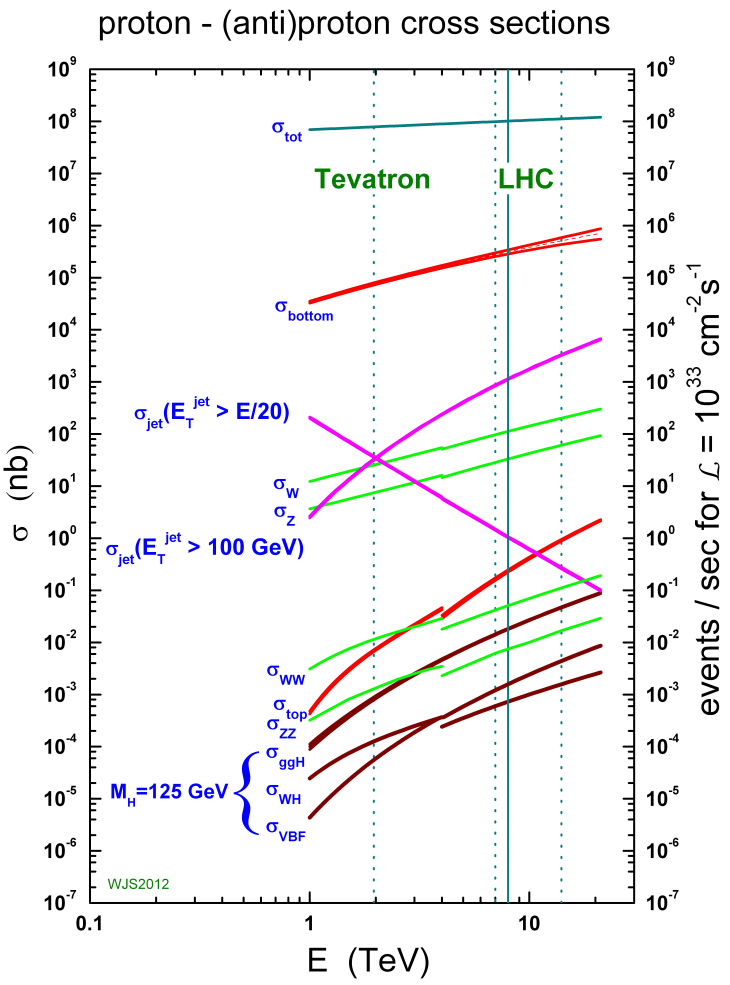


Figure 1.8: Particle production in proton-proton collision. The dashed line furthest to the right is the energy at which the LHC operates in run 2. From [10].

1.3 Machine learning

Machine learning (ML) is the discipline of analyzing data and finding patterns using algorithms without explicitly programming them. Two types of tasks are often solved using machine learning, namely classification and regression problems. In this thesis the task has been a binary classification problem, so only ML related to classification problems are covered.

1.3.1 Toy data example

In Figure 1.9.a a classical text book case of a classification problem is shown. The task is to be able to predict whether a data point is an X or an O. The data are described by two variables, $x1$ and $x2$. So given the data in the figure, create a model that predicts if new data points are X or O.

In Figure 1.9.b the straight red line is a good division of phase space for separation of Xs and Os, though one O event is placed on the X side of the line.

One could use a more complex separation line, which would place all the events correct. But the O that is on the wrong side, might be a statistical fluctuation and not a general feature of the X and O distribution. If a more complicated model is used to learn statistical fluctuation, the model is over trained.

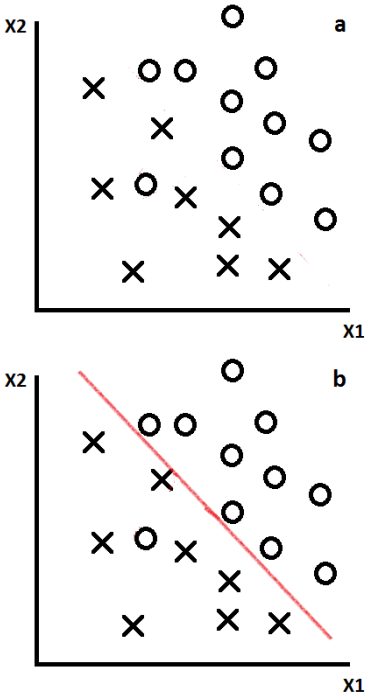


Figure 1.9: Toy classification problem. In a, the data points are shown with their respective labels, X and O. In b a line separating X and O are made.

1.3.2 General concepts in machine learning

In this sub-section general concepts of machine learning will be presented.

Data

Data is the most important part in ML. The quality of the data is important and in the case of ATLAS data, the preparation and quality control is very high. The amount of data is also crucial for the amount of information or patterns that an algorithm can learn. The more data, the less likely it is that over training occurs and the more subtle features in data can be learned if present. The data sample can be split into sub-samples for training and testing to avoid over training.

Input variables

Data is described by variables. The number of variables can vary from a few to thousands if not millions. Regardless of the number of variables, they should contribute with information in the process of predicting an event. Often a transformation of the variables can be useful. Later, a simple transformation used in this work is described.

Labels

For each data point a label follows stating if the event is signal or background. This is called supervised learning. Unsupervised learning also exists, which is when no label information is present. In this thesis, each event has a label and therefore methods for supervised learning are used. It is important that the labels are correct when training. Some algorithms are more robust against mis-labeled data compared to others. In this thesis mis-labeled data and the removal of such has been important in order to use ML for electron identification.

Algorithms

There exists many different algorithms and they all have different advantages and weaknesses. For each problem different types of algorithms might be optimal. The choice of algorithm depends on the complexity and structure of data, the amount of available data, number of variables and the correlations between the variables. For this work the data consists of $8 - 12$ variables and the number of data points ranges from thousands to a few million events.

For this work, boosted decision trees (BDT) [11] and neural networks (NN) [12] have been considered. This was due to earlier work done in ATLAS, where BDTs and NNs often are seen performing well. The BDTs have been used for all the cases. The NNs have only been used for data where the number of events available for training are high.

A Fisher's discriminant is also used in this thesis [13]. It is based on linear correlation analysis. The red line on Figure 1.9.b could have been based on a Fisher's discriminant.

Training & Test

In supervised learning the chosen algorithm needs to be trained on data to learn the features in order to predict the label of data. That means finding the parameters for the chosen algorithm that predicts the best without learning statistical fluctuations. The number of parameters are often too large to scan the whole parameter space. Therefore a minimization algorithm is chosen to find the optimal set of parameters that minimizes the cost function (defined below). Depending on the chosen ML algorithm the minimization algorithm varies, especially for NN many different minimization algorithms exists. Often computing time is limited, and a trade-off occurs between having a minimizer that quickly finds a minimum or a minimizer that is good at escaping local minima.

The training is often done on one sub-sample of the data. Usually this part is the largest fraction of data. The other part is called a test sample. The test sample is used to test whether the algorithm is doing as well as in the training sample. If there are significant discrepancies between the performance on the training sample and the test sample, over training have occurred.

If there are discrepancies between test and training, a simpler algorithm e.g. fewer parameters should be employed, or for NN the number of training iterations or the number of neurons should be decreased.

Cost functions

The cost function is the function that has to be minimized such the the ML algorithm performs as well as possible. For different ML algorithm different cost functions can be useful. In Figure 1.10 three different cost functions are shown for BDTs. p_1 is the probability of being correct in a given node. In this thesis cross entropy (CE) were used for NN. The error-rate (E), is defined as the fraction of mis-classified events out of the total number of events. As an example, a tree consisting of two end-nodes with $N_{sig} = 400, N_{bkg} = 200$ and $N_{sig} = 200, N_{bkg} = 400$ and a tree with $N_{sig} = 200, N_{bkg} = 0$ and $N_{sig} = 400, N_{bkg} = 600$ would yield the same error-rate (0.33) but the CE cost function would favor the second tree.

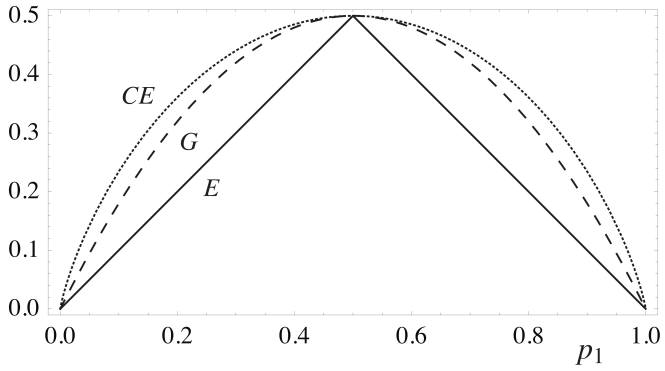


Figure 1.10: Cost functions for BDTs. p_1 is the probability of being correct in a given node. CE is the cross entropy, G is the Gini index. E is the error-rate. The first two favors pure nodes in trees where the latter only takes the total mis-classified events into account. From `http://efavdb.com/notes-on-trees/`

Evaluation

The evaluation of the algorithm's performance can be done in several ways and is usually done on the test data. In this thesis, the end result is evaluated with a receiver operating characteristic (ROC) curve. An example of an ROC curve can be seen in Figure 1.11 bottom. For better separation between the two distributions, the ROC curve will go towards the $(0, 1)$ corner of the plot. If there is no separation between two distributions the ROC-curve will be a straight line from $(0, 0)$ to $(1, 1)$. Other options are the accuracy or some combination of the error rate for signal and background.

In the case of electrons and non-electrons, the ROC curve describes the amount of non-electrons that are accepted as electrons for a specific amount of electron acceptance also called fakes.

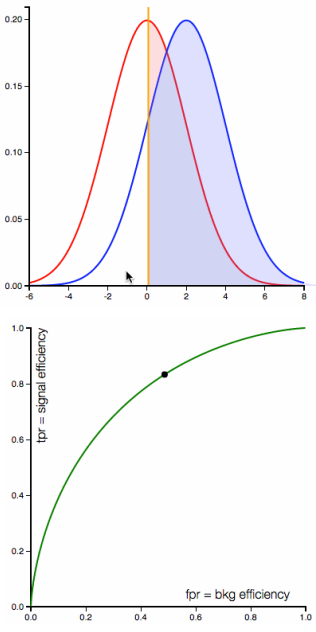


Figure 1.11: Example of a ROC curve. The yellow line corresponds to the black dot on the ROC curve. From [14]

1.3.3 Boosted Decisions Trees

The idea behind a binary classification decision tree is to partition input space into regions where one label is dominant. After one

partition of input space there will be some mis-classified events. The boosting algorithm assigns a higher weight to the mis-classified events according to the boosting method used, and a new tree is trained on the re-weighted data. After repeating this procedure, a whole bunch of trees are created that all predicts if an event is signal or background, and an average prediction is calculated stating if an event is signal like or background like. In this work the TMVA BDTs were used [15]. The boosting algorithm used were adaBoost [16] and Gradient boosting [17].

Decision tree

An example of a simple decision tree is illustrated in Figure 1.12. In this example data has three input variables, gender, age and number of spouses or siblings aboard (sibsp) and it morbidly describes the survival of passengers on Titanic. The first number in each leave is the percentage of surviving passengers and the second number is the percentage of passengers in each leave.

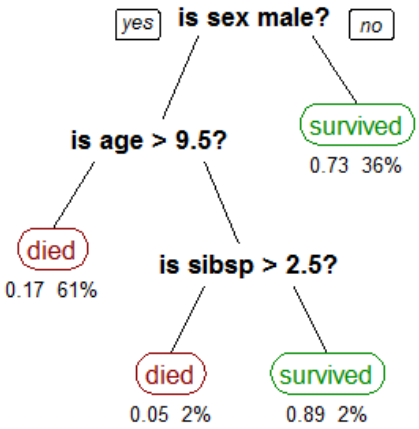


Figure 1.12: Decision tree showing the survival on Titanic. The first number in each leave is the percentage of surviving and the second number is the percentage of data in each leave. From [18].

Adaptive Boosting & Gradient boosting

For Adaptive Boosting (adaBoost) the loss function is an exponential loss function. For each tree all the misclassified events get a higher weight that is calculated the following way,

$$\alpha_m = \frac{1 - err}{err} \tag{1.1}$$

where err is the error-rate of the tree, $err = N_{miss}/N_{total}$. The weights are then renormalized such that the sum of weights stays constant. The overall prediction is calculated the following way,

$$Y_M(x) = \frac{1}{M} \sum_{m=1}^{M} \ln(\alpha_m) y_m(x) \tag{1.2}$$

Here $Y_M(x)$ is the output number from the M trees. $y_m(x)$ is the prediction of the m'th tree. The final number, $Y_M(x)$, is the sum of all the individual trees, weighted with the same weights as the misclassified events are assigned with in the training process.

For Gradient boosting in TMVA the loss function is a binomial log-likelihood function, $L(F, y) = \ln(1 + \exp(-2F(x)y))$. The boosting procedure changes from adaBoost, and will not be presented here. One advantage with a Gradient boosting compared to adaBoost is that it is less sensitive to mis-labeled events [15].

1.3.4 Fisher's discriminant

A Fisher's discriminant is an analysis of linear correlation between input variables. In Figure 1.9, the red line could come from a Fisher's discriminant. The Fisher discriminant is calculated the following way,

$$F = w_0 + \bar{w}\bar{x} \tag{1.3}$$

where the weight vector, $\bar{w}$, is calculated using the covariance matrix for signal and background, $\sum_{S(B)}$ and the mean, $\mu_{S(B)}$. w_0 is the bias, often used to shift the values such that background is below 0 and signal is above 0.

$$\bar{w} = (\Sigma_S + \Sigma_B)^{-1} (\mu_S - \mu_B) \tag{1.4}$$

After calculating F for signal and background, the two distributions can be used to calculate a ROC curve.

1.3.5 Neural networks

A neural network (NN) is another ML algorithm that can be used for classification or regression. It was originally inspired by real neurons and how they communicate. In this thesis only the feed forward NN was used though other types exist.

A feed forward NN works from left to right as seen in Figure 1.13. A single neuron gets activated by an input coming from the left and outputs a single value to the right. This value works as input for all the neurons in the next layer. The mapping from input to output is through an activation function. Originally, the activation function was inspired by real neurons with a sigmoid function. This results in a neuron having a threshold value for activation. Now, the activation functions can take many different shapes. The architecture, meaning the number of neurons and layers, are important for the performance of a network. The more neurons and layers the more complicated features can be learned by the network.

An example of a NN

In Figure 1.13 a NN with an input layer, a hidden layer and an output layer is shown. Each input variable is fed into the input layer. Each y_n^1 represent a neuron from the input layer. They then feed their

activation into all the neurons in the next layer with a weight, $w^1_{i,j}$, which can be different. The superscripts describes which layer it belongs to, i describes from which neuron in the layer the input comes from and j to which neuron in the next layer that the input belongs to. For each neuron in the next layer, a sum of all the input is made, and the result is put into the activation function, which then outputs a number. This procedure is continued until the output layer is reached. For a binary classification problem, the output layer is either a single neuron or two neurons depending on the activation function used for the output layer.

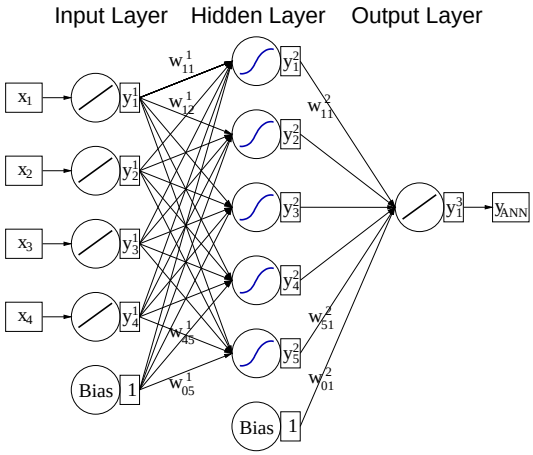


Figure 1.13: An neural network with one hidden layer. From [15].

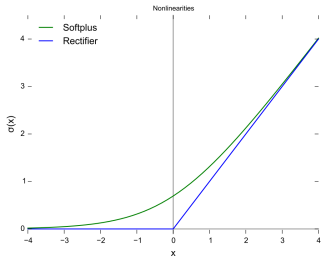


Figure 1.14: The shape of the ReLu and softplus activation functions

Activation functions

There are different activation functions. In this thesis several different functions have been tried, but the ones with the most success was Rectified linear unit (ReLu) and softplus.

ReLu is defined the following way,

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}. \tag{1.5}$$

Softplus is defined the following way,

$$f(x) = \ln(1 + e^x). \tag{1.6}$$

ReLu is fast to compute, but it's derivative, which is important for optimization of the weights is ill defined in 0. The softplus activation has the same feature as the ReLu activation, and the derivative is defined in 0. See Figure 1.14.

For the output layer both a sigmoid and softmax activation function have been used. The softmax generally worked better than the sigmoid. The softmax is defined the following way,

$$F_i(x) = \frac{\exp(x_i)}{\Sigma_{j=1}^{J} \exp(x_j)} \tag{1.7}$$

Here x_i is the total input to neuron i in the output layer, and the numerator is the sum of all the output from the neurons in the output layer. This ensures that the output layer is normalized, such that the sum of $\Sigma_{j=1}^{J} F_i = 1$. In the binary classification case there are two neurons in the output layer for softmax.

Transformation of input variables

For many ML algorithms a transformation of the input variables can change the performance of the algorithm dramatically. NNs are sensitive to preprocessing of the input. The simplest way is a transformation such that the minimum value and the maximum value is -1 and 1. Other transformations exists, such as changing the shape of a distribution, but for this work only this transformation was done.

Cost function

The cost function of a NN is important for the performance of a NN. There exists different cost functions and a preliminary study showed that the binary cross entropy performed the best. This cost function is shown in Figure 1.10.

Back-propagation and minimization algorithms

In order to find the optimal weights for a NN different approaches can be used. To do this a the back propagation algorithm is used. It is a way of propagating the errors from the output neurons to the input neurons. This gives a way to update all the weights based on the gradient of the cost function such that a minimum is found. In this work a stochastic gradient decent optimizer have been used. This method is less likely to end in a local minima of the cost function. Both the adam and nadam optimizers have been tried, where nadam performed slightly better than adam. Both of them are stochastic gradient based optimizer. See [19] for further information.

1.4 The ATLAS likelihood

At present, the electron identification in ATLAS is based on a likelihood (LH) method. It is based on the variables shown in Figure 1.15. The distribution of the variables will be shown in Section 2.2.

1.4.1 Construction of the likelihood

The LH method is based on one-dimensional Probability Density Functions (PDFs) for each of the variables originating from histograms. It does not take any correlations into account. The PDFs are at present constructed based on MC simulations. It is based on $Z \rightarrow ee$ and $J/\Psi \rightarrow ee$ for signal and $JF17$ for background[1]. The background is a di-jet called $JF17$ where 17 is related to energy in GeV. The LH is constructed in the following way [20],

[1] This information comes from private correspondence with Joey Reichert, but I have not been able to find any documents confirming this.

$$d_L = \frac{L_s}{L_s + L_b}, \tag{1.8}$$

where the $L_{s(b)}$ is the LH value for signal (background) and is calculated the following way,

$$L_{s(b)} = \prod_{i=1}^{n} P_{s(b),i}(x_i). \tag{1.9}$$

$\bar{x}$ is the input vector with all the variables. $P_{s(b),i}$ is the PDF constructed for signal (background). The PDFs constructed from MC are shifted linearly and the widths are changed to be more consistent with data. If no correlation between the variables are present, the LH is under general circumstances the most powerful discriminant (Neyman-Pearson Lemma [21]).

The LH has some advantages compared to some of the ML algorithms described in the previous section. Firstly, it is simple to construct and it does not involve any training step apart from creating histograms. Secondly, it is not as sensitive to smaller data samples, since only a one-dimensional PDF needs to be created per variable for signal and background. ML algorithms might exploit more of the parameter space but consequently also need more statistics to create effective classifiers.

1.4.2 Binning of the likelihood

As mentioned earlier, the LH method is the most powerful discriminant if there are no correlation between variables. This is not the case, but by binning in η and E_T the correlations between the variables are decreasing and thereby the LH becomes closer at being optimal. It has 14 bins in pT, and 22 bins in eta [2] [20]. For each combination of phase-space (PS), a PDF for signal (background) is constructed.

[2] η: (-2.47, -2.37, -2.01, -1.81, -1.52, -1.37, -1.15, -0.8, -0.6, -0.1, 0, 0.1, 0.6, 0.8, 1.15, 1.37, 1.52, 1.81, 2.01, 2.37, 2.47)
E_T: (7, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 80, 150)

Type	Description	Name
Hadronic leakage	Ratio of E_{T} in the first layer of the hadronic calorimeter to E_{T} of the EM cluster (used over the range $\lvert\eta\rvert < 0.8$ or $\lvert\eta\rvert > 1.37$)	R_{had1}
	Ratio of E_{T} in the hadronic calorimeter to E_{T} of the EM cluster (used over the range $0.8 < \lvert\eta\rvert < 1.37$)	R_{had}
Back layer of EM calorimeter	Ratio of the energy in the back layer to the total energy in the EM accordion calorimeter. This variable is only used below 100 GeV because it is known to be inefficient at high energies.	f_3
Middle layer of EM calorimeter	Lateral shower width, $\sqrt{(\Sigma E_i \eta_i^2)/(\Sigma E_i) - ((\Sigma E_i \eta_i)/(\Sigma E_i))^2}$, where E_i is the energy and η_i is the pseudorapidity of cell i and the sum is calculated within a window of 3×5 cells	$w_{\eta 2}$
	Ratio of the energy in 3×3 cells over the energy in 3×7 cells centered at the electron cluster position	R_ϕ
	Ratio of the energy in 3×7 cells over the energy in 7×7 cells centered at the electron cluster position	R_η
Strip layer of EM calorimeter	Shower width, $\sqrt{(\Sigma E_i (i - i_{\mathrm{max}})^2)/(\Sigma E_i)}$, where i runs over all strips in a window of $\Delta\eta \times \Delta\phi \approx 0.0625 \times 0.2$, corresponding typically to 20 strips in η, and i_{max} is the index of the highest-energy strip	w_{stot}
	Ratio of the energy difference between the largest and second largest energy deposits in the cluster over the sum of these energies	E_{ratio}
	Ratio of the energy in the strip layer to the total energy in the EM accordion calorimeter	f_1
Track conditions	Number of hits in the innermost pixel layer; discriminates against photon conversions	n_{Blayer}
	Number of hits in the pixel detector	n_{Pixel}
	Number of total hits in the pixel and SCT detectors	n_{Si}
	Transverse impact parameter with respect to the beam-line	d_0
	Significance of transverse impact parameter defined as the ratio of d_0 and its uncertainty	d_0/σ_{d_0}
	Momentum lost by the track between the perigee and the last measurement point divided by the original momentum	$\Delta p/p$
TRT	Likelihood probability based on transition radiation in the TRT	eProbabilityHT
Track-cluster matching	$\Delta\eta$ between the cluster position in the strip layer and the extrapolated track	$\Delta\eta_1$
	$\Delta\phi$ between the cluster position in the middle layer and the track extrapolated from the perigee	$\Delta\phi_2$
	Defined as $\Delta\phi_2$, but the track momentum is rescaled to the cluster energy before extrapolating the track from the perigee to the middle layer of the calorimeter	$\Delta\phi_{\mathrm{res}}$
	Ratio of the cluster energy to the track momentum	E/p

Figure 1.15: Variables used for the likelihood. From [20]. w_{stot} was not used for the ML. At the beginning of the project, the list of variables was taken from Run 1, where this variable was not included.

1.5 Event selection

In order to obtain electrons (signal) and non-electrons (background) for training of the ML algorithms, event selections are needed. The way of selecting events are the same as for the LH method [20].

As described in Section 1.2, electrons from Z and W are usually of interest. Therefore, it is desirable to select electrons from the decay of one of these particles. However, the selection of electrons needs to be unbiased. Thus, the selection of electrons cannot be based on any information that is related to identification, e.g. triggers. For each event there are many reconstructed particles and it is impossible to find the electrons picking random particles.

In the case of a $Z \to ee$ in an event, one electron can be identified with a tight identification requirement. This leaves the second electron unbiased, and if this can be found it can be selected. This procedure is called Tag & Probe (T&P), and it will be explained in further details later in this chapter. Electrons from the W are produced with almost a factor 10 more than the Z electrons and therefore provides an electron source 10 times as high. But it does not have one electron to trigger on like the Z but only a neutrino. They cannot be measured by the detector and they are only seen indirectly through missing energy in events.

For background selection, selection criteria involve vetos against electron sources namely, Z and W. This will explained further in Section 1.5.2.

1.5.1 Tag & Probe

For selection of signal the T&P method is applied on each event. Every electron candidate is tested as a tag particle and as a probe particle. They are shown below,

Selection criteria for T&P:

1. Veto LAr Error (Event level).
2. Pass Good Runs List (Event level).
3. Number of verticies > 0 (Event level).
4. $|\eta| < 2.47$ (Tag and Probe).
5. Veto on $1.37 < |\eta| < 1.52$ (Tag).
6. $E_T > 25$ GeV (Tag).
7. Pass tight LH (Tag).
8. $E_T > 15$ GeV (Probe).
9. Opposite charge of Tag particle (Probe).
10. B jet veto (not implemented).
11. Only one T&P pair within $Z_m \pm 10$ GeV.

The first two criteria relate to the functionality of the detector. The third requires tracks originating from at least one interaction point (vertex).

The next step is to look at the electron candidates that have been constructed for an event. Only electron candidates within $|\eta| < 2.47$ are considered due to the coverage of the ID.

The fifth criterion relates to the crack region of the calorimeter. The calorimeter is not fully equipped in $1.37 < |\eta| < 1.52$, resulting in a less certain identification.

The next step is an $E_T > 25$ GeV requirement on the tag electron. For higher E_T, the ability to identify electrons increases, and significantly so in the calorimeter. The tag electron also needs to pass tight LH. This ensures that the tag electron almost certainly is an electron.

The probe electron candidate has an $E_T > 15$ GeV requirement. For $E_T < 15$ GeV, the number of electrons originating from a Z compared to background rapidly decreases. This of course limits the identification of electrons to 15 GeV. For lower energies, another electron source is used (not implemented for the ML algorithms in this work, but this is the case for the LH).

The next step is demanding that the T&P pair has opposite signs. Due to conservation laws in physics, this must be fulfilled.

Also, a b-quark veto is applied. As described in Section 1.2, the b-quark can decay into a c-quark and a W which then further can decay into an electron. This type of event does not have interest in terms of electron identification.

Finally, the T&P pair needs to have an invariant mass close to the Z_{mass} (± 10 GeV). If more than one pair fulfills this requirement the whole event is skipped to avoid getting fake electrons.

To sum up the procedure, most of the criteria are related to detector constraints. The tight identification of one electron, the opposite sign and the invariant mass demand is what makes up the selection of unbiased electrons. This almost ensure that it was a Z initially and therefore, if it decayed to electrons, the probe electron candidate will be an electron.

Finally, if the probe also passes the tag cuts, the two candidates can change roles, and both cases are selected.

In Figure 1.16, an example of $Z \rightarrow ee$ is shown. The two green tracks are electrons and the yellow tracks are typically hadrons. The T&P procedure makes it possible to find one green track given the other and knowledge about the Z.

1.5.2 Background selection

For the selection of background, the idea is to exclude any electron sources, namely electrons from Z and W. The selection criteria are shown below.

Selection for background:

1. Veto LAr Error (Event level).
2. Pass Good Runs List (Event level).
3. Number of verticies > 0 (Event level).
4. MeT < 25 GeV (Event level).

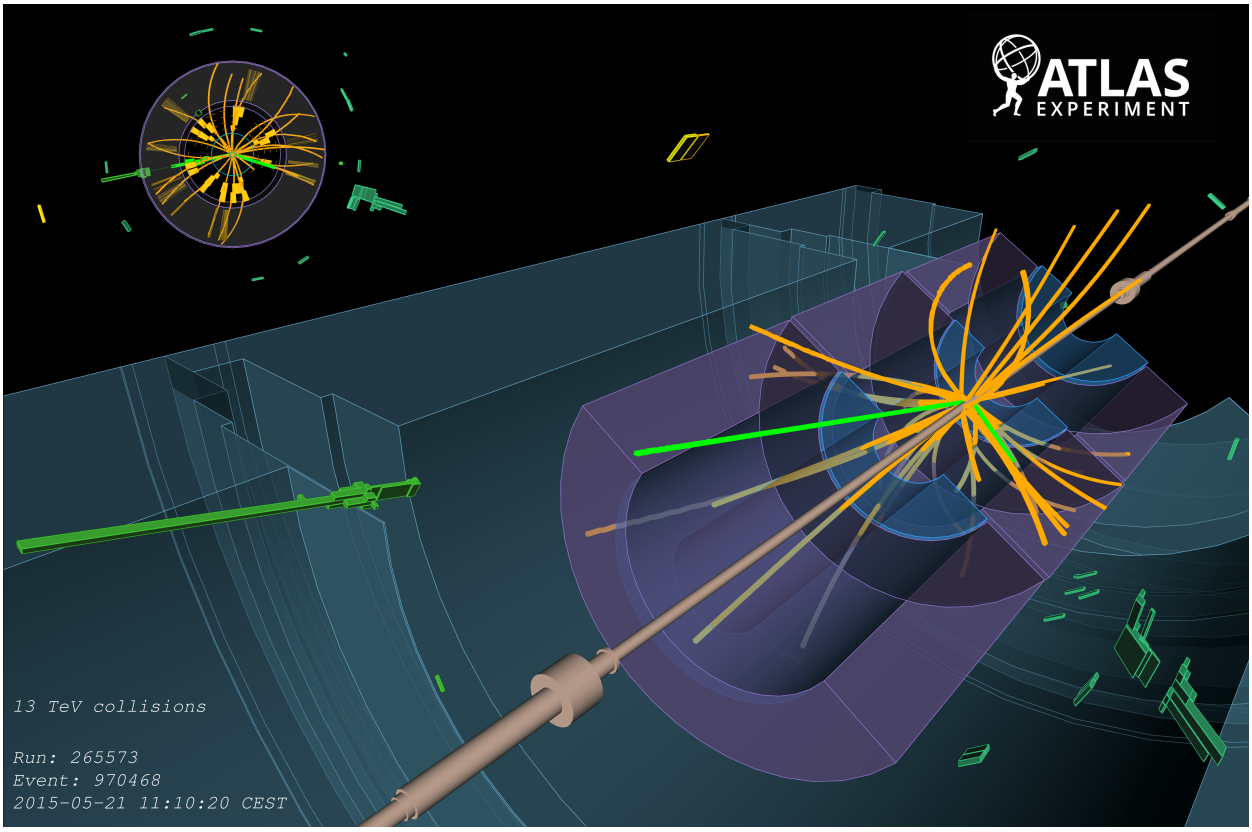


Figure 1.16: Example of an $Z \rightarrow ee$ event in the ATLAS detector. Two electron tracks in green are shown. The yellow tracks are other reconstructed particles. If both particles pass the tag and probe criteria both, will be kept. From [22].

5. $|\eta| < 2.47$ (Particle level).
6. $P_t > 15\,\text{GeV}$ (Particle level).
7. $mT < 40\,\text{GeV}$ (Particle level) [3].
8. Z_m veto,$\pm 20\,\text{GeV}$, paired with particle passing medium likelihood (Particle level).

[3] $m_T = \sqrt{2E_T MeT(1-\cos\theta)}$, where θ is the angle between the particle and the MeT in the transverse plane

Most of the requirements are the same as for signal. There are basically two vetos, one against Z particles and one against W.

The Z veto is the last cut. All electron candidates are paired together to check if any pairs have an invariant mass close to the Z_m.

The MeT and mT are W vetos. MeT is the transverse missing energy. For W decaying to an electron and a neutrino, the E_T of the neutrino will not be measured, but only seen as missing E_T therefore the cut on MeT. The measurement is not very certain though. The transverse mass veto is also a W veto.

1.5.3 Data sets

For this analysis, data from 2016 was used at $\sqrt{s} = 13$ TeV. The EGAM1 derivation produced by the EGamma group was used as the electron sample, which is selected to have $Z \to ee$ events [23].

For background events, the EGAM7 derivation was used. It contains events where at least one HLT e/gamma trigger has fired and at least one electron has been reconstructed [23]. By adding vetos against Z and W, the electron candidates left will likely be other objects reconstructed as electrons. The identification task is exactly to distinguish between reconstructed real electrons and reconstructed fake electrons.

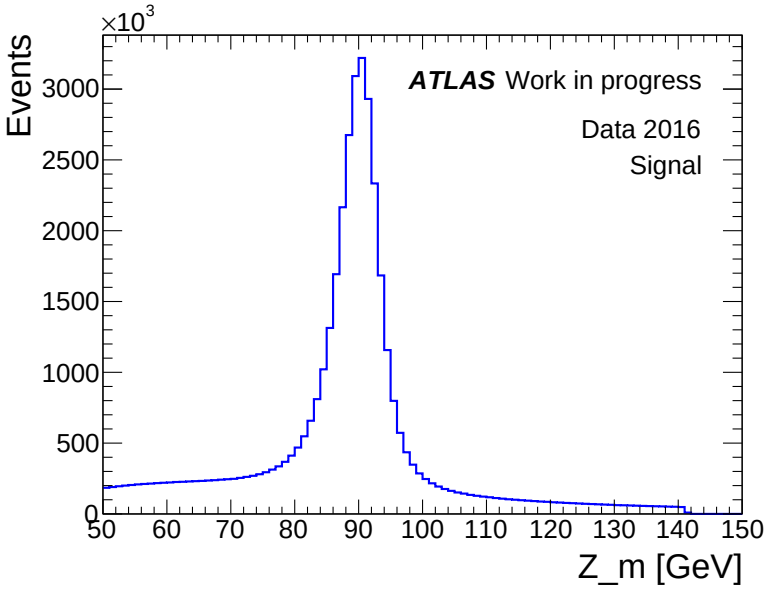


Figure 1.17: The Z peak from T&P for 2016 data. It contains 4.00×10^7 electron pairs.

For MC signal samples $Z \to ee$ samples have been used. An MC background corresponding to real background is difficult to achieve.

As for the LH, JF17 samples have been used as background. In order to get background events with higher E_T, JF35 and JF50 was used as well. The TRT conditions were not updated for the JF50. This contributes to the inconsistency between data and MC and between MC from different files.

For data, the result from T&P selection is shown in Figure 1.17 before the invariant mass cut. From the tails of the distribution it is hinted that not all of the events are $Z \to ee$. This will be confirmed in the next chapter.

The T&P and background selection gives 30 million signal and background candidate events for data. For MC signal 2.8 million candidates and for MC background 17 million candidates are selected.

Name
data16_13TeV:data16_13TeV.00304006.physics_Main.merge.DAOD_EGAM1.f716_m1620_p2689/
data16_13TeV:data16_13TeV.00311170.physics_Main.merge.DAOD_EGAM7.f758_m1710_p2840/
mc15_13TeV.361106.PowhegPythia8EvtGen_AZNLOCTEQ6L1_Zee.merge.AOD.e3601_s2876_r7917_r7676
mc15_13TeV.423300.Pythia8EvtGen_A14NNPDF23LO_perf_JF17.merge.AOD.e3848_s2876_r7917_r7676/
mc15_13TeV.423302.Pythia8EvtGen_A14NNPDF23LO_perf_JF35.merge.AOD.e3848_s2876_r7886_r7676/
mc15_13TeV.423303.Pythia8EvtGen_A14NNPDF23LO_perf_JF50.merge.AOD.e3848_s2608_s2183_r7773_r7676/

Table 1.1: File used for this study.

II Analysis

2.1 Introduction to the analysis

The identification of electrons in this work has had three main objectives:

- Implementation of ML methods to improve ID of electrons.
- Make an ID tool based on data alone.
- Include additional variables such as η and $\langle\mu\rangle$.

The identification will be divided into 25 PS bins in η and E_T. This results in 25 classifiers and it makes reporting of every PS bin difficult. Therefore, four PS bins covering different η and E_T are presented in details for most steps in the analysis. The other PS bins have been inspected during the analysis but only their final results are reported.

The results will be presented in the following order:

- Performance of BDT based classifiers on MC.
- Performance of BDT based classifiers on data, including a method to purify data to allow for a data-based training and the improvements of including additional variables compared to the LH.
- Results from implementation of NN on data.

Finally, the results are summarized in a conclusion and the outlook of the project is presented.

2.2 Setup

In this section the general setup of the analysis and variables are presented.

2.2.1 Phase-space binning

As mentioned in Section 1.4, the LH does not take correlations into account. To decrease the correlation between variables, a binning of PS in E_T and $|\eta|$ is done. For the same reasons as with the LH, binning of PS was done for the ML methods but with fewer bins. The bin boundaries for $|\eta|$ are listed in Table 2.1.

$\lvert\eta\rvert$
0.8, 1.37, 1.52, 2.01, 2.47

Table 2.1: Bin boundaries for $|\eta|$.

There are fewer PS bins to increase the statistics, and due to the fact that the MVA methods can handle correlations, but by having some PS bin boundaries, the features that the classifiers need to learn are simpler, and thereby they reach an optimal solution easier. Furthermore, it is assumed that the detector identical for positive and negative η. The PS bin boundary values are all contained in the likelihood bin boundaries. For E_T the PS bin boundaries are shown in Table 2.2.

E_T [GeV]
15, 20, 30, 40, 50

Table 2.2: Bin boundaries for E_T.

An example of the effect of binning is shown in Figure 2.1. The left plot shows the correlation between two calorimeter variables, f_1 and *Eratio*, for $E_T < 20$ GeV and $|\eta| < 1.37$. The correlations are significant and non-linear. By adding a bin boundary in $|\eta| = 0.8$, the non-linear correlation is significantly reduced. The middle plot shows $|\eta| < 0.8$ and the right plot shows $0.8 < |\eta| < 1.37$.

The binning of PS results in a different number of events for each PS bin for training. In Table 2.3, the number of signal and background events in data for each PS bin are shown. In Table 2.4, the number of MC events are shown. Note that from MC to data the range of events is a few thousands to a few millions.

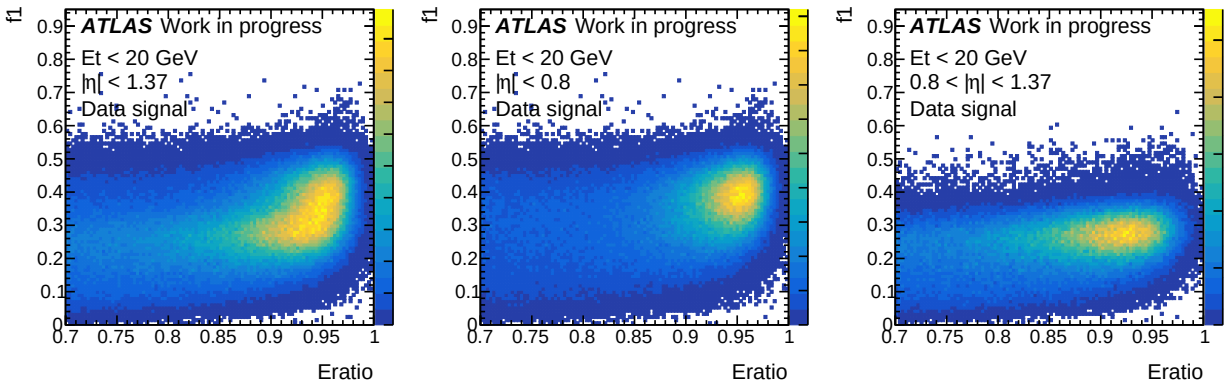


Figure 2.1: Distributions of f_1 and *Eratio*. Left: $E_T < 20$ GeV and $|\eta| < 1.37$. Middle: $E_T < 20$ GeV and $|\eta| < 0.8$. Right: $E_T < 20$ GeV and $0.8 < |\eta| < 1.37$. Note the reduction non-linear correlation between the variables by adding a bin in $|\eta| = 0.8$.

Phase-space bin	N_{sig}	N_{bkg}
$\|\eta\|$: 0.0-0.8, E_T : $15-20$ GeV	469906	1685462
$\|\eta\|$: 0.0-0.8, E_T :20-30 GeV	1108839	1818264
$\|\eta\|$: 0.0-0.8, E_T :30-40 GeV	1980751	885217
$\|\eta\|$: 0.0-0.8, E_T :40-50 GeV	2009917	496354
$\|\eta\|$: 0.0-0.8, E_T :> 50	772805	1271295
$\|\eta\|$: 0.8-1.37, E_T :15-20 GeV	270343	1208465
$\|\eta\|$: 0.8-1.37, E_T :20-30 GeV	667370	1235842
$\|\eta\|$: 0.8-1.37, E_T :30-40 GeV	1266422	589155
$\|\eta\|$: 0.8-1.37, E_T :40-50 GeV	1251563	329322
$\|\eta\|$: 0.8-1.37, E_T :> 50 GeV	472112	891831
$\|\eta\|$: 1.37-1.52, E_T :15-20 GeV	69597	390567
$\|\eta\|$: 1.37-1.52, E_T :20-30 GeV	162579	390509
$\|\eta\|$: 1.37-1.52, E_T :30-40 GeV	274980	172204
$\|\eta\|$: 1.37-1.52, E_T :40-50 GeV	281559	92881
$\|\eta\|$: 1.37-1.52, E_T :> 50	116749	261649
$\|\eta\|$: 1.52-2.01, E_T :15-20 GeV	196565	946920
$\|\eta\|$: 1.52-2.01, E_T :20-30 GeV	470203	974158
$\|\eta\|$: 1.52-2.01, E_T :30-40 GeV	754292	440165
$\|\eta\|$: 1.52-2.01, E_T :40-50 GeV	762926	233544
$\|\eta\|$: 1.52-2.01, E_T :> 50	299217	508931
$\|\eta\|$: 2.01-2.47, E_T :15-20 GeV	150401	802763
$\|\eta\|$: 2.01-2.47, E_T :20-30 GeV	340632	884479
$\|\eta\|$: 2.01-2.47, E_T :30-40 GeV	518046	364099
$\|\eta\|$: 2.01-2.47, E_T :40-50 GeV	539061	180897
$\|\eta\|$: 2.01-2.47, E_T :> 50	210429	340732

Table 2.3: Number of signal and background events for data training in each PS bin from T&P and background selection.

Table 2.4: Number of signal and background events for MC training in each PS bin from T&P and background selection.

Phase-space bin	N_{sig}	N_{bkg}
$\|\eta\|$: 0.0-0.8, E_T :15-20 GeV	9006	276504
$\|\eta\|$: 0.0-0.8, E_T :20-30 GeV	49123	236844
$\|\eta\|$: 0.0-0.8, E_T :30-40 GeV	117961	81508
$\|\eta\|$: 0.0-0.8, E_T :40-50 GeV	124466	27715
$\|\eta\|$: 0.0-0.8, E_T :> 50	44665	17340
$\|\eta\|$: 0.8-1.37, E_T :15-20 GeV	5570	1208465
$\|\eta\|$: 0.8-1.37, E_T :20-30 GeV	28133	174201
$\|\eta\|$: 0.8-1.37, E_T :30-40 GeV	73269	62716
$\|\eta\|$: 0.8-1.37, E_T :40-50 GeV	78908	21797
$\|\eta\|$: 0.8-1.37, E_T :> 50 GeV	27746	14178
$\|\eta\|$: 1.37-1.52, E_T :15-20 GeV	1442	50112
$\|\eta\|$: 1.37-1.52, E_T :20-30 GeV	6521	48177
$\|\eta\|$: 1.37-1.52, E_T :30-40 GeV	15653	19916
$\|\eta\|$: 1.37-1.52, E_T :40-50 GeV	17878	7601
$\|\eta\|$: 1.37-1.52, E_T :> 50	6674	5404
$\|\eta\|$: 1.52-2.01, E_T :15-20 GeV	4603	145905
$\|\eta\|$: 1.52-2.01, E_T :20-30 GeV	20031	129039
$\|\eta\|$: 1.52-2.01, E_T :30-40 GeV	43631	46221
$\|\eta\|$: 1.52-2.01, E_T :40-50 GeV	49300	16057
$\|\eta\|$: 1.52-2.01, E_T :> 50	17569	9576
$\|\eta\|$: 2.01-2.47, E_T :15-20 GeV	3885	114159
$\|\eta\|$: 2.01-2.47, E_T :20-30 GeV	16052	98368
$\|\eta\|$: 2.01-2.47, E_T :30-40 GeV	32121	34875
$\|\eta\|$: 2.01-2.47, E_T :40-50 GeV	34016	11943
$\|\eta\|$: 2.01-2.47, E_T :> 50	11147	6898

2.2.2 Input variables

The input variables for classification of electrons will be the same as for the LH to benchmark against the LH. They will be divided into calorimeter and ID variables. The reasons for that is to obtain two sub-classifiers that will aid in purification of data. The sub-classifiers will be combined into one classifier that will be comparable to the LH. An isolation classifier will also be constructed based on isolation variables from the calorimeter and ID. The ID variables will be called track from now on. Finally, additional variables will be included in the classifiers to enhance the performance further.

All the variables are high-level variables. As an example $f1$ describes the ratio between the energy in the first layer against the total energy of a particle. It is a characteristic of the shower shape, and it is discriminating between different particles. It consists of energy measurements from different strips and cells and reduces many measurements into one number. It does so with great success, and with all the other high-level variables the problem of identifying electrons is reduced to 8 variables from many variables (strip and cell measurements in the calorimeter). For most ML algorithms, a much higher number of variables is not a problem. A future study of electron identification with low-level variables (strip and cell information) and high-level variables would possibly improve the calorimeter sub-classifier.

The calorimeter variables are shown in Table 2.5.

Calorimeter variables
RHad1, RHad, f3, weta2, Rphi, Reta, Eratio, f1

Table 2.5: Calorimeter variables.

The additional calorimeter variables are:

- η, used to provide indirect information about detector geometry for the ML algorithm.
- averageInteractionPerCrossing, used to provide information about pileup. For higher pileup, more noise is present in the detector. The averageInteractionPerCrossing was used instead of actualInteractionPerCrossing since this variable had negative values (known bug).

The tracking variables are shown in Table 2.6.

Track variables
nOfInnermostPixHits, nOfPixHits, nOfSCTHits, d0, d0Oversigmad0, dPOverP, deltaEta1, deltaPhiRescaled2 , E/p, TRTPID

Table 2.6: Track variables.

The addtional track varaibles are:

- nOfTRTHits (TRT information)

- nOfTRTXenonHits (TRT information)

The LH variables a described in details in Figure 1.15.

The isolation variables are all based on energy nearby the particle in the $\phi \times \eta$ plane. A cone30 is a distance, $R = \sqrt{\Delta\eta^2 + \Delta\phi^2} < 0.3$, and etcone30 is the sum of energy from all other particles within a distance of $R < 0.3$.

Isolation variables:

- etcone20 (Calorimeter isolation)
- etcone30 (Calorimeter isolation)
- etcone40 (Calorimeter isolation)
- etcone20ptCorr (Calorimeter isolation)
- etcone30ptCorr (Calorimeter isolation)
- etcone40ptCorr (Calorimeter isolation)
- ptcone20 (Track isolation)
- ptcone30 (Track isolation)
- ptcone40 (Track isolation)

New isolation variable:

- ptPU30 (Pileup variable)

To account for the increasing pileup, ptPU30 was created. It is calculated summing over E_T for all tracks within a cone30 coming from a vertex different from the particle's vertex.

Instead of using one of the variables for isolation, all of them are combined to determine better degree of isolation including the new isolation variable.

In the following, the input variables for data for a given PS bin are shown. In Figure 2.2, the calorimeter variables are shown. In Figure 2.3, the isolation variables are shown. In Figure 2.4, the tracking variables are shown. All of the variables are shown after event selection. The linear correlations between the variables for $0.8 < |\eta| < 1.37$ and $30 < E_T < 40$ GeV PS bin are shown in Appendix A.1. The additional variables are weighted such that $\langle\mu\rangle$ and $|\eta|$ are the same for signal and background. The reasons are explained in Section 2.4.

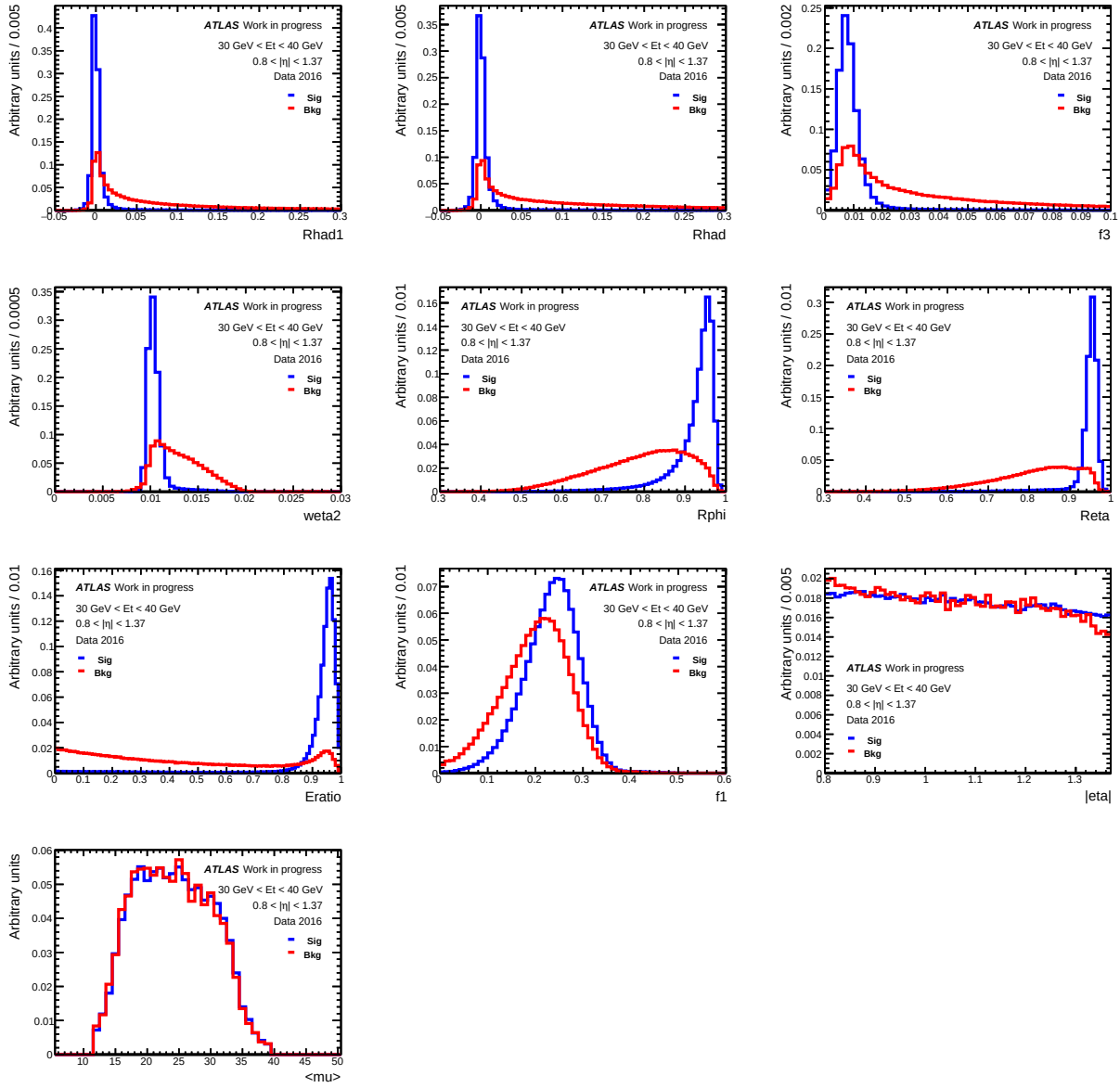


Figure 2.2: Calorimeter variables for data for $30 < E_T < 40$ GeV and $0.8 < |\eta| < 1.37$. This is data obtained from event selection. $\langle\mu\rangle$ and η are weighted to be identical for signal and background for reasons explained later.

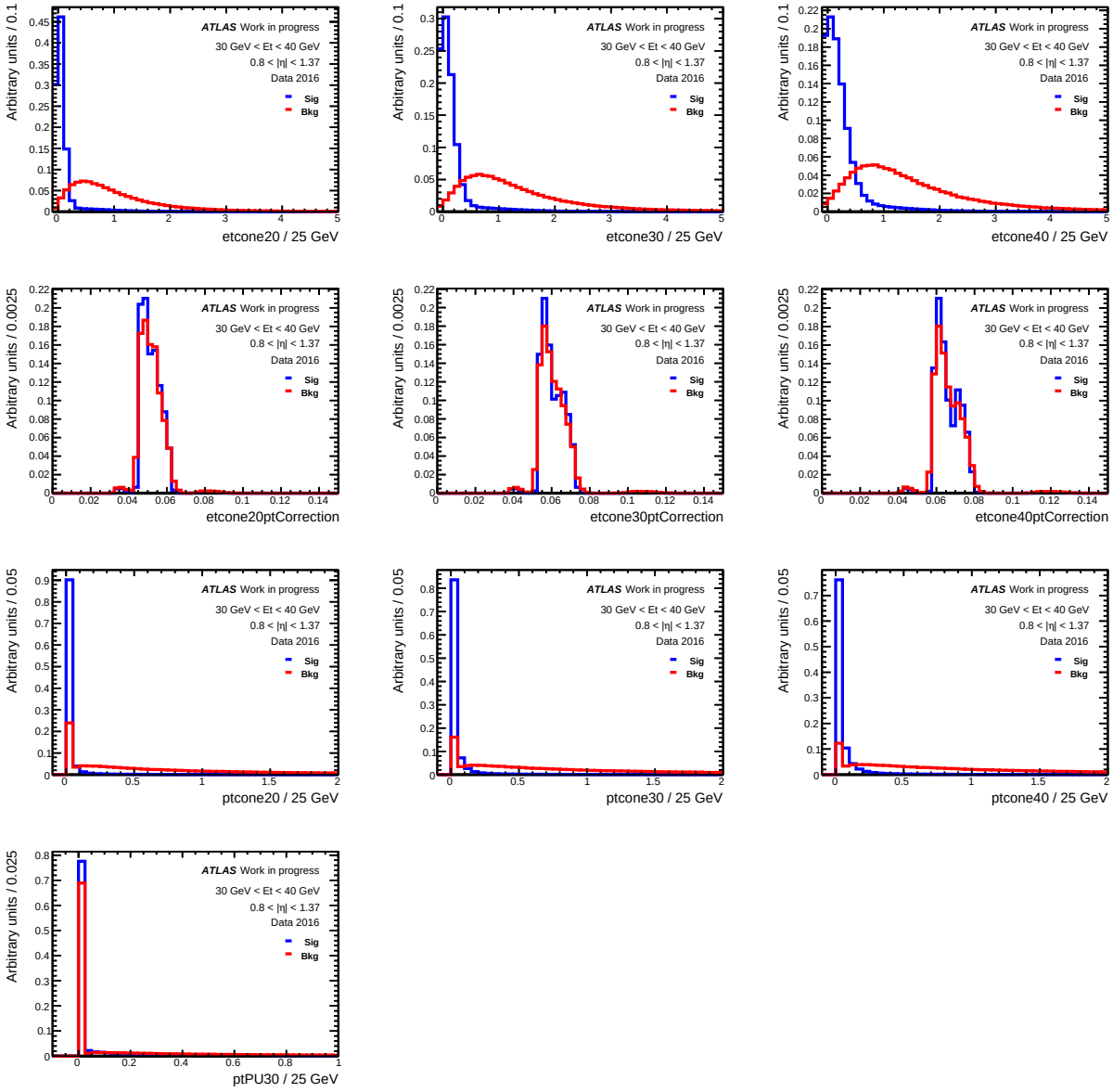


Figure 2.3: Isolation variables for data for $30 < E_T < 40$ GeV and $0.8 < |\eta| < 1.37$. This is data obtained from event selection.

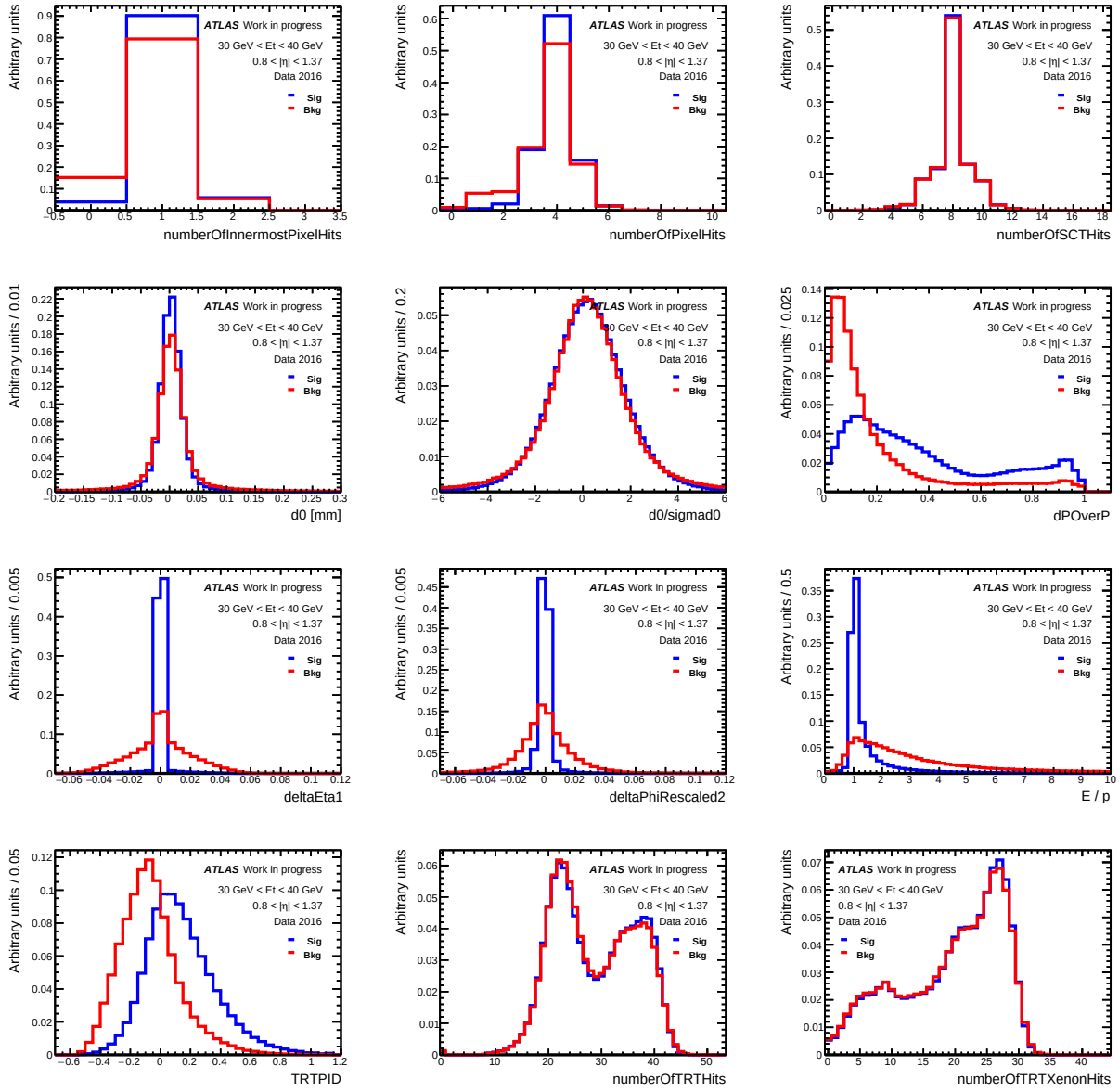


Figure 2.4: Track variables for data for $30 < E_T < 40$ GeV and $0.8 < |\eta| < 1.37$. This is data obtained from event selection. numberOfTRTHits and numberOfTRTXenonHits are weighted with same weight used for $\langle\mu\rangle$ and η.

2.3 MC signal and MC background

In this section the implementation of BDT based classifiers trained on MC is presented. The variables used for the classifier are the same as for the LH.

2.3.1 Method

The work-flow for construction of a BDT based classifier trained on MC is shown in Figure 2.5.

First, a signal sample is created based on T&P on Zee simulation including all the noise (pile-up) that is present in data. No truth matching is done, originally in order to keep it as close to data as possible, but missing all the background present in the signal sample for data, this was in retrospect a poor choice. A background sample is created using the background selection on $JF17/35/50$ simulations with no truth matching as well. The samples created after the event selections are called Ntuples. The LH was constructed based on these signal and background simulation and therefore this is good starting point for comparison of the LH and BDTs based method.

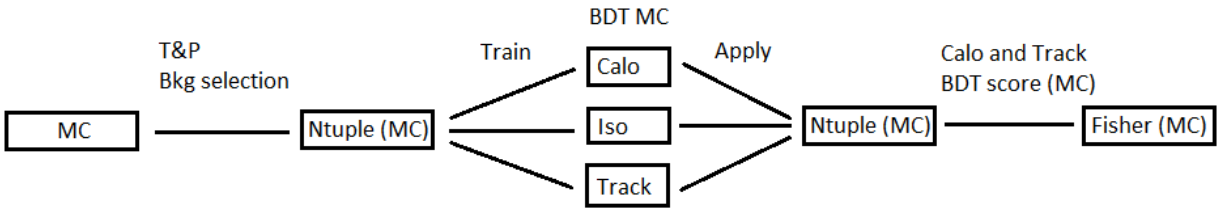


Figure 2.5: Diagram for construction of a classifier for electron identification. The BDTs are trained on three groups of variables, calorimeter, isolation and track variables. Finally, the calorimeter and track scores are combined with a Fisher's discriminant to create one classifier.

After obtaining the Ntuples, three different BDTs are trained on the calorimeter, isolation and tracking variables on 80% of the Ntuples. The number of events for training is shown in Table 2.4. Afterwards, the BDTs are applied on the last 20% of the Ntuples. The reason for splitting up the variables into calorimeter, isolation and track is due to the data-driven method and will be explained later.

The last step is to combine the sub-classifiers into one classifier to compare with the LH. This has been done for the calorimeter BDT scores and the tracking BDT scores with a Fisher's discriminant. For reasons explained in section 1.4, the isolation is left out of the Fisher's such that the LH and the BDT based classifier can be compared on equal terms.

2.3.2 BDT configuration in TMVA

The settings for the training of the BDTs are the same for each PS bin and for all three sub-classifiers. Ideally, the hyper-parameters for the BDTs should be optimized for each classifier and for each PS bin in $|\eta|$ and E_T. The options chosen for training are the default options in

TMVA for BDTs with exception of the number of trees, the minimum amount of data in each node of the BDT and the maximum depth of the trees [15].

The number of trees have been chosen based on training and test for the $|\eta| < 0.8$, $20\,\text{GeV} < E_T < 30\,\text{GeV}$ PS bin.

In Figure 2.6, ROC curves from several different NTrees options are shown. The blue curve is 50 trees, magenta is 100 trees, red is 200 trees and green is 400 trees. It shows that there is almost no improvements from adding more trees, but specially from 200 to 400 there is no gain. Therefore, 200 trees has been picked. In this particular PS bin the amount of events to train on is of average size. For PS bins with fewer events, fewer trees might be optimal, but as seen, more trees than needed does not degrade the ROC curve. The black curve shows the ROC curve from a Fisher's discriminant based on the calorimeter input variables. The Fisher's performs significantly worse than the BDTs, as expected. It is included to show the gain in performance when using a classifier that takes non-linear correlations into account (BDTs) compared to a linear-correlation classifier (Fisher's).

The other options for the BDTs were optimized in preliminary studies with MinNodeSize=4 and MaxDepth=4 being optimal.

In the following sections, the results from the calorimeter, isolation and track BDTs are presented.

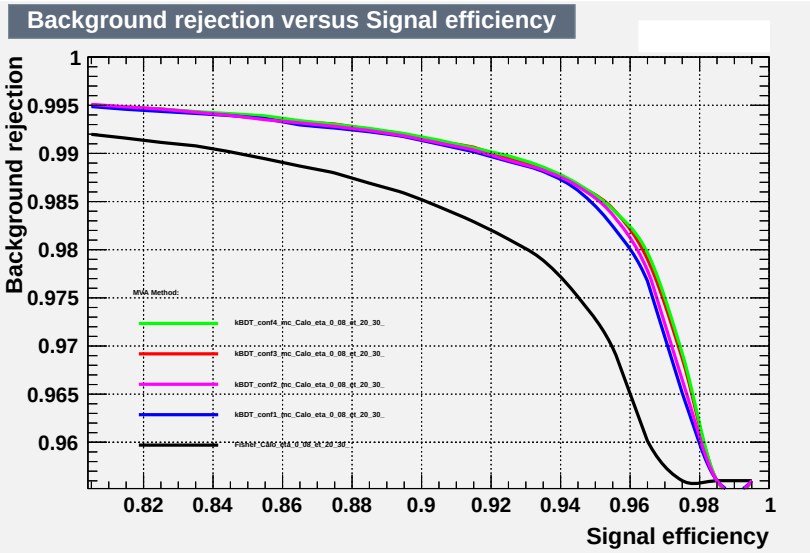


Figure 2.6: The ROC curves for the test sample after training are shown for the $|\eta| < 0.8$ and $20\,\text{GeV} < E_T < 30\,\text{GeV}$ bin. The black curve is a Fisher, the blue is a BDT with 50 trees, magenta 100 trees, red 200 trees, green is 400. All BDTs perform well in terms of separating electrons from non-electrons.

2.3.3 Calorimeter BDT

In Figure 2.7, the distributions of the calorimeter BDTs for four chosen PS bins are shown. For $|\eta| < 0.8$ and $E_T < 20$ GeV, there is a small peak in the background region indicating that the signal is not completely pure. For the crack region the tail from background is long within the signal region due to the lower discrimination power of the calorimeter in the crack region.

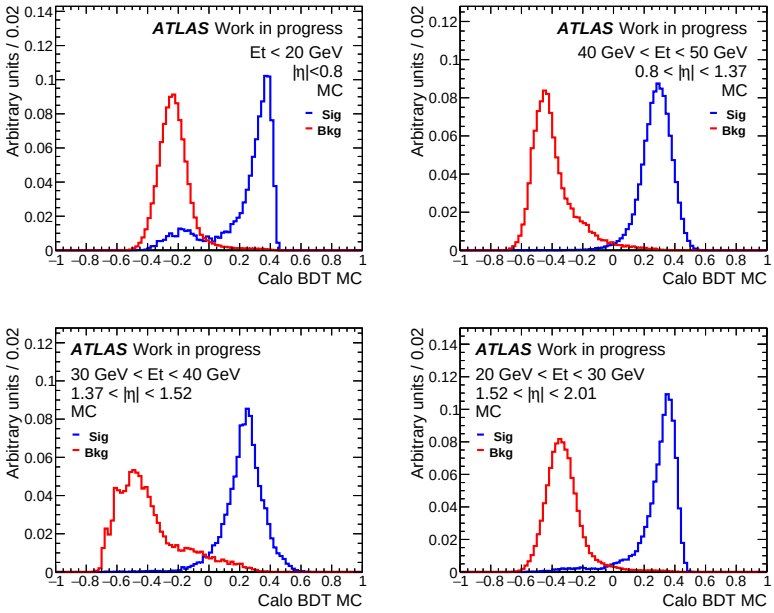


Figure 2.7: Distribution of the calorimeter BDT scores for four different bins. For low energy and central electrons, the signal sample is not completely pure.

The corresponding ROC curves are shown in Figure 2.8. The ROC curve is shown as, Bkg acc. $= 1-$ Bkg rej. (Figure 2.6), on the y-axis. This will be the case throughout this thesis.

Notice how the small subset of mis-labeled events gives the ROC curve a different shape for the $|\eta| < 0.8$ and $E_T < 20\,\text{GeV}$ case. The actual background acceptance for that PS bin is smaller than what the ROC curve shows due to the mis-labeled events. Also, the lacking performance of the calorimeter in the crack region is clear.

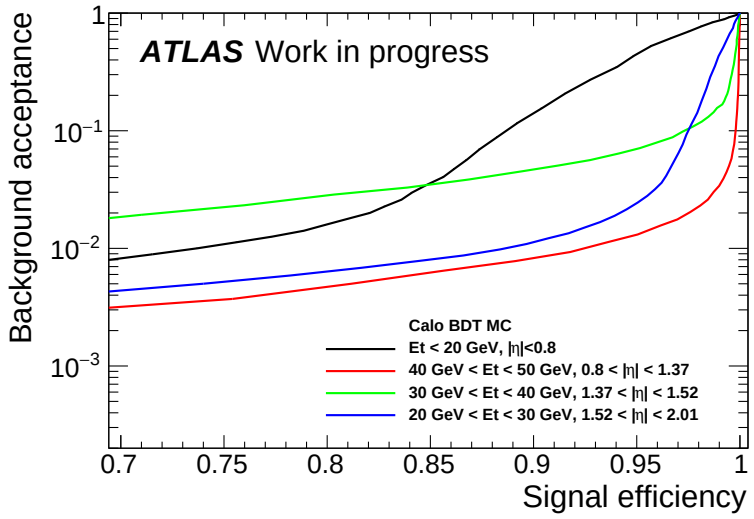


Figure 2.8: The corresponding ROC curves for Figure 2.7. Notice how the background contamination in the signal sample changes the shape of the ROC curve for low energy central electrons. For signal efficiencies above 0.85, the ROC curve gives a conservative value for background acceptance.

2.3.4 Isolation BDT

In Figure 2.9, the results for the isolation BDTs are shown. As with the calorimeter scores the isolation also shows that there are mislabeled events in signal for $|\eta| < 0.8$ and $E_T < 20$ GeV. The isolation distributions are not as smooth as the calorimeter distribution. This can be due to some of the variables often being zero. Figure 2.10 shows the corresponding ROC curves. The isolation performance in the crack region does not suffer as the calorimeter performance.

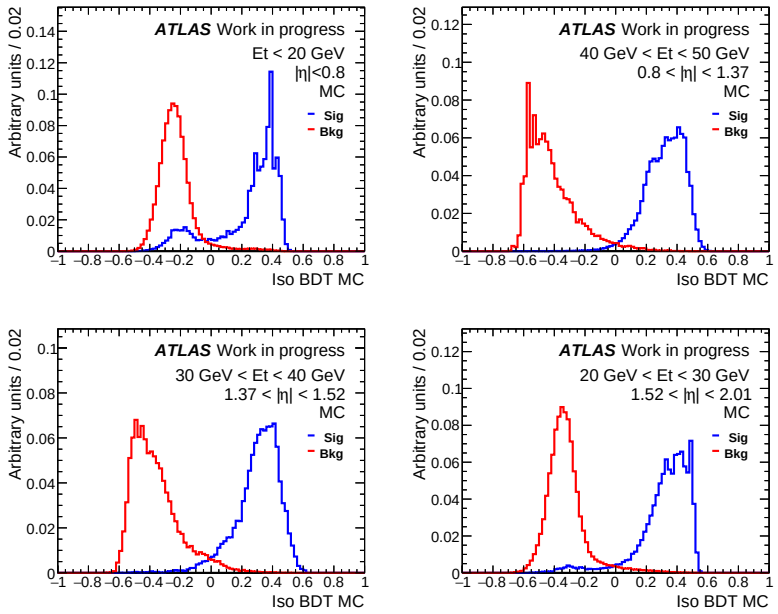


Figure 2.9: Distribution of the isolation BDT scores for four different PS bins. For low energy and central electrons, the signal sample is not completely pure.

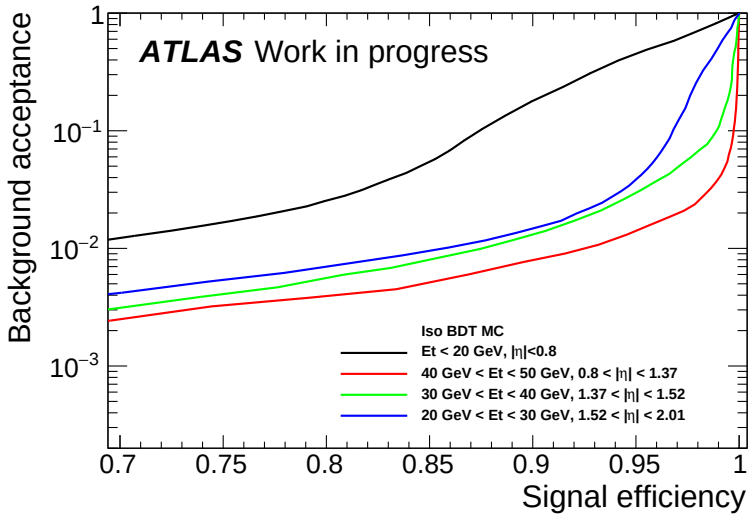


Figure 2.10: The corresponding ROC curves for Figure 2.9. Notice how the crack (green) is slightly better than the end-cap (blue). This is opposite for the calorimeter because of the crack.

2.3.5 Track BDT

In Figure 2.11, the results for the track BDTs are shown. In Figure 2.12, the corresponding ROC curves are shown. The ROC curves shows that the inner detector has less discriminating power than the calorimeter except in the crack region. For the top right plot, the background has two slight peaks. It is not clear why, but the mix of $JF17/35/50$ can be the cause of this. As mentioned in section 1.5, the TRT (track related variable) conditions are different for the $JF50$ simulation compared to the two others.

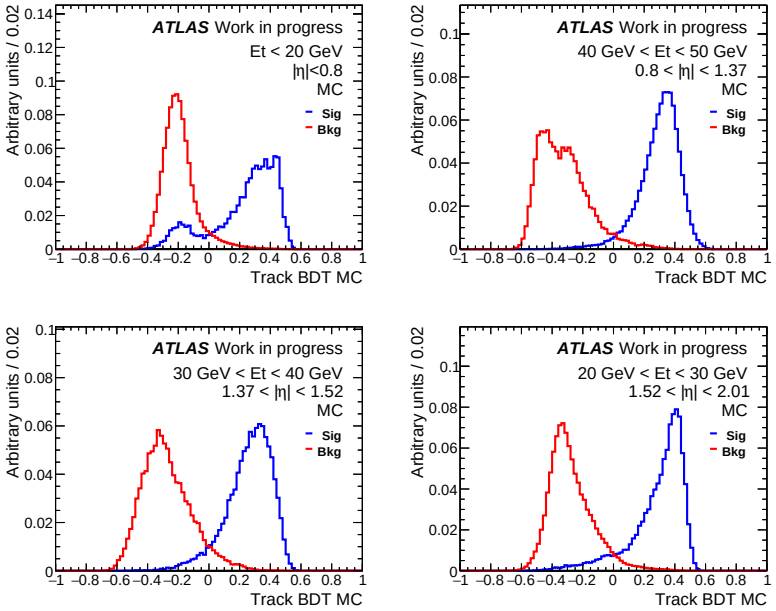


Figure 2.11: Distribution of the track BDT scores for four different bins. For low energy and central electrons, the signal sample is not completely pure. Note two slight peaks for background in top right plot.

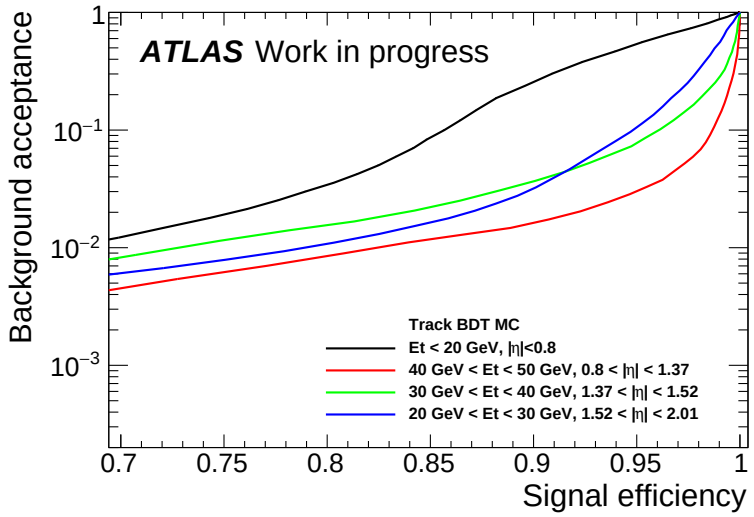


Figure 2.12: The corresponding ROC curves for Figure 2.11. Notice how the separations are worse compared to the calorimeter and isolation. This shows that the inner detector is less powerful than the calorimeter at detecting electrons.

2.3.6 Combining calorimeter and track BDTs

After construction of the 3 classifiers, the calorimeter BDT and the track BDT are combined with a Fisher's discriminant. The distributions of the two BDT scores for two selected bins are shown in Figure 2.13 for signal and background. The correlations between the two variables varies for the signal (positive BDT score) and background (negative BDT score) region depending on the PS bin, but as no significant non-linear features are visible it is reasonable to use a Fisher's discriminant to combine the two. For background $40 < E_T < 50$ GeV and $0.8 < |\eta| < 1.37$, the two slight peaks are seen with the track score but not with the calorimeter.

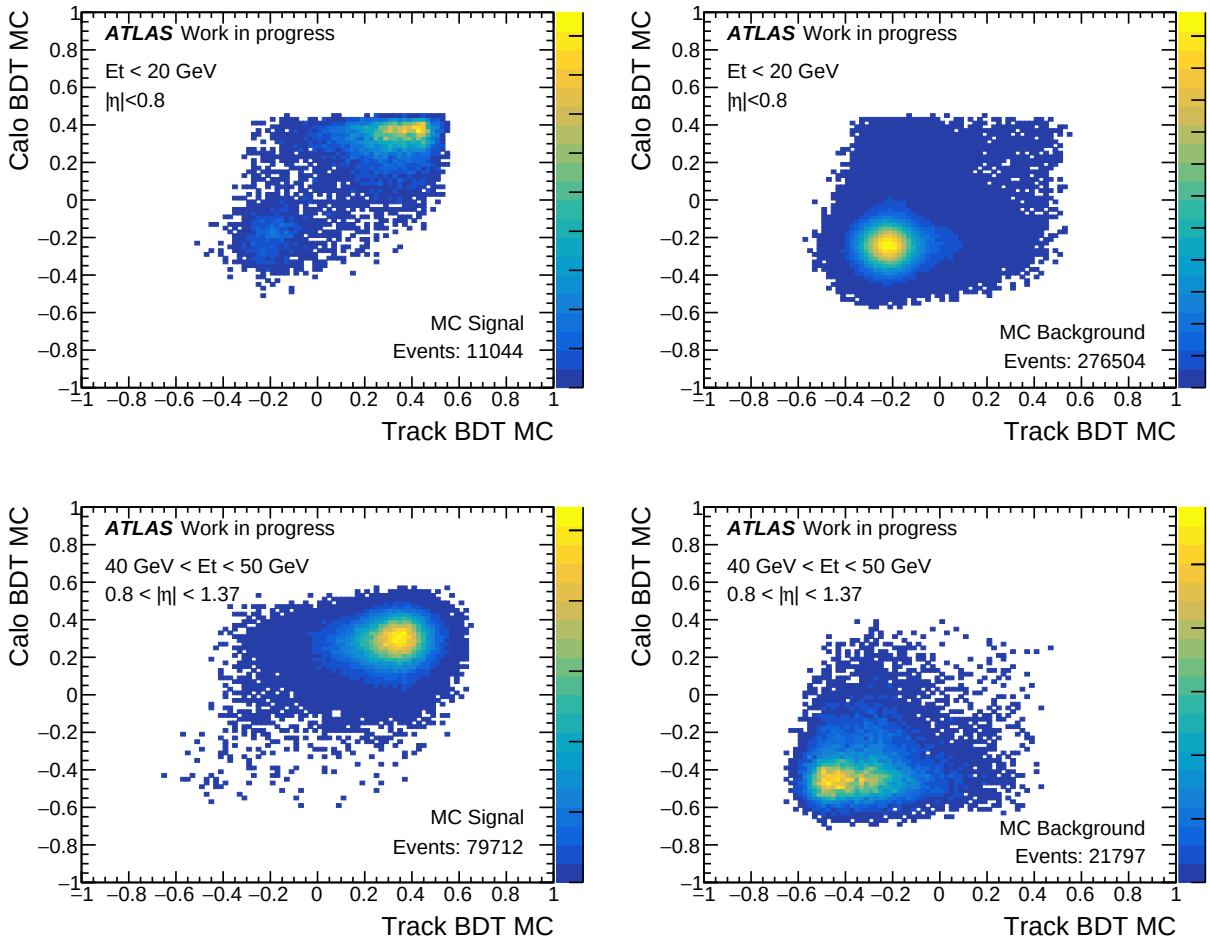


Figure 2.13: Distribution of the calorimeter and tracking BDT scores. The correlations in signal and background region are small and linear.

The distribution of the Fisher's scores are shown in Figure 2.14. The corresponding ROC curves are shown in Figure 2.15. The ROC curves for the LH for each of the four PS bins are shown with a dashed line. For the low energy and central bin, the combined BDTs perform worse than the LH, possibly due to mis-labeled signal events and yields a conservative background acceptance above 85% signal efficiency. For the crack region the relative improvement is the largest. Since the crack region has a more complicated geometry, more non-

linear effects are expected. This is the reason for the BDT based classifier to perform relatively better in the crack region.

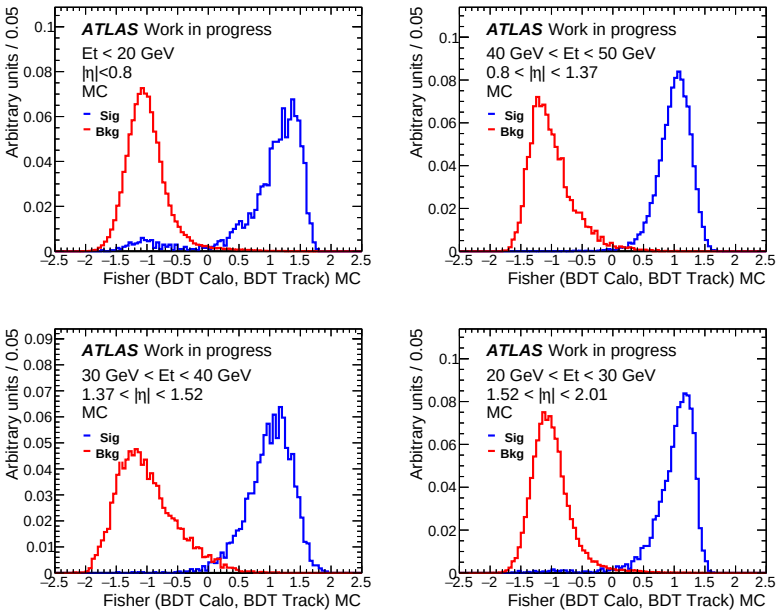


Figure 2.14: The distribution of the Fisher's discriminant.

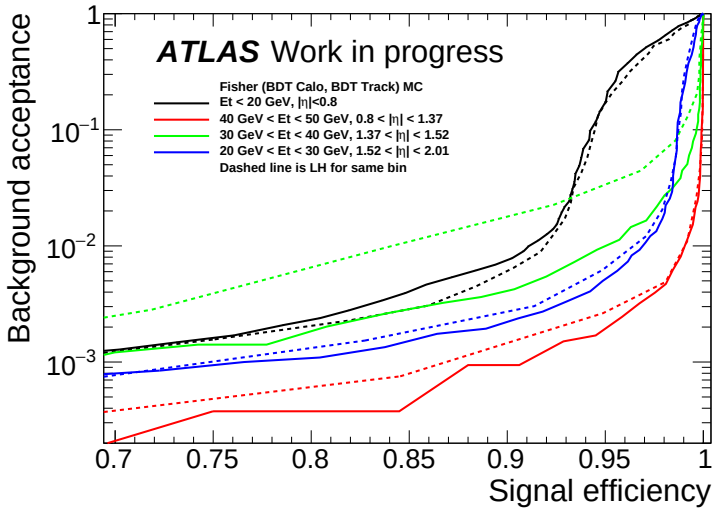


Figure 2.15: The corresponding ROC curves for the Fisher distributions from Figure 2.14 with the LH ROC curves in dashed for the same bins. For $|\eta| < 0.8$ and $E_T < 20\,\text{GeV}$, the LH is doing better than the Fisher's. The biggest improvements are in the crack region.

2.3.7 Results

The results for all 25 PS bins are shown in Figure 2.16, and they are shown as relative performance improvements. The ROC curves are evaluated at 92% signal efficiency, and the relative performance improvements are ratios between background acceptance for the LH and Fisher's. It approximately corresponds to the average medium likelihood efficiency [20]. This is how most results will be reported throughout this thesis. The crack region sees the largest improvements.

Most low energy BDTs perform slightly better than the LH. Since the final goal of electron identification is finding electrons in data, the MC based classifiers have been tested in data. In Figure 2.17, the results from applying the classifier in data are shown. The method of how to apply the classifier in data is presented in Section 2.4. For most PS bins, there are a decrease in performance and specially for the crack region. To overcome this effect data based training has to be used.

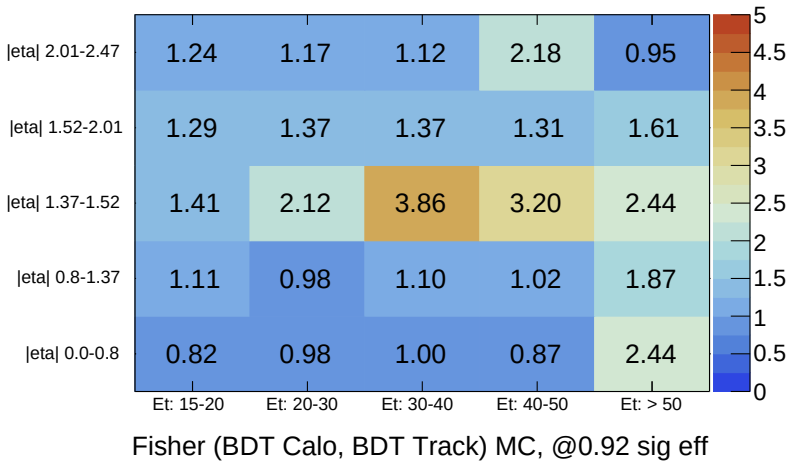


Figure 2.16: Performance of the BDT based classifier relatively to the the LH performance at 0.92 signal efficiency in MC. The performance is calculated as the ratio of background acceptance of the LH compared to the Fisher's.

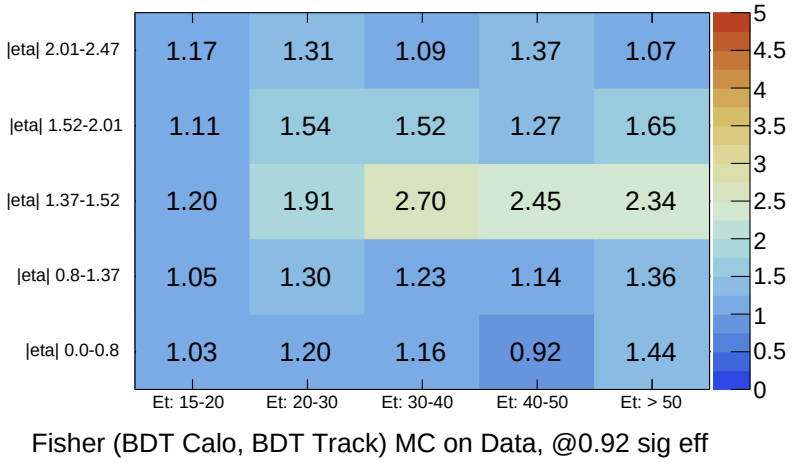


Figure 2.17: Performance of the BDT based classifier relatively to the the LH performance at 0.92 signal efficiency in data. The method on how to evaluate in data is presented in Section 2.4. For most PS bins, the gain from the MC based classifiers are decreasing.

2.4 Data-driven

In this section, a method to make the training process data-driven is presented. The motivation for doing so is the minor inconsistencies between MC and data in the tails of the variable distributions and however much work was put into the MC, it is difficult to get the tails and correlations right in a many dimensional space. As seen in Section 2.3, the MC based classifiers applied in data show some decrease in performance compared to the LH.

For T&P selection, specially for low-energy electrons, many misclassified events are present in the training sample. The main challenge of the data-driven method is to purify the samples such that an ML algorithm can train on the data and discriminate electrons from non-electrons.

2.4.1 Method

In Figure 2.18, an illustration of the implementation strategy for the data-driven method using BDTs is shown. The EGAM1 ($Z \rightarrow ee$) and EGAM7 (fake electrons) samples, derived from data taken during 2016, are used.

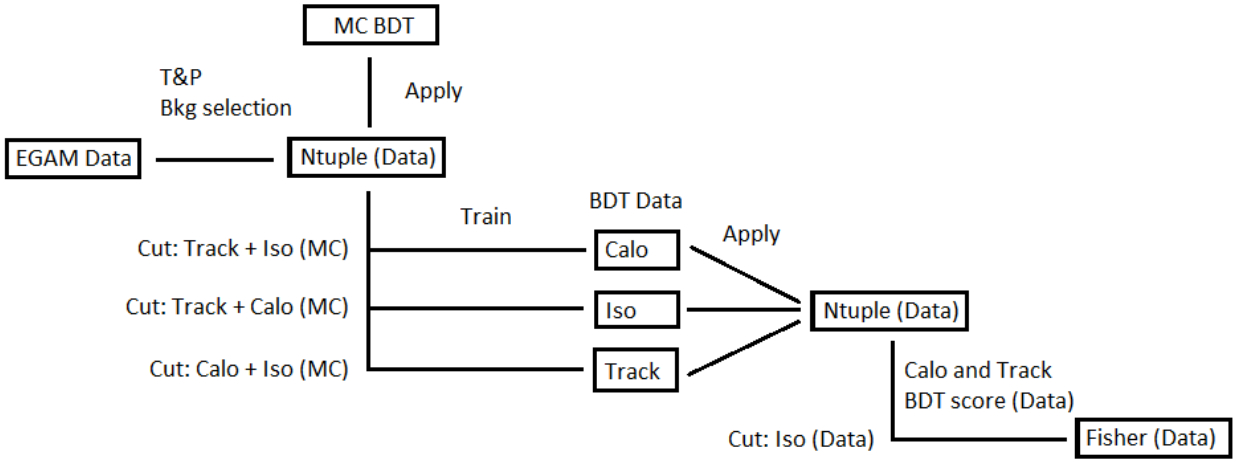


Figure 2.18: Diagram for implementation of BDTs on data is shown. Note how the MC trained BDTs are used for purification. See text for further explanation.

The BDTs trained on MC from section 2.3 are applied on the data Ntuples created from event selection. That gives a calorimeter, isolation and track BDT score for each event based on the MC trained BDTs. From these scores, it is possible to cut on two of them to create a more pure data sample for training of the third sub-classifier. This is done for each sub-classifier. One further iteration where the data trained sub-classifiers are applied on data for purification would be of interest to check if the methods converges in the number of electrons and background events seen by the sub-classifiers in each sample. This was not done in this thesis though. Another approach similar to [24] could be used. Instead of training on labeled data, samples where the average number of signal and background events are known, can

be used for training. This method yield almost as good results as fully supervised learning methods and mis-labeled events are of no concern.

For example, for training of the calorimeter BDTs in data, if both the track and isolation BDTs indicates the particle at hand in the signal (background) sample is an electron (non-electron), it is likely that the particle is signal (background). If this is not the case, the event will not be part of the training sample, and thereby most mis-labeled events are removed from the signal (background) samples and only few correctly labeled events are lost. Ideally, no bias is introduced in the electron and background samples when cutting on the BDT scores. A bias could be introduced if the BDT scores are correlated. The correlations and the effects of cutting on two sub-classifiers will be studied later in Section 2.4.9.

After training of the data BDTs, they are applied on a different data Ntuple. Before training of the Fisher's discriminant, a cut on the data BDT isolation score is applied to purify the samples. The calorimeter BDT and track BDT are then combined into a Fisher's. Since the Fisher's contains both the calorimeter BDT and track BDT, only isolation is left to use for purification. When only cutting on the isolation BDT, the purity will not be as high as before, but a Fisher's discriminant is likely to be less sensitive to mis-classified events.

2.4.2 Purification of data

The cut values for the purification of data on the MC trained BDT scores have been picked for every bin and for every classifier to be $c_s = 0.025$ for signal samples and $c_b = -0.03$ for background samples. These values have been optimized to give clean samples without cutting hard on the sub-classifiers. If the cut is too tight, statistics are lost and the effect from cutting in terms of biasing the sample, if any, is increased.

The invariant mass of the T&P pair is shown in Figure 2.19 before cutting on the Z mass (T&P cut 11).

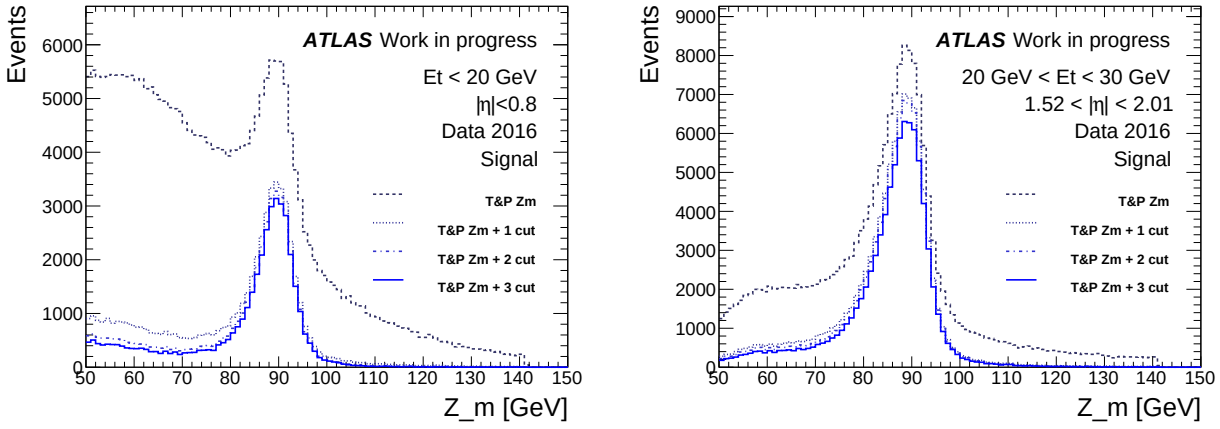


Figure 2.19: Distribution of the invariant mass of the T&P pair in data before and after cutting on the BDTs scores from MC for different PS bins.

For $|\eta| < 0.8$ and $E_T < 20$ GeV, the number of mis-labeled events is large. Notice the difference from the first to second cut is small and from the second to the third cut is very small. This implies that cutting on two classifiers is sufficient to get very clean samples. In Figure 2.20, the distributions of the BDT scores from MC trained classifiers applied on data are shown, before and after cutting on the two other BDT scores for $|\eta| < 0.8$ and $E_T < 20$ GeV. In the signal case, the two cuts remove the background events from signal, leaving out the signal peak. For background, the cuts remove the peak in the calorimeter case, and removes most of the tails towards the signal region for the two other cases. The background sample in the calorimeter has a peak in the signal region before cutting, but for isolation and track no peak is seen. This can be explained by converted photons which becomes electrons and therefore behave like electrons in the calorimeter, but do not have electron-like tracks through all of the ID, and the origin of the photons can be from non-isolated events.

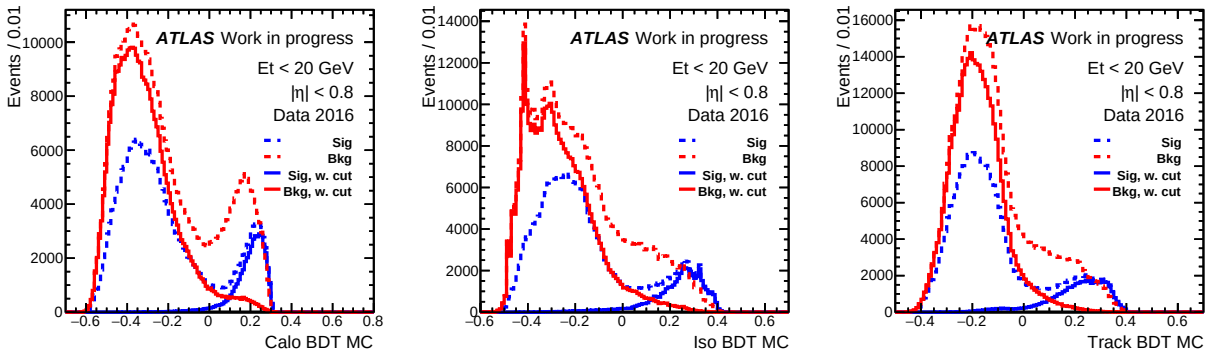


Figure 2.20: Distribution of the MC trained BDT scores for $|\eta| < 0.8$ and $E_T < 20$ applied on data. The dashed line is the distribution of signal and background before cutting on the two sub-classifiers. The full line is after the cuts.

Purities & Efficiencies

For each PS bin, estimates of the purities of signal and background samples are presented. For one PS bin, the number of signal events in a signal sample is calculated the following way,

$$N^{sig,sig} = \frac{N^{sig,sig}_{3cuts}}{\epsilon^{sig,sig}_{calo}\epsilon^{sig,sig}_{iso}\epsilon^{sig,sig}_{track}}, \tag{2.1}$$

where, $\epsilon^{sig,sig}_{calo}$ is the signal efficiency of the calorimeter sub-classifier on a signal sample. It is calculated as the amount of signal that is kept when cutting on the calorimeter after cutting on the two other sub-classifiers. $N^{sig,sig}_{3cuts}$ is the number of events after cutting on all three sub-classifiers. As seen in Figure 2.20, and in Figure 2.19, the purities of signal and background are high after applying two cuts. Thus, it can be assumed that after two cuts, the samples are close at being pure. The number of mis-labeled events in a signal sample is

calculated the following way,

$$N^{bkg,sig} = N^{sig} - N^{sig,sig}. \tag{2.2}$$

N^{sig} is the number of events in a signal sample. The amount of signal and background in a signal sample for training a sub-classifier can be calculated using the efficiencies of the sub-classifiers. The total number of signal events in a signal sample is,

$$N^{sig,sig}_{calo} = N^{sig,sig}\epsilon^{sig,sig}_{iso}\epsilon^{sig,sig}_{track}. \tag{2.3}$$

The amount of background in a signal sample is estimated using the acceptance calculated for background at $c_b = 0.025$. The estimated number of background events in a signal sample is,

$$N^{bkg,sig}_{calo} = N^{bkg,sig}(1-\epsilon^{bkg,sig}_{iso})(1-\epsilon^{bkg,sig}_{track}). \tag{2.4}$$

The same procedure is used to estimate the purities in the background samples but with a cut in $c_b = -0.03$.

$$N^{bkg,bkg} = \frac{N^{bkg,bkg}_{3cuts}}{\epsilon^{bkg,bkg}_{calo}\epsilon^{bkg,bkg}_{iso}\epsilon^{bkg,bkg}_{track}}, \tag{2.5}$$

$$N^{sig,bkg} = N^{bkg} - N^{bkg,bkg}, \tag{2.6}$$

$$N^{bkg,bkg}_{calo} = N^{bkg,bkg}\epsilon^{bkg,bkg}_{iso}\epsilon^{bkg,bkg}_{track}, \tag{2.7}$$

$$N^{sig,bkg}_{calo} = N^{sig,bkg}(1-\epsilon^{sig,bkg}_{iso})(1-\epsilon^{sig,bkg}_{track}). \tag{2.8}$$

Figure 2.21 shows the purities from the selection of signal and background and the obtained purities after cleaning the samples for training of the sub-classifiers. The purities reached are $> 99\%$ for most bins for signal and background. Except the crack region with low energy where the purity for signal is (95%). That PS bin could be treated separately with cuts optimized for that region to gain higher signal purity. This has not been done to keep uniformity of the analysis. In Figure 2.22, the efficiencies for signal and background for each sub-classifier are shown. For most bins, the efficiencies are high $(85-95\%)$ and they also result in pure samples. For the crack region, the efficiencies are lower.

A study of the impact on the performance of the sub-classifiers from mis-labeled data would be interesting. This could be done by truth matching candidates in MC demanding the signal to be electrons and add different amounts of background into the signal sample and train the sub-classifier for each level of impurity. Depending on the results, the purities for training could be optimized by changing the cut values, if the study shows that the sub-classifiers suffer at present levels of mis-labeled events.

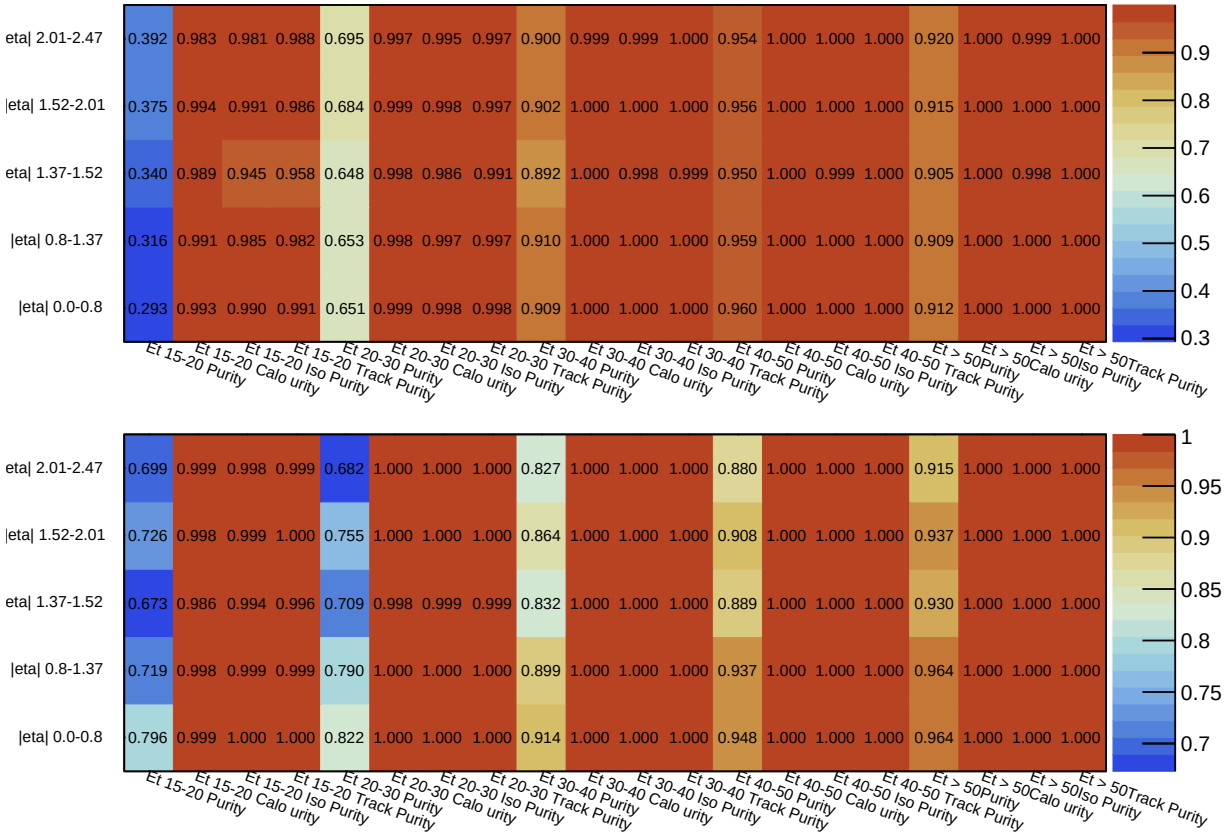


Figure 2.21: Top: Signal purity from T&P (first column) and the purities for training of each sub-classifier (next three columns) for each E_T bin. Bottom: Background purities with the same structure. For signal $E_T < 20$ GeV and $1.37 < |\eta| < 1.52$, the purity for training of isolation and track is low (95%). They suffer from the lack of discriminating power of the calorimeter in the crack region. For the rest of the cases the purities for signal and background are $> 99\%$. Note different scales.

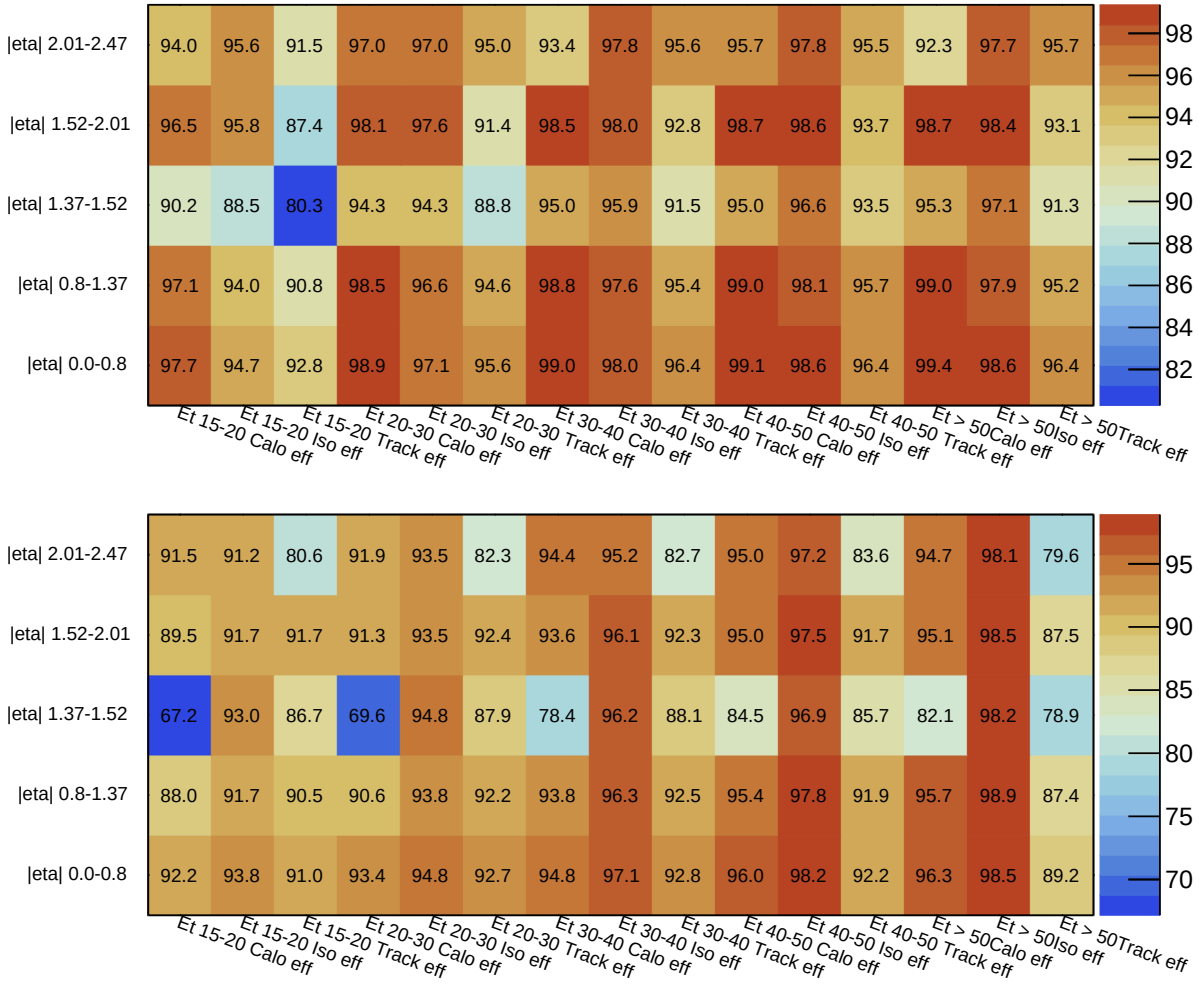


Figure 2.22: Top: Efficiency of the three sub-classifiers for signal (%). Bottom: Efficiency of the three sub-classifiers for background (%). Note different scales.

2.4.3 BDT configuration in TMVA for data training

As for the MC case, the training configurations of the BDTs for the data case are the same for each sub-classifier and for each PS bin. The only option that differs from the MC case is the number of trees. In Figure 2.23, the performances of different trees are shown. From red to cyan is 100, 200, 400, 800 and 1600 trees, respectively. There is no difference in performances from 800 to 1600 trees and therefore the 800 trees option is chosen.

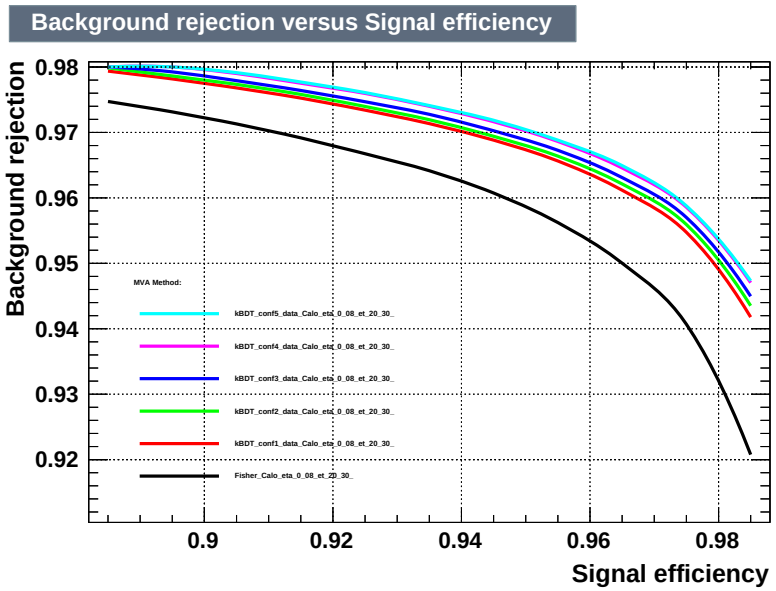


Figure 2.23: For $|\eta| < 0.8$ and $20\,\text{GeV} < E_T < 30\,\text{GeV}$, the ROC curves for test samples after training. The black curve is a Fisher's, red is a BDT with 100 trees, green 200 trees, blue 400 trees, magenta 800 trees and cyan 1600 trees. All of them perform well in terms of separating electrons from non-electrons. No improvement is seen from 800 trees to 1600 trees.

Different types of boosting with and without bagging[1] has been examined. In Figure 2.24, the results are shown. Both bagging and random (UseRandomisedTrees option in TMVA [15]) has been tried with different boosting algorithms. The UseRandomisedTrees option uses a subset of the input variables for the training of each tree instead of all variables. The gradient boosting (green) performs equally well with the random gradient boosting, and has therefore been used for the entire analysis together with adaBoost. The results from gradient boosting will be shown later. The following presented results are based on adaBoost.

[1] Bagging is a technique that samples a random subset of data to train each tree. It is a technique to avoid over-training [15].

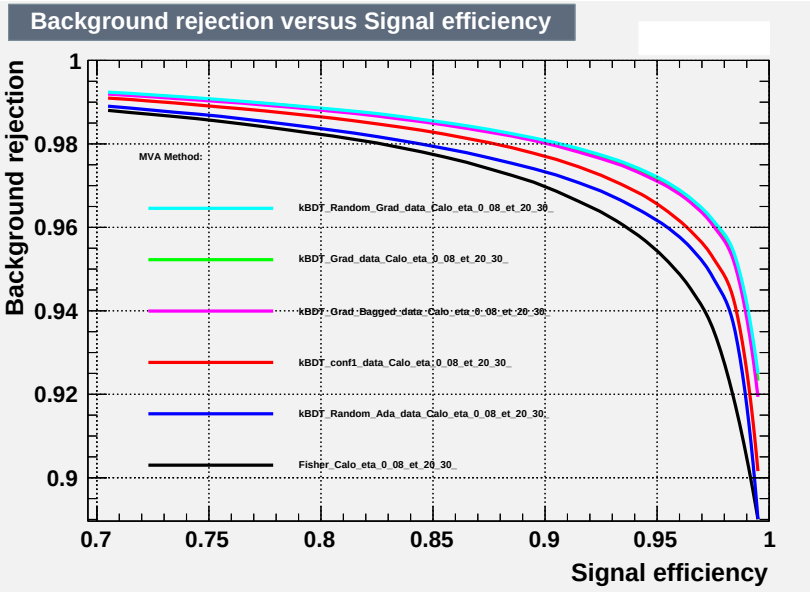


Figure 2.24: Performance of different boosting configurations for $|\eta| < 0.8$ and $20 < E_T < 30$ GeV. The red curve is adaBoost. The green curve is Gradient boosting, magenta is gradient combined with bagging and cyan is gradient with UseRandomisedTrees. Blue is UseRandomisedTrees combined with adaBoost. The green is beneath the cyan, and is performing equally well. Based on that, gradient boosting without bagging or the use of random subset of variables has been implemented for the analysis.

2.4.4 Calorimeter BDT

In this section, the results of the calorimeter BDTs are presented. After applying the isolation and track cuts, the calorimeter data BDTs are trained for each PS bin. The results are shown in Figure 2.25 for four different bins after cutting on the two other sub-classifiers data based values. For $1.37 < |\eta| < 1.52$ and $30 < E_T < 40$ GeV, the calorimeter is not separating electrons from non-electrons well due to the crack region. For all four PS bins, the tail from the background is long and ends well within the signal region. For the MC case, the tails were smaller and not as electron-like as with background from data.

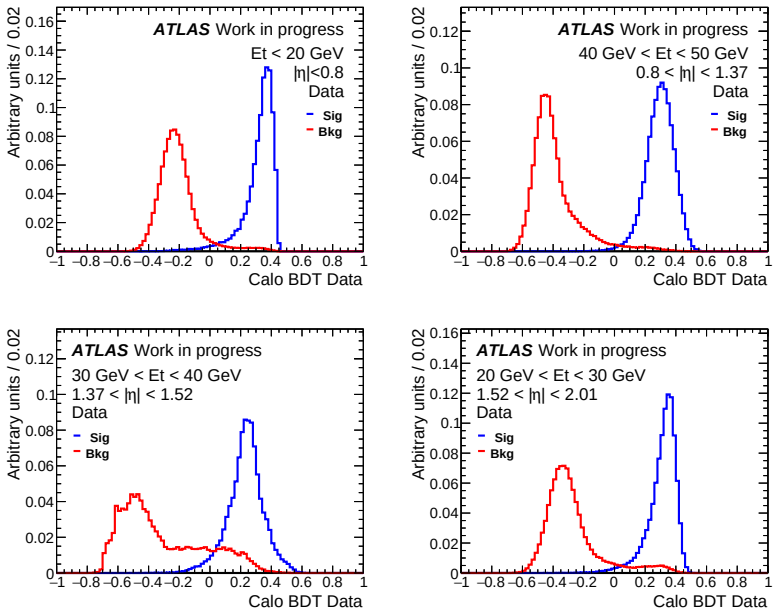


Figure 2.25: Distribution of the calorimeter BDT scores for four different bins. They are shown after cutting on the isolation and track BDT data based scores.

The corresponding ROC curves are shown in Figure 2.26. The crack region is performing poorly relatively to the other PS bins as expected.

2.4.5 Isolation BDT

In Figure 2.27, the results for the isolation BDTs are shown the same way as for the calorimeter BDTs. The distributions have spikes due to the isolation variables often being zero. In Figure 2.28, the corresponding ROC curves are shown. The ROC curve for high E_T is separating better than any of the sub-classifiers, indicating the isolation criteria is a better discriminator for electrons from non-electrons. However, this is likely an artifact of high energy electrons being more clean already.

2.4.6 Track BDT

In Figure 2.29, the results for the track BDTs are shown the same way as for the calorimeter BDTs. In Figure 2.30, the corresponding

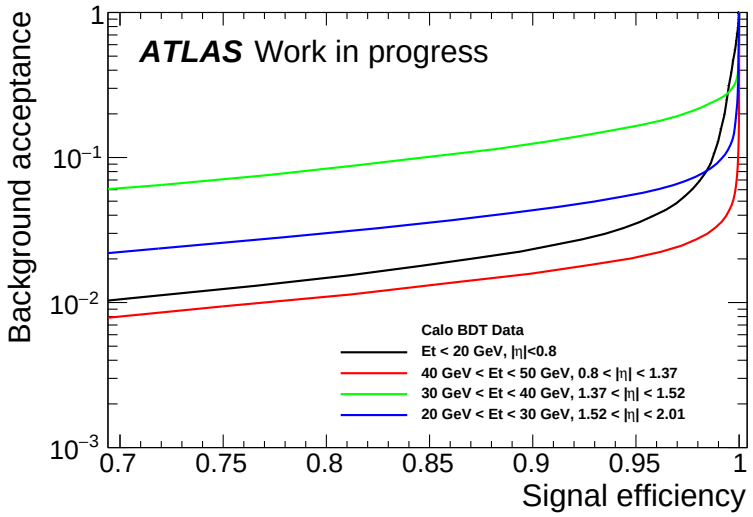


Figure 2.26: The corresponding ROC curves for Figure 2.25. The ROC curve from the crack region is performing poorly compared to the other bins.

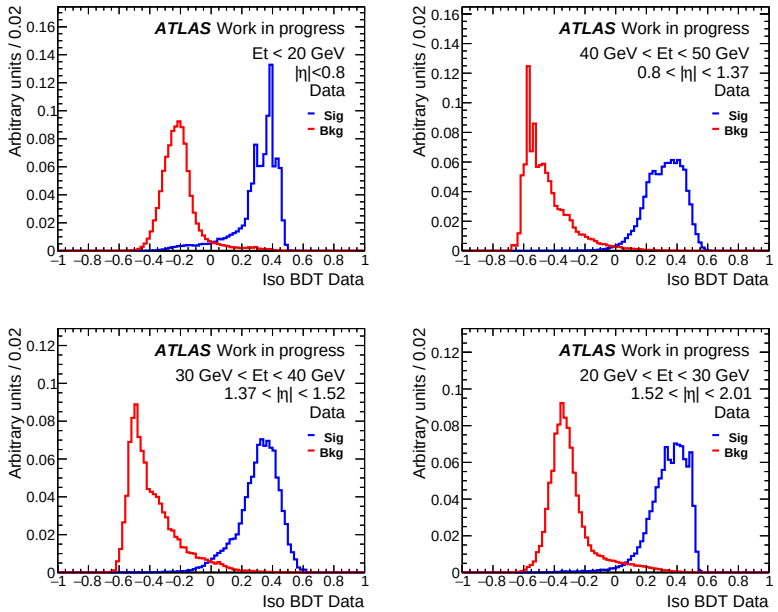


Figure 2.27: Distribution of the data trained isolation BDT scores for four different bins. They are shown after cutting on the calorimeter and track BDT data based scores.

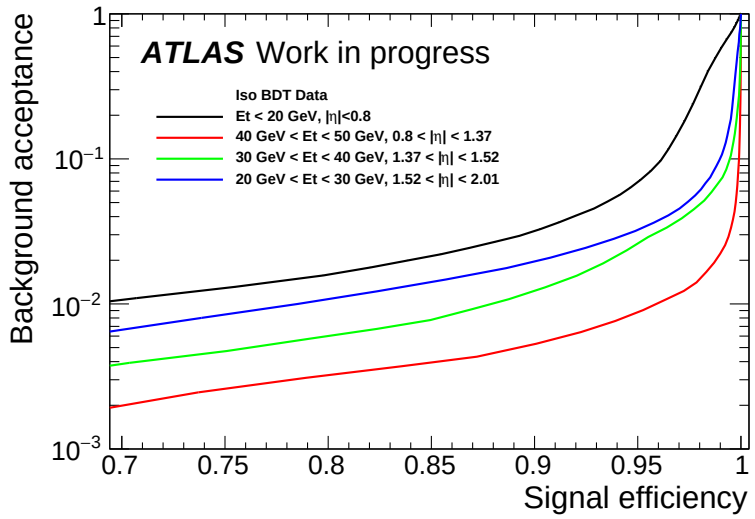


Figure 2.28: The corresponding ROC curves for the distributions in Figure 2.27. For high energy the isolation BDTs becomes more discriminating.

ROC curves are shown. The variation in performances are smaller compared to the calorimeter and isolation performances.

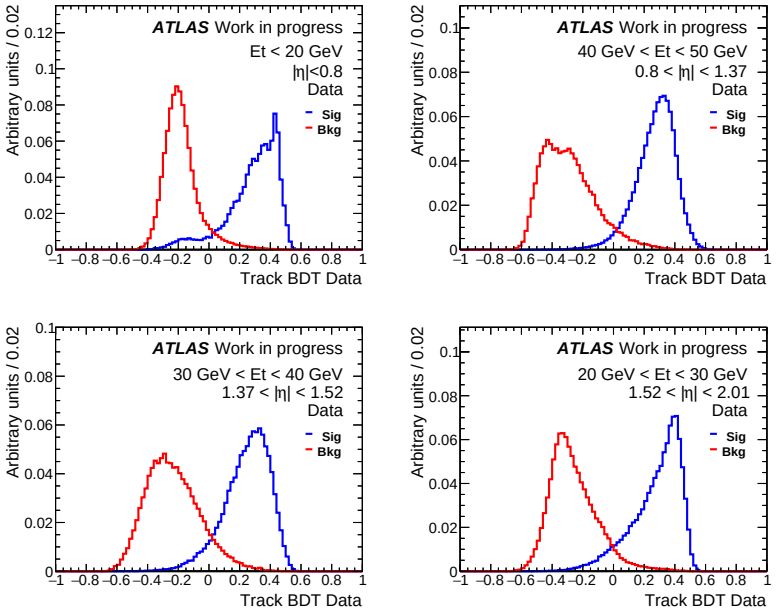


Figure 2.29: Distribution of the data trained track BDT scores for 4 different bins. They are shown after cutting on the calorimeter and isolation data trained scores.

2.4.7 Combining calorimeter and track BDTs

After applying the three sub-classifiers on a new Ntuple with data (test sample) and cutting on the isolation score, the calorimeter and track BDT scores are combined with a Fisher's discriminant as in the MC case. The distribution of the calorimeter and track BDTs in two bins are shown in Figure 2.31 for signal and background, after cutting on the isolation data BDT with same cut values as described earlier

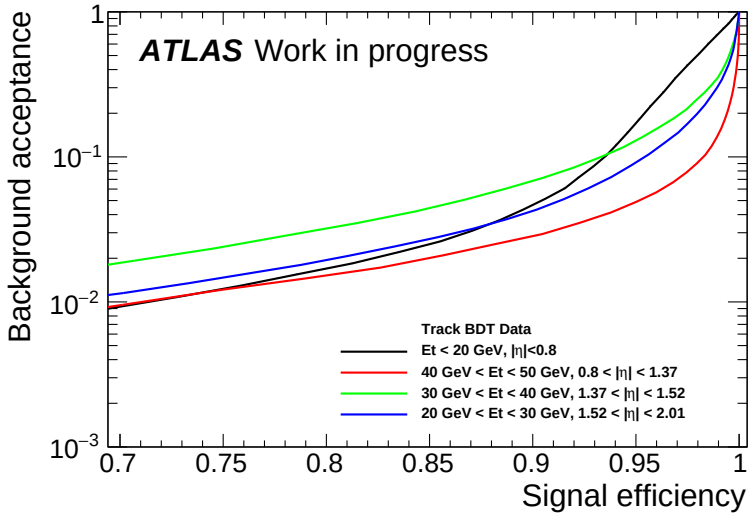


Figure 2.30: ROC curves for the distributions in Figure Figure 2.29. The ROC curves have less variation than for the two other classifiers.

for signal and background.

Combining the two scores with a Fisher's is reasonable given the signal and background distributions. The distributions of the Fisher's discriminants are shown in Figure 2.32. The corresponding ROC curves are shown in Figure 2.33. The isolation cut is not providing samples with the same high purities as for training of the three sub-classifiers. Specially for the low energy cases this is the case. As mentioned earlier, training of the Fisher's should not suffer from mis-labeled events but this has not been tested.

2.4.8 Results

The results for every PS bin are shown in Figure 2.34. For every PS bin, except $0.8 < |\eta| < 1.37$ and $E_T < 20$ GeV, the classifiers are performing better than the LH. The performances are measured at 92% signal efficiency corresponding to medium LH. Figure 2.35 shows results from the gradient boosting based training. In Figure 2.36, the results from adaBoost with additional variables are shown. Later, further details on how to include more variables are presented. Comparing adaBoost and adaBoost with additional variables, a few PS bins show lower performances for the additional variables case which is surprising. The gradient boosting classifier is performing better in the crack region where most mis-labeled events are present.

For all cases, the overall pattern for increasing E_T and $|\eta|$, is increasing relative performances compared to the LH. While the latter was expected, the former was a surprise.

A total measure of improvements is obtained by transforming the signal and the background distributions such that they have a mean of 1 and -1, respectively. This transformation does not change the ROC curve for a particular PS bin. It can be done with two parameters,

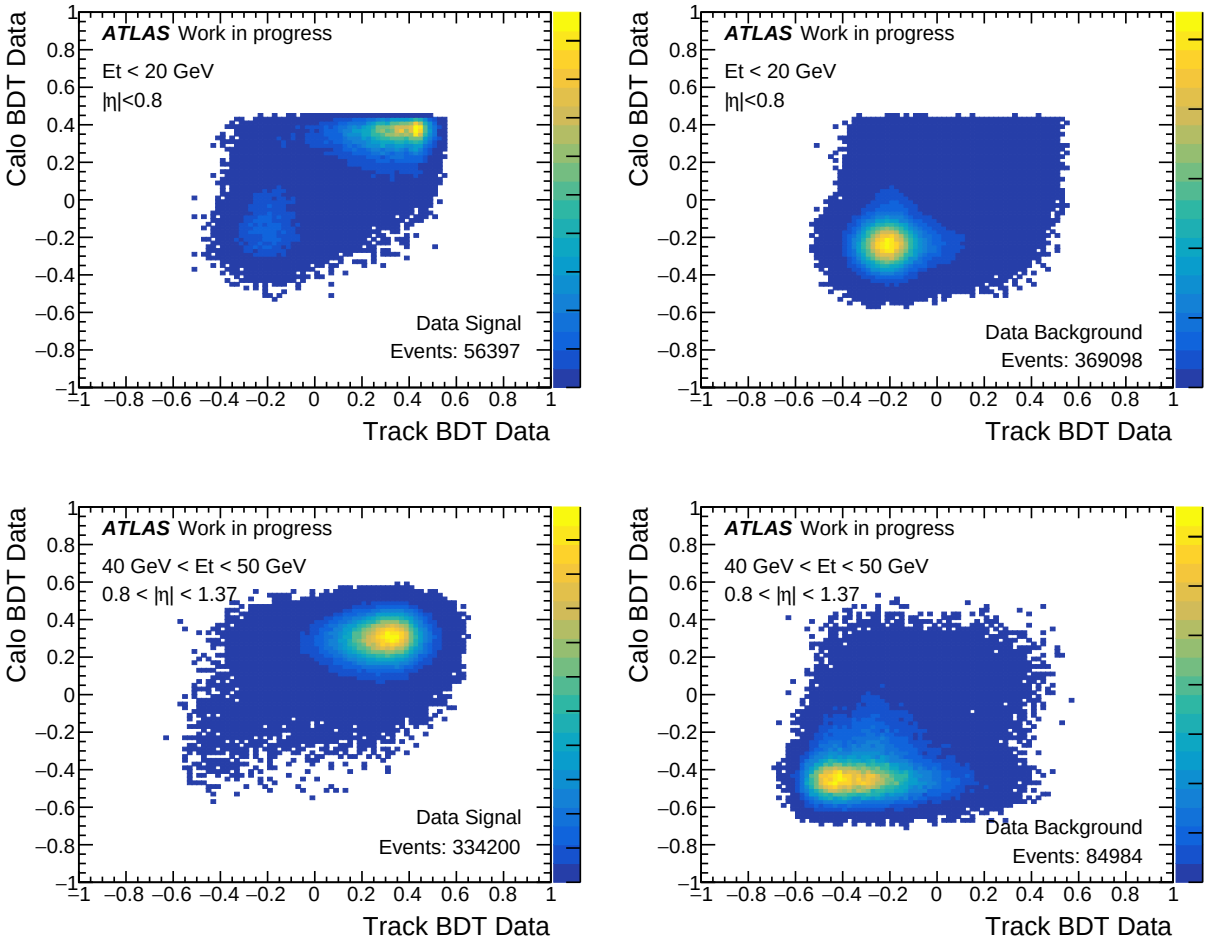


Figure 2.31: Distribution of calorimeter and track scores for signal and background. For both bins, a linear cut is reasonable.

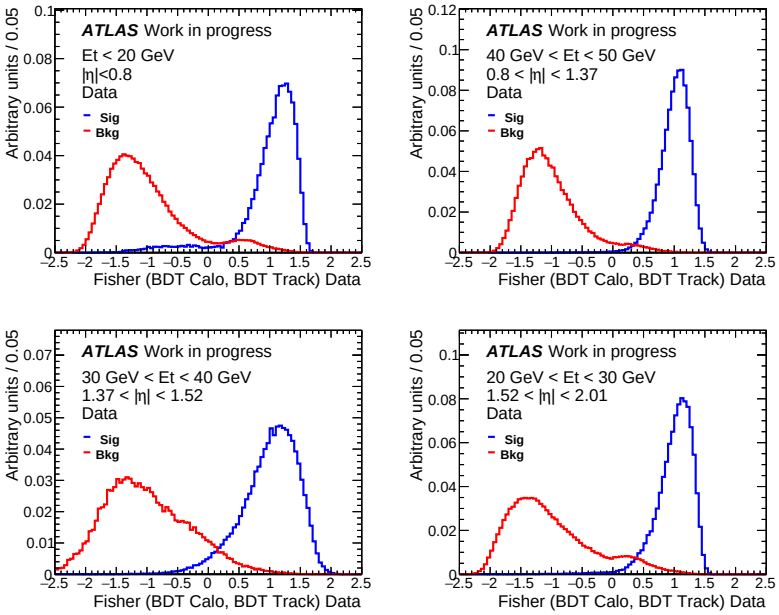


Figure 2.32: Fisher's distribution for signal and background for four different bins after cutting on isolation. The samples are not completely pure, especially for lower energies.

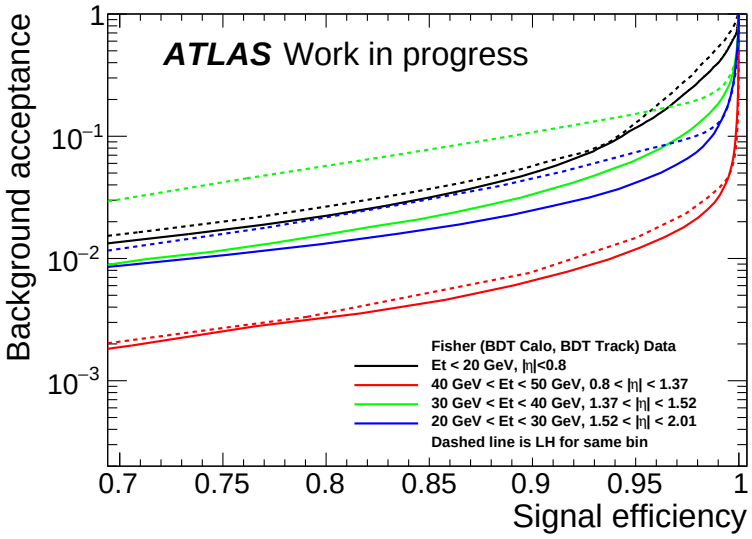


Figure 2.33: The corresponding ROC curves for the distributions from Figure 2.32. The dashed line is the LH for the corresponding bin. The improvement is biggest for the crack-region.

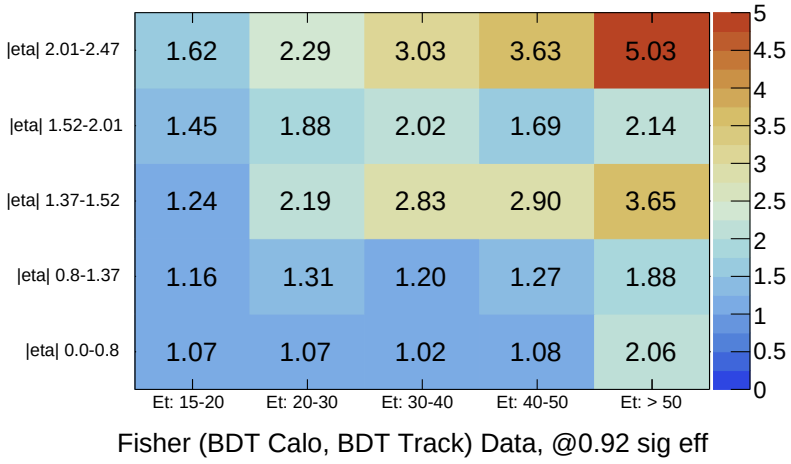


Figure 2.34: The performance of the combined calorimeter and track BDTs in a Fisher's for every bin based on the adaBoost trained BDTs.

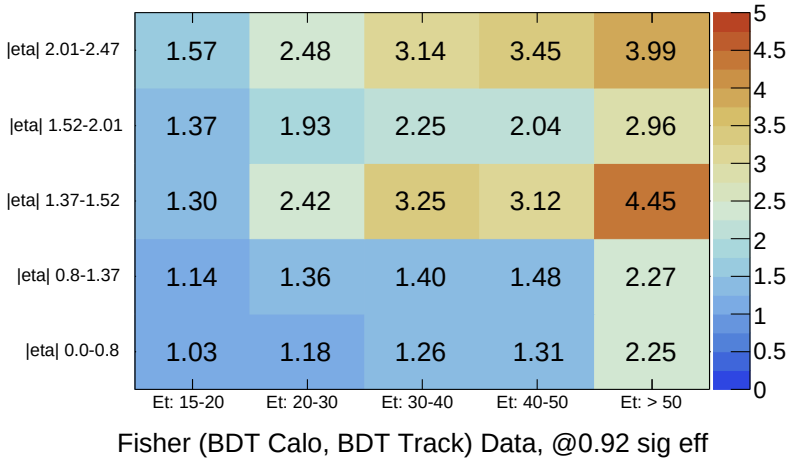


Figure 2.35: The performance of the combined calorimeter and track BDTs in a Fisher's for every bin based on the gradient boosting trained BDTs.

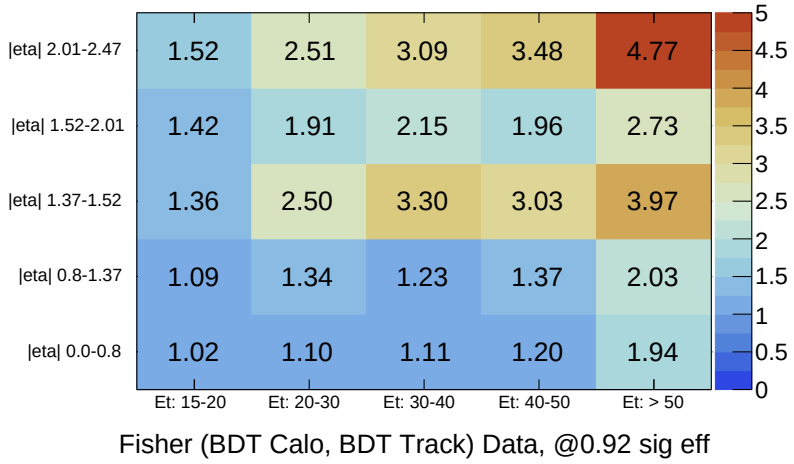


Figure 2.36: Results from adaBoost using additional variables where the events are weighted such that $\langle\mu\rangle$ and η are the same for signal and background. Note that for a few of the PS bins the performances are slightly worse than the classifiers without the additional bins.

namely,

$$1 = B(A + \overline{F}_{sig}), \tag{2.9}$$

$$-1 = B(A + \overline{F}_{bkg}). \tag{2.10}$$

where,

$$A = \frac{1}{B} - \overline{F}_{sig}, \tag{2.11}$$

$$B = \frac{2}{\overline{F}_{sig} + \overline{F}_{bkg}}, \tag{2.12}$$

with $\overline{F}_{sig}$ ($\overline{F}_{bkg}$) being the average of the distribution from the classifier for signal (background). This ensures that every bin has a signal and background mean at the same value, and therefore the distributions can be stacked across different PS bins. The stacked distributions from the transformed adaBoost are shown in Figure 2.37. The corresponding ROC curve is shown in Figure 2.38. The figure also contains the ROC curve from the stacked LH distributions, gradient boosting and adaBoost with additional variables. The ROC curves show that for a signal efficiency at 92%, the total improvement is 94% for adaBoost and 104%for gradient boosting. adaBoost with additional variables has a gain of 109%. The improvements are relatively flat for tighter signal efficiencies. This can be translated into signal efficiency gain of 4-5% at the same background acceptance.

This transformation is the simplest way of adding the distributions for all the PS bins. And for a different transformation a different ROC curve might be obtained. Another way could be to change the distribution such the 92% of signal would be at 1. The first one was chosen for its simplicity. The total ROC curve can vary depending on the chosen transformation.

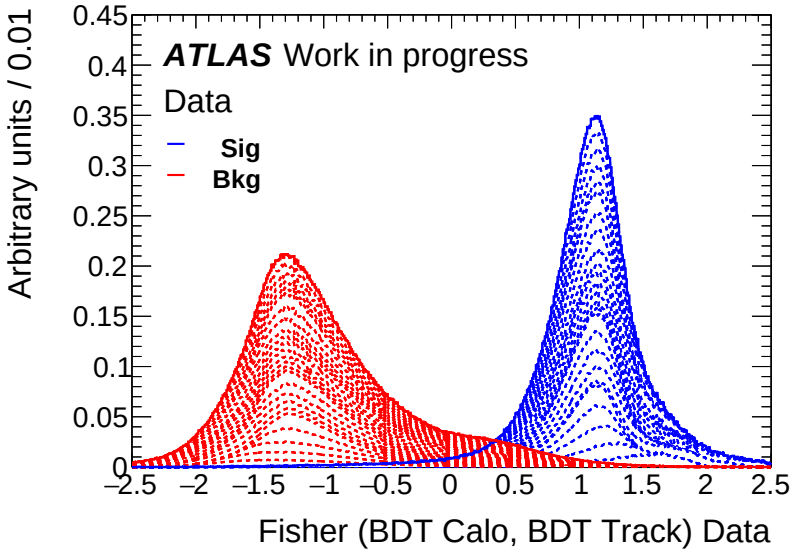


Figure 2.37: Stacked distribution for each PS bin transformed such that the mean for signal and background is 1 and -1, respectively.

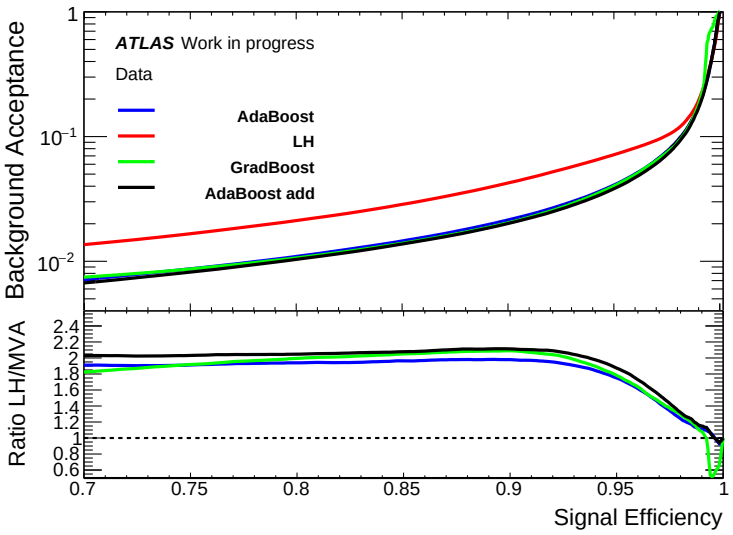


Figure 2.38: ROC curves for LH, adaBoost, gradient boosting and adaBoost with additional variables. All distributions from each bin in $|\eta|$ and E_T are transformed to make a stacked distribution, see text for details. The overall relative performance gain at 92% signal efficiency is 94% for adaBoost and 104% for gradient boosting and 109% for adaBoost with additional variables.

2.4.9 Effects of correlations between sub-classifiers

The linear correlation between the sub-classifiers for each PS bin shown in Figure 2.39 after purification in data. The linear correlation between calorimeter and track, and track and isolation are small. For calorimeter and isolation the correlations are larger.

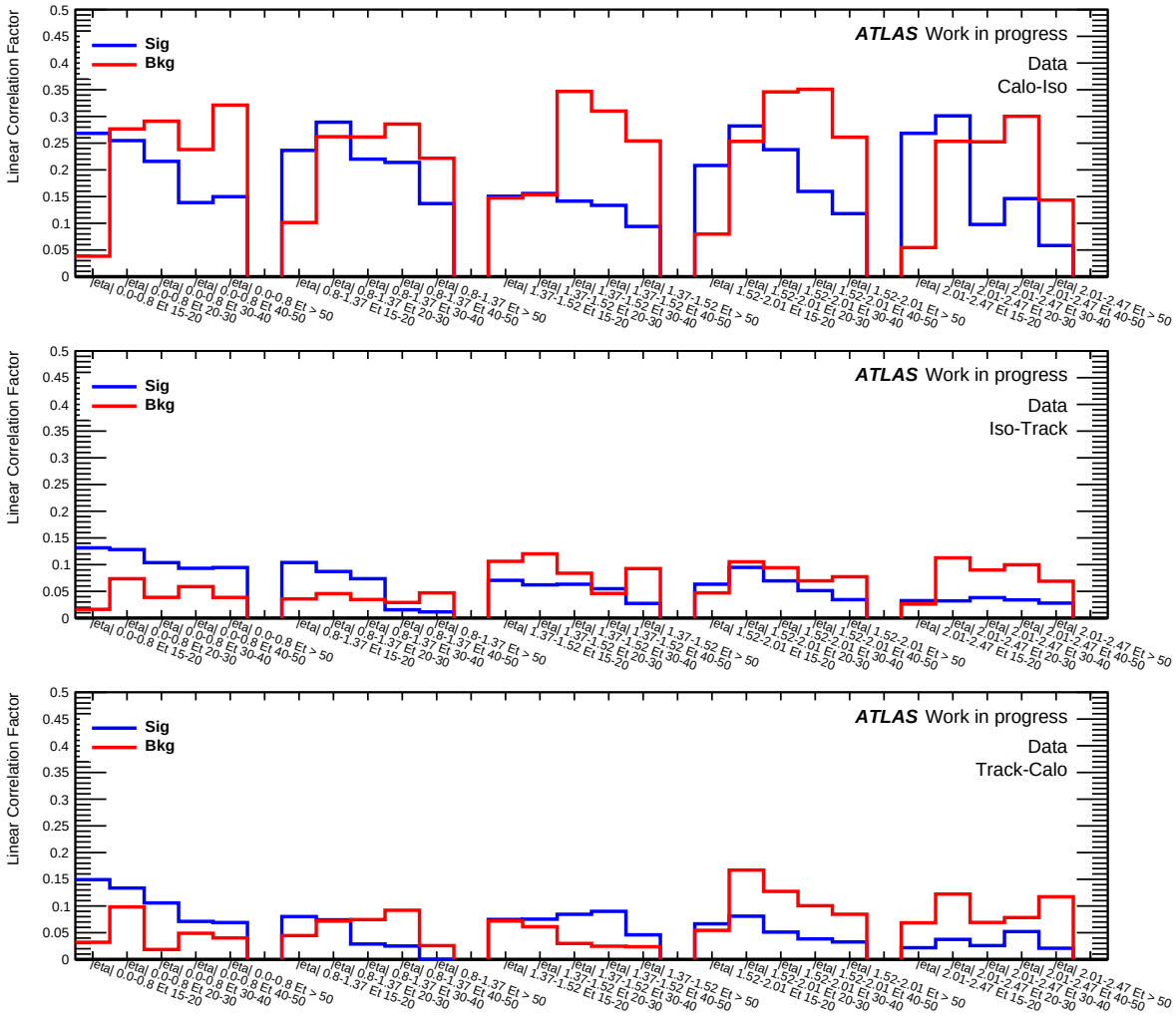


Figure 2.39: Linear correlation between sub-classifiers after cutting on all three sub-classifiers. Top: Linear correlation between the calorimeter BDT score and isolation BDT scores. Middle: Linear correlation between isolation and track. Bottom: Linear correlation between calorimeter and track.

The effect from the purification process and the correlations between the sub-classifiers can be studied in MC for signal and background, where the truth of the particles are known. The method from Figure 2.5 (MC method) and Figure 2.18 (data-driven method) have both been applied on MC to see the differences in performance. If the performances of the two cases are the same, it implies that the purification process does not introduce any biases, and specially that the correlations between the calorimeter and isolation sub-classifiers does not change the results. The results of the two methods are shown in Figure 2.40. It is shown for the additional variables case and it is

the background acceptance at 92% signal efficiency.

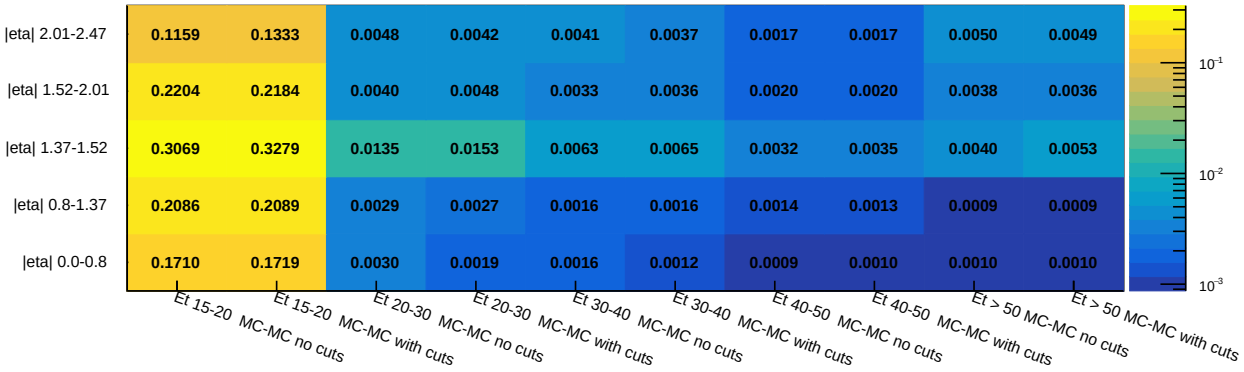


Figure 2.40: Background acceptance at 92% signal efficiency for a classifier following the pure MC method from section 2.3 (no cuts) and a classifier following the data-driven method from section 2.4 (with cuts). Both of them are trained on MC. The results are similar, and the performances of the classifiers does not change significantly, implying that cutting on two sub-classifiers do not have an effect on the results when the samples are clean.

The background rejection at 92% signal efficiency is similar for the two methods. For 10/25 PS bins the performances improve when following the data-driven method, and for 10/25 PS bins the performances decrease. 5 of the PS bins have the same background acceptance. It does not make a significant changes in performances whether the MC method is used or the data-driven method is used. Ideally, it should be tested in the data-data case. The very reason for using the purification process, impurities in signal and background, is not allowing a study of the effects in data, since it is not possible to train on mis-labeled training samples if the mis-labeling is too large. Also, when too many mis-labeled events are present the final ROC curves do not show the actual separation.

2.5 Additional variables

In order to include η and $\langle\mu\rangle$, a re-weighting of signal and background for the two variables are needed such that they are the same for signal and background. They do not discriminate in themselves and therefore they need to be weighted such that they are the same for signal and background. That ensure that the ML algorithms does not gain additional discrimination due to the variables themselves but it allow them to possibly increase the discriminating power of the other variables.

2.5.1 Re-weighting

For the calorimeter sub-classifier, the additional variables are η and $\langle\mu\rangle$. They can contribute to the discrimination of electrons to non-electrons with information on the activity in the detector ($\langle\mu\rangle$) or the exact detector geometry (η) for an event.

For the track classifier, additional information from the TRT is included. For both variables, numberOfTRTHits and numberOfXenonHits the distributions are the same for signal and background.

In section 2.2, η and $\langle\mu\rangle$ are shown re-weighted. In Figure 2.41, the weight distribution is shown together with η and $\langle\mu\rangle$.

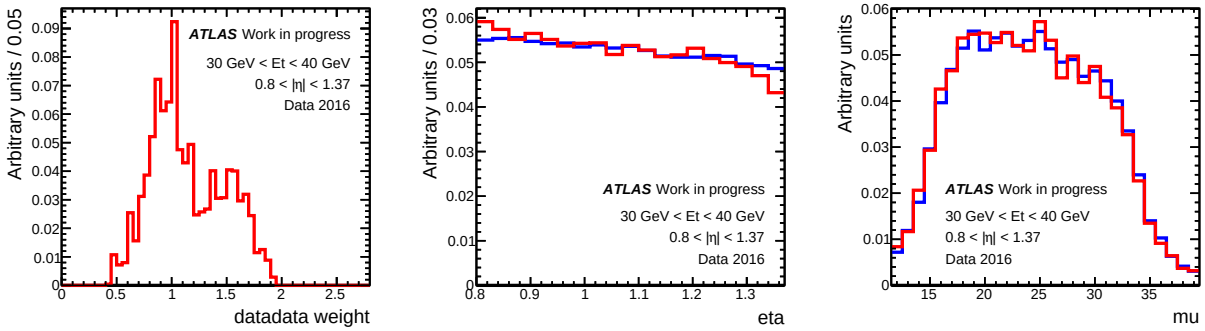


Figure 2.41: Blue is signal and red is background. Left: Weight distribution. Middle: $|\eta|$ distribution. Right: $\langle\mu\rangle$. All are shown for $30 < E_T < 40$ GeV and $0.8 < |\eta| < 1.37$ bin. Note no huge weights.

2.5.2 Performance as a variable of eta, mu and Et

One concern doing electron identification, is the performance as a function of different variables. The ever increasing luminosity of LHC, forces the classifiers to be as flat as possible in $\langle\mu\rangle$. It is also desirable to have a classifier which is efficient in $|\eta|$ and E_T. The behavior of the classifiers has been studied with adaBoost, adaBoost with additional variables and for the LH.

Figure 2.42 shows the performance as a function of $|\eta|$, $\langle\mu\rangle$ and E_T. For $|\eta|$ the shape of the performance curve is the same as for the LH. The adaBoost with additional variables performs slightly better compared to the adaBoost with LH variables as expected. For higher $|\eta|$ the distribution is much more flat compared to the LH. The increasing improvement with increasing $|\eta|$, shown earlier is also seen in the figure. For $\langle\mu\rangle$, the performance is decreasing with increasing $\langle\mu\rangle$ which is expected since increasing $\langle\mu\rangle$ results in more activity in the detector. The performance curve is more flat than the LH, and with the increasing $\langle\mu\rangle$ in the LHC the relative performance of a BDT based classifier would increase more than reported here. The E_T performance curve follows that of the LH.

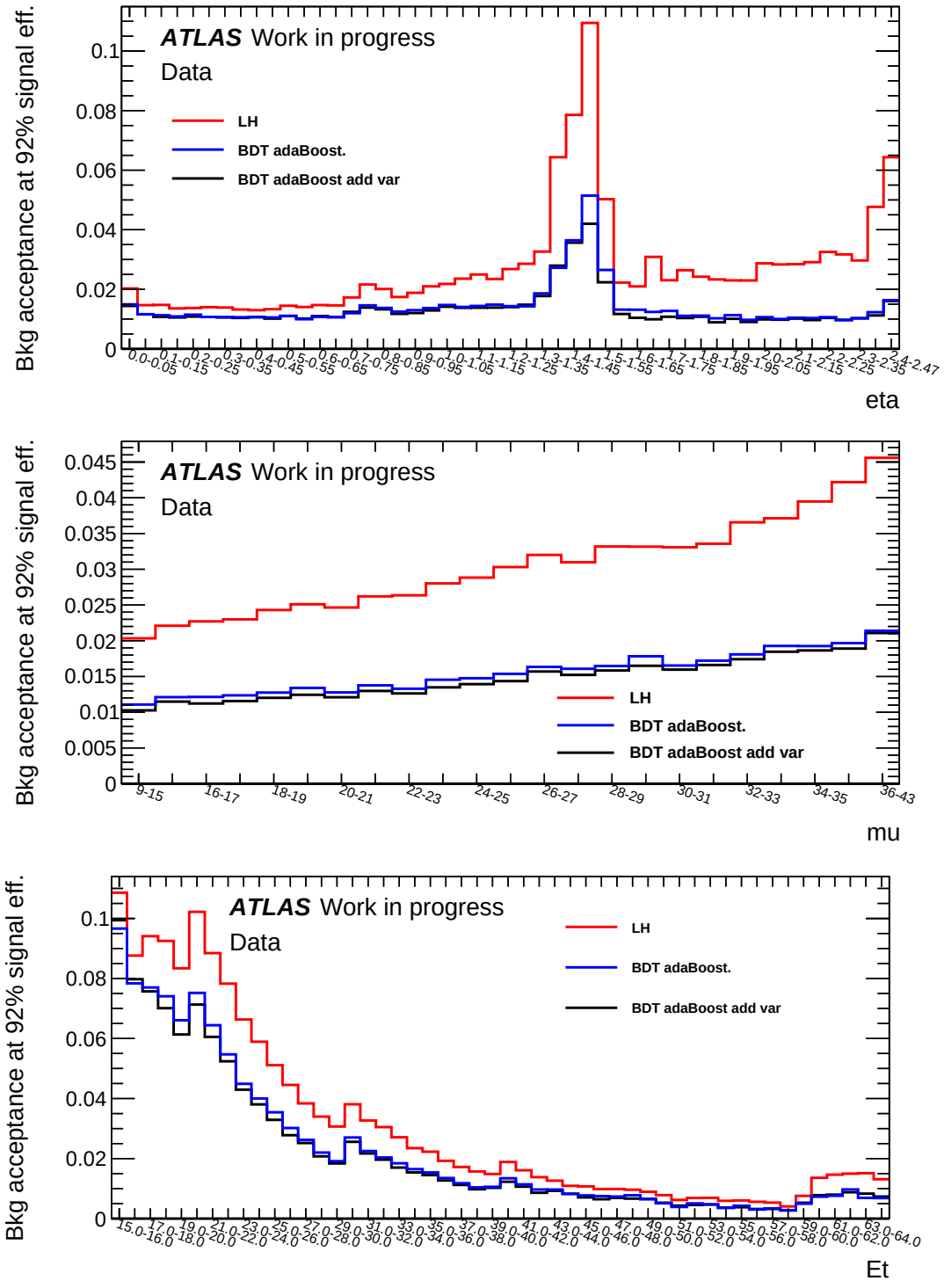


Figure 2.42: performance check. Top: Performance as a function of η. The BDT classifier is more flat then LH. Middle: Performance as a function of $\langle\mu\rangle$. Curve is more flat than the LH. This is very desirable due to the increasing $\langle\mu\rangle$ in the LHC. Bottom: Performance as a function of E_T. The shape is the same as the LH.

2.6 Isolation

In many analyses, searching for new particles involve isolated electrons. The isolation is typically based on a cut in the etcone30 variable [20]. In this section the gain of combining all the isolation variables is presented. It is only shown for the data case and for adaBoost. The distributions of the isolation BDTs can be seen in Section 2.4 and the distribution of etcone30 is shown in Section 2.2.

The results are shown in Figure 2.43. The figure shows the relative improvement of the isolation BDT compared to etcone30 at 92% signal efficiency. This is after purifying data using the calorimeter and track data trained BDTs. As expected the relative performances are much better, and using an isolation BDT instead of etcone30 would decrease the background significantly when searching for new particles with ATLAS.

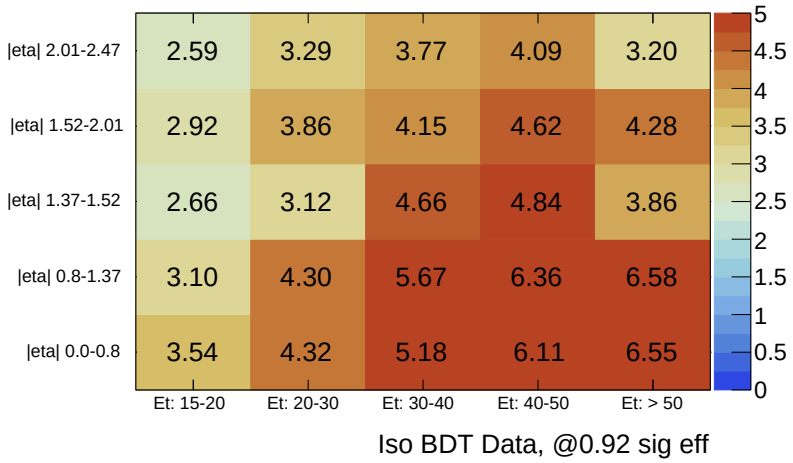


Figure 2.43: The relative performance of the data isolation BDT compared to the etcone30 variable.

2.7 Neural networks

The training of the NN follows the same scheme as the one presented in section 2.4, Figure 2.18. The BDTs trained on MC have been used to purify the data samples. Afterwards, three NNs have been trained in data with the calorimeter, isolation and tracking variables. Only the NNs trained in the calorimeter and track are presented.

The architecture of the NNs are the same for each sub-classifier and for each PS bin. The number of neurons in the first layer is equal to the number of input variables with ReLu activation. The second layer has 20 neurons with ReLu activation. The third layer has 20 neurons with softplus activation and, finally, the output layer has two neurons with softmax activation. The loss function is a binary cross entropy. The optimizer used is nadam [19]. In preliminary studies, different architectures were tried with more layers with varying number of neurons, fewer layers and different optimizers. Ideally, the optimization would be done for each sub-classifier and for each PS bin. The NNs require more tuning for optimal solutions than the BDTs do, but due to limited time this was not done.

In Figure 2.44, the distributions of the calorimeter and track NNs for the $40 < E_T < 50$ GeV and $0.8 < |\eta| < 1.37$ PS bin are shown for both training and test. Training and test for signal and background should have the same shape. The distributions are not normalized, but the shapes are similar for training and test, and the corresponding ROC curves are shown in Figure 2.45. The ROC curves are similar for training and test which shows that the shapes of the distributions are the same and indicates no over training. Note the track distributions are less separated than the calorimeter distributions.

In Figure 2.46, the relative performances of the calorimeter NN compared to the calorimeter BDT at 92% signal efficiency are shown. The comparison is between the adaBoost data trained BDTs. The relative performances are the ratios between the BDT background acceptances compared to the NN background acceptances. For most PS bins except the crack region, the NNs have a higher performance than the BDT. In the crack region the performances are worse, and one possible explanation is the statistics being relatively low in the crack region.

In Figure 2.47, the relative performances of the track NNs compared to the track BDTs are shown. The track NNs are performing poorly to compared to the calorimeter NNs. In general the calorimeter variables are more continuous and less categorical. The track variables consists of more integer variables which can be one explanation for the poor performances of the track NNs. An optimization of the architecture and the activation functions of the neurons could be one way to overcome the lack of performance.

The combination of the calorimeter NNs and track NNs into a Fisher's does not yield good results. This is due to the lack of performance from the track NNs. A combination of the calorimeter NNs and track BDTs would likely perform better than the pure BDT based

classifiers. This was not done in this study.

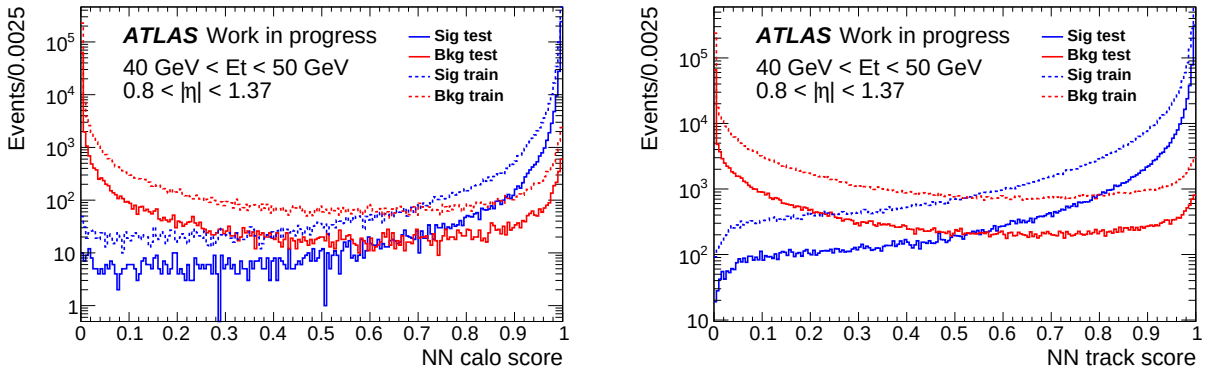


Figure 2.44: Distribution of the NN output from the calorimeter and track for both the training sample

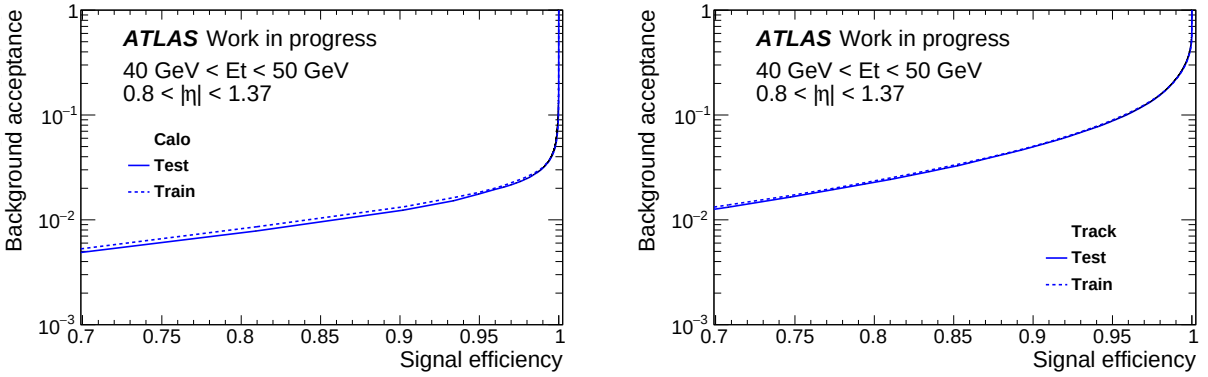


Figure 2.45: Distribution of the NN output from the calorimeter and track for both the training sample and the test sample for $40 < E_T < 50$ GeV and $0.8 < |\eta| < 1.37$. It is shown after cleaning using the MC trained BDTs following same procedure as presented in Figure 2.18.

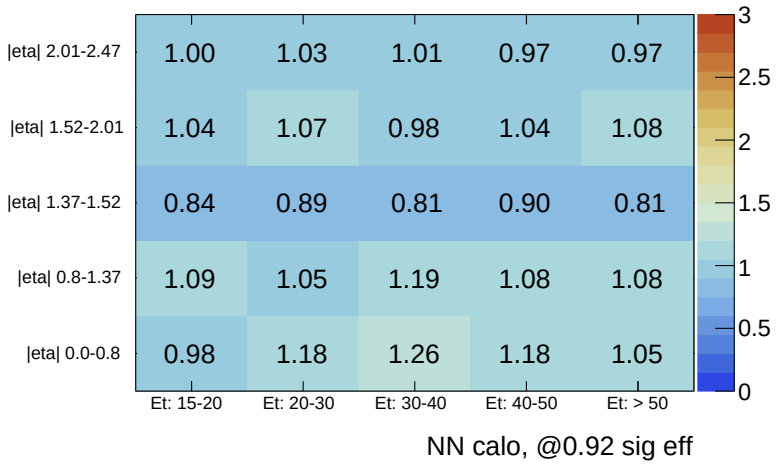


Figure 2.46: Relative performance of the calorimeter NN compared to the calorimeter BDT. For the crack region the calorimeter NN is performing worse than the BDT. This region has lower statistics than the other PS bins.

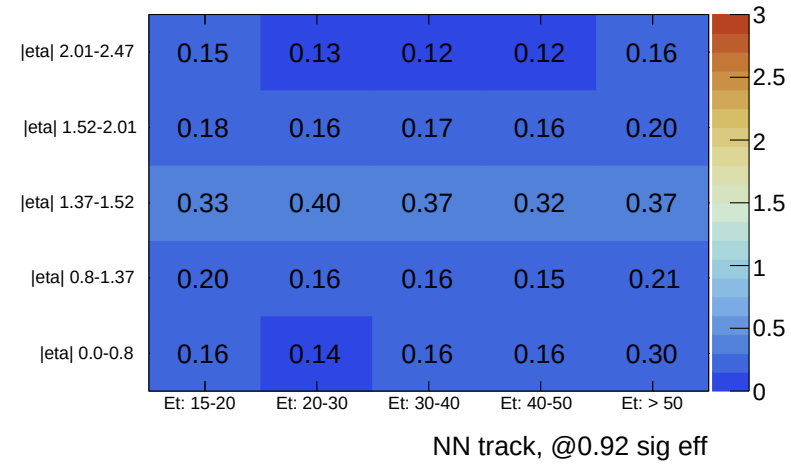


Figure 2.47: Relative performance of the track NN compared to the track BDT. The performance is poor compared to the BDTs. The performance in the crack region is not as poor as for other PS bins.

2.8 Concluding remarks

2.8.1 Summary

This thesis has presented results from implementing machine learning in electron identification in the ATLAS experiment.

Results from implementing a BDT based classifier trained on Monte Carlo (MC) yields improvements compared to the current likelihood (LH) based identification method. The gain in performance decreases when evaluating the classifier on data. To overcome this effect, a data-driven training method has been developed.

In data, the number of mis-labeled electrons and non-electrons from event selections are significant, and a method to remove mis-labeled events in data is needed. The removal is done by splitting up the identification into two sub-classifiers, one for the calorimeter and one for the inner detector. Also, an isolation classifier has been constructed based on isolation in the calorimeter and inner detector. When training one sub-classifier, the two other sub-classifiers are used to remove mis-labeled events, producing data samples with purities above 99%. This step involved using the MC trained classifiers from previously.

After purification, the sub-classifiers are trained on data. Finally, the calorimeter and track BDTs are combined with a Fisher's discriminant into one classifier. The BDTs trained on data have been tested with two different boosting algorithms, adaptive boosting and gradient boosting. The results from the adaptive boosting BDT based classifier gives an improvement compared to the LH of 94% more background rejection at 92% signal efficiency corresponding to 4% increased signal efficiency at the same background rejection. The gradient boosting based classifier gives an improvement of 104% more background rejection. Adding more variables to the classifiers gives an improvement of 109% for the adaptive boosting based classifier.

Implementing neural networks (NN) gives the best calorimeter sub-classifier performance but a poor track sub-classifier performance. The NN calorimeter classifiers perform up to 20% better than the BDT calorimeter classifiers. Except in the crack where they perform 15% worse.

2.8.2 Outlook

This work opens up many different possibilities related to electron identification in ATLAS. The data-driven training method allows for more complicated machine learning algorithms to be used for electron identification. The minor differences in MC and data turns out to decrease the performance gained from implementing machine learning on MC for electron identification.

To obtain a complete ML identification for electrons, lower energy and forward electrons need to be included in the analysis.

For lower energy electrons, the background increases drastically

compared to the signal. The method presented in this thesis to remove mis-labeled events will probably not be sufficient. Though, this method together with [24] which allows for mis-labeled events could solve the issue. Otherwise, a pure MC implementation is an option.

For the forward region, the challenge is different. The method presented in this thesis is based on option of using the inner detector to purify when training the calorimeter sub-classifier, but this is not possible for the forward region since no tracking information is available. An inspection of the calorimeter variables could aid in a division of the calorimeter variables, such that two calorimeter classifiers were constructed enabling a purification scheme similar to what has been presented in this thesis.

In my opinion, the biggest improvements will be gained from including low-level variables e.g. in the NN implementation for the calorimeter. Based on [25], NN performs better when using both high-level and low-level variables, and given that the NN implementation already performed better with only high-level variables, the classifier should be much better.

Appendices

A.1 Correlation between variables in data

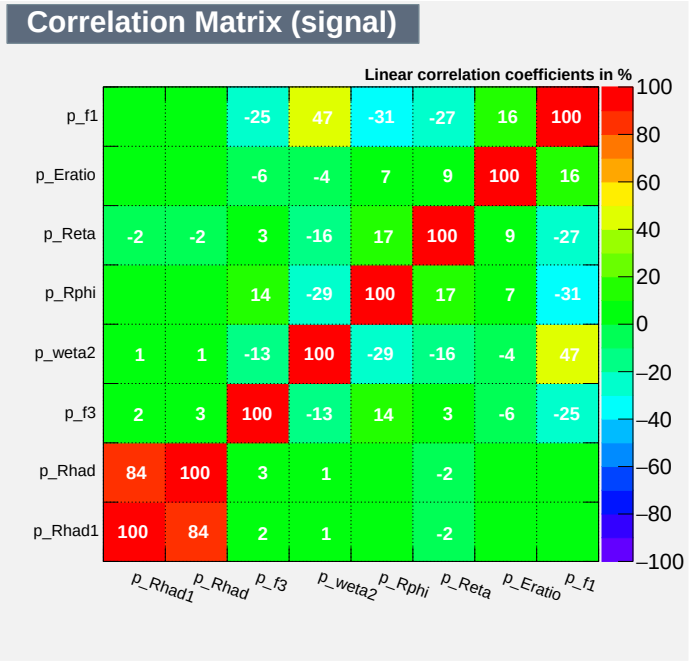


Figure A.1: Correlation between the calorimeter variables for signal from data for $0.8 < |\eta| < 1.37$ and $30 < E_T < 40$ GeV.

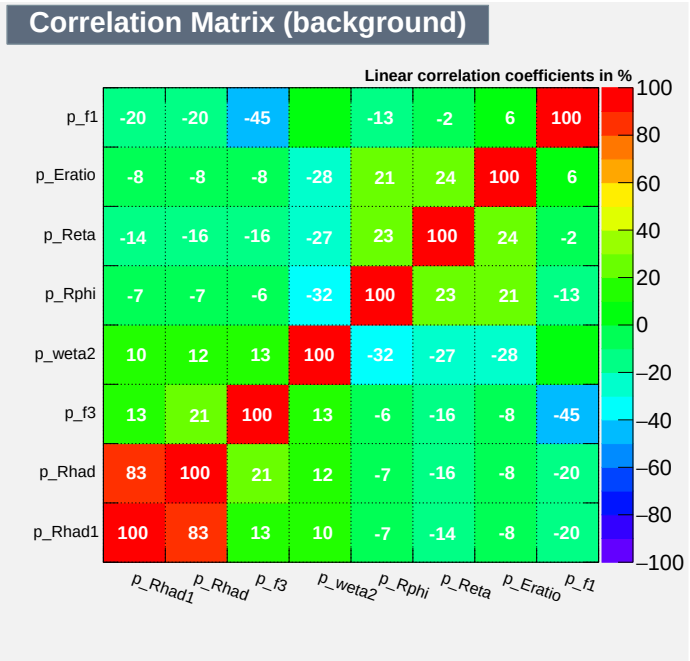


Figure A.2: Correlation between the calorimeter variables for background from data for $0.8 < |\eta| < 1.37$ and $30 < E_T < 40$ GeV.

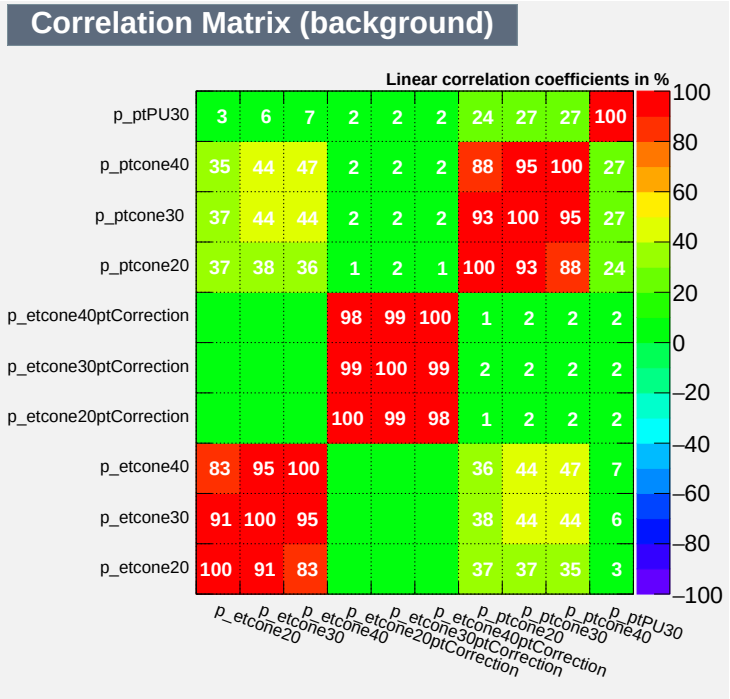


Figure A.3: Correlation between the isolation variables for signal from data for $0.8 < |\eta| < 1.37$ and $30 < E_T < 40$ GeV.

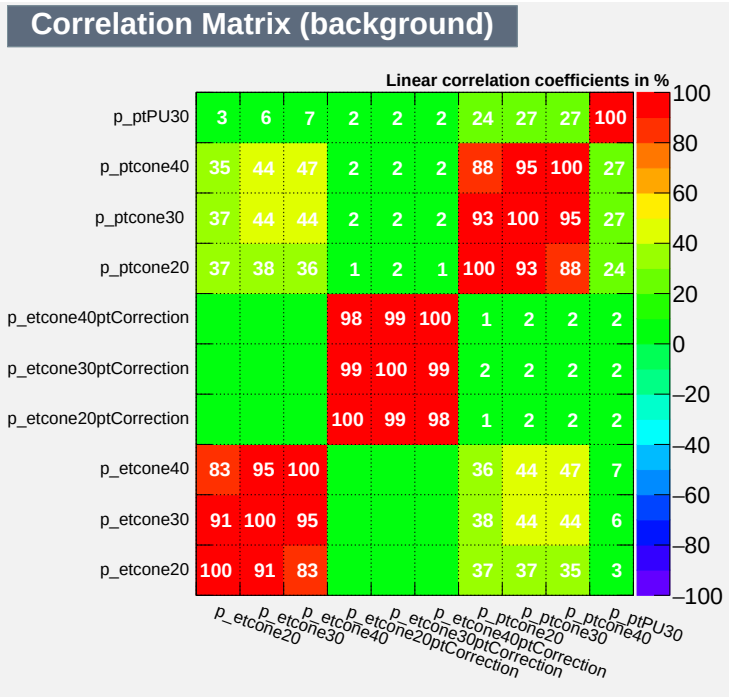


Figure A.4: Correlation between the isolation variables for background from data for $0.8 < |\eta| < 1.37$ and $30 < E_T < 40$ GeV.

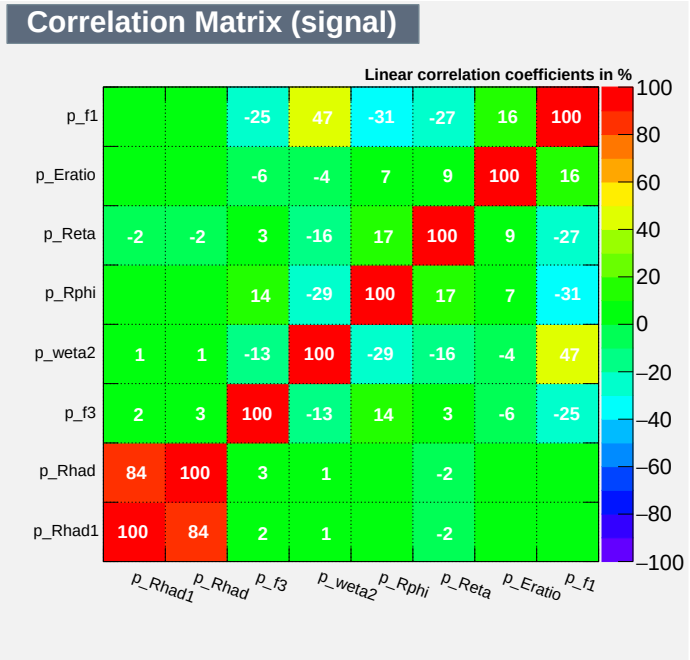


Figure A.5: Correlation between the track variables for signal from data for $0.8 < |\eta| < 1.37$ and $30 < E_T < 40$ GeV.

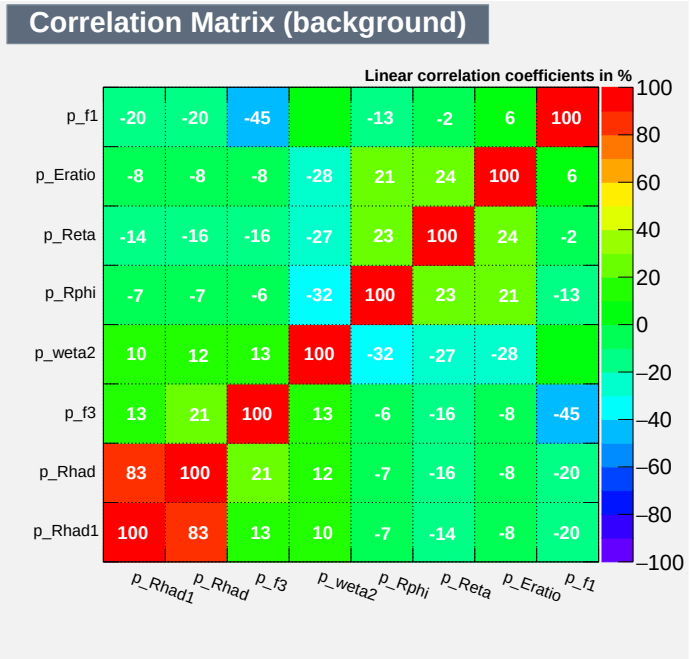


Figure A.6: Correlation between the track variables for background from data for $0.8 < |\eta| < 1.37$ and $30 < E_T < 40$ GeV.

References

[1] CERN, Antonio Saba, and Peter Ginter. CERN Accelerator Complex, 2008. [Online; accessed 24-August-2016].

[2] LHC Machine Outreach. [Online; accessed 07-October-2016].

[3] G. Aad et al. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008.

[4] ATLAS Collaboration. Performance of the atlas transition radiation tracker in run 1 of the lhc: tracker properties. *Journal of Instrumentation*, 12(05):P05002, 2017.

[5] ATLAS Collaboration. ATLAS Insertable B-Layer Technical Design Report. Technical Report CERN-LHCC-2010-013, CERN, Geneva, Sep 2010.

[6] ATLAS Collaboration. Improved electron reconstruction in ATLAS using the Gaussian Sum Filter-based model for bremsstrahlung. Technical Report ATLAS-CONF-2012-047, CERN, Geneva, May 2012.

[7] W Lampl, S Laplace, D Lelas, P Loch, H Ma, S Menke, S Rajagopalan, D Rousseau, S Snyder, and G Unal. Calorimeter Clustering Algorithms: Description and Performance. Technical Report ATL-LARG-PUB-2008-002. ATL-COM-LARG-2008-003, CERN, Geneva, Apr 2008.

[8] T Cornelissen, M Elsing, S Fleischmann, W Liebig, E Moyse, and A Salzburger. Concepts, Design and Implementation of the ATLAS New Tracking (NEWT). Technical Report ATL-SOFT-PUB-2007-007. ATL-COM-SOFT-2007-002, CERN, Geneva, Mar 2007.

[9] T G Cornelissen, M Elsing, I Gavrilenko, J-F Laporte, W Liebig, M Limper, K Nikolopoulos, A Poppleton, and A Salzburger. The global ÏĞ 2 track fitter in atlas. *Journal of Physics: Conference Series*, 119(3):032013, 2008.

[10] Cross sections, http://www.hep.ph.ic.ac.uk/ wstirlin/plots/crosssections2013.jpg.

[11] J. R. Quinlan. Bagging, boosting, and c4.s. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 1*, AAAI'96, pages 725–730. AAAI Press, 1996.

[12] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Neurocomputing: Foundations of research. chapter Learning Representations by Back-propagating Errors, pages 696–699. MIT Press, Cambridge, MA, USA, 1988.

[13] Sir 1890-1962 Fisher, Ronald Aylmer. 138: The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[14] Alex Rogozhnikov. Roc curve demonstration, 2017. [Online; accessed 21-August-2017].

[15] Andreas Hoecker, Peter Speckmayer, Joerg Stelzer, Jan Therhaag, Eckhard von Toerne, and Helge Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[16] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting, 1997.

[17] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[18] Wikipedia. Decision tree learning, 2017. [Online; accessed 18-August-2017].

[19] Keras. Optimizers - keras documentation, 2017. [Online; accessed 18-August-2017].

[20] Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton-proton collision data. Technical Report ATLAS-CONF-2016-024, CERN, Geneva, Jun 2016.

[21] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

[22] ATLAS event at 13 TeV - Zee candidate, https://cds.cern.ch/record/2019370.

[23] EGamma. Egamma xaod derivations, 2017. [Internal note; accessed 22-August-2017].

[24] Lucio Mwinmaarong Dery, Benjamin Nachman, Francesco Rubbo, and Ariel Schwartzman. Weakly supervised classification in high energy physics. *Journal of High Energy Physics*, 2017(5):145, May 2017.

[25] Whiteson Baldi, Sadowski. Searching for exotic particles in high-energy physics with deep learning. 2014.

ELECTRON IDENTIFICATION USING MACHINE LEARNING IN THE ATLAS EXPERIMENT WITH 2016 DATA