|  |  |
| --- | --- |
| Christiaan Gerhardus Viljoen | Dr Thomas Dietel |
| STUDENT NUMBER: VLJCHR004 | SUPERVISOR |
| [christiaan.viljoen@cern.ch](mailto:christiaan.viljoen@cern.ch) | [thomas.dietel@cern.ch](mailto:thomas.dietel@cern.ch) |
| Department of Statistical Sciences | Department of Physics |
| PD Hahn Building | RW James Building |
| University Avenue North | University Avenue |
| University of Cape Town | University of Cape Town |
| Rondebosch, Cape Town, South Africa | Rondebosch, Cape Town, South Africa |

# Masters Dissertation

## PHY5008W

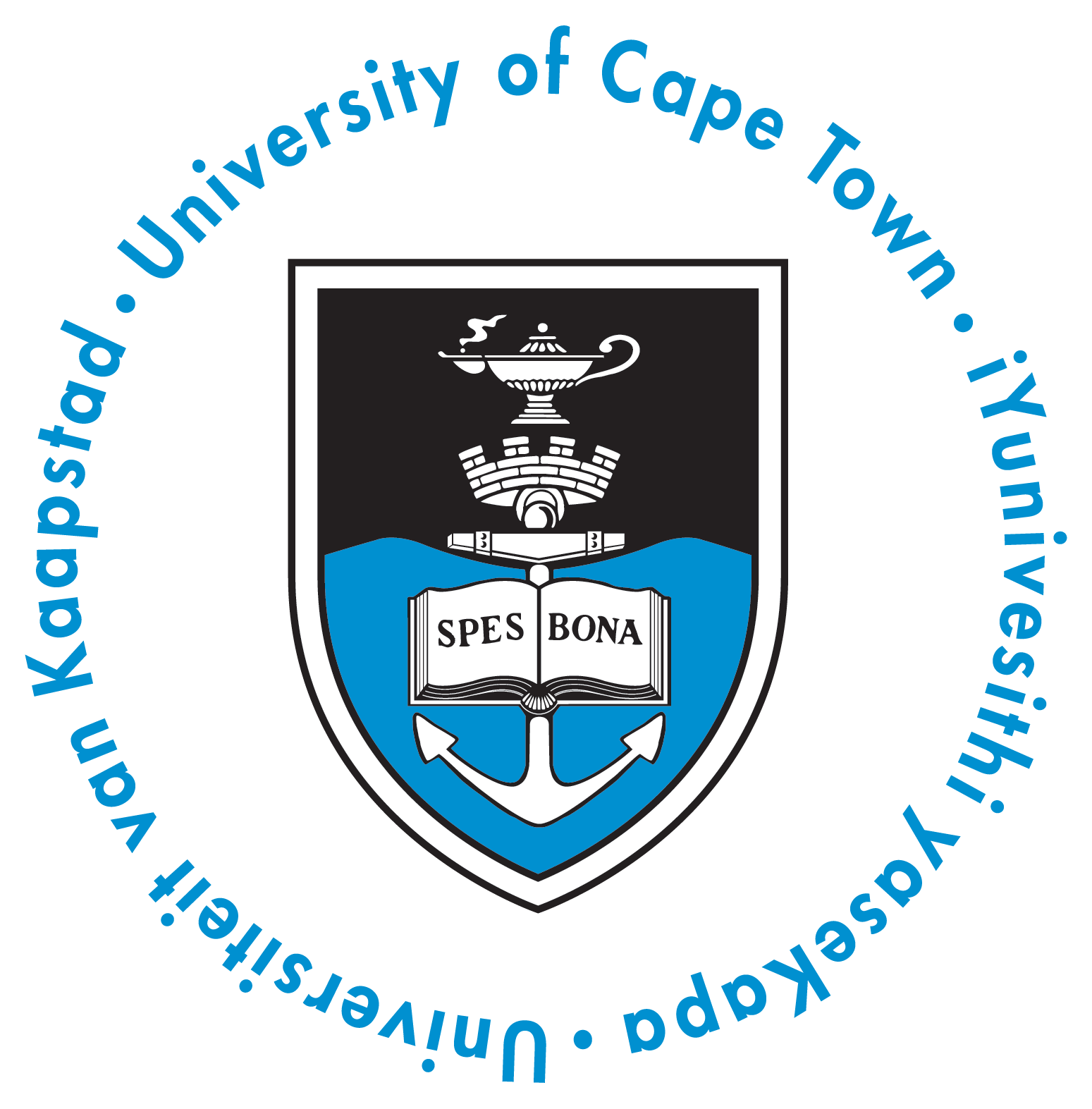
The Application of Machine Learning Techniques towards the Optimization of High Energy Physics Event Simulations in the at

† A Large Ion Collider Experiment

‡ European Organization for Nuclear Research/ Organisation Européenne  
pour la Recherche Nucléaire

͛ Transition Radiation Detector

# Submitted in Fulfilment of the Degree: MSc Data Science



Contents

[Christiaan Gerhardus Viljoen 1](#_Toc534563944)

[Dr Thomas Dietel 1](#_Toc534563945)

[Masters Dissertation 1](#_Toc534563946)

[PHY5008W 1](#_Toc534563947)

[Submitted in Fulfilment of the Degree: MSc Data Science 1](#_Toc534563948)

[Github Repository 8](#_Toc534563949)

[Dedication 8](#_Toc534563950)

[Acknowledgements 8](#_Toc534563951)

[Plagiarism Declaration 9](#_Toc534563952)

[Abstract 10](#_Toc534563953)

[Background 11](#_Toc534563954)

[A Large Ion Collider Experiment (ALICE) 11](#_Toc534563955)

[Introduction to the ALICE Experiment 11](#_Toc534563956)

[The ALICE Detector Set-Up 12](#_Toc534563957)

[The ALICE Transition Radiation Detector (TRD) 12](#_Toc534563958)

[Introduction to the TRD 12](#_Toc534563959)

[TRD Physical Properties 12](#_Toc534563960)

[TRD Measurement Mechanism 14](#_Toc534563961)

[Identifying Electrons 15](#_Toc534563962)

[Current TRD Accuracy 16](#_Toc534563963)

[Existing Particle Physics Software 17](#_Toc534563964)

[Data Analysis Framework: ROOT 17](#_Toc534563965)

[ALIROOT 17](#_Toc534563966)

[Integration with other Programming Languages 18](#_Toc534563967)

[ROOT File Structure & I/O 18](#_Toc534563968)

[The Logical Root File & Direct Access 19](#_Toc534563969)

[The Current Directory 19](#_Toc534563970)

[The Tree Object: 20](#_Toc534563971)

[Getting Event data from TFile: 20](#_Toc534563972)

[Generative Adversarial Networks 24](#_Toc534563973)

[Introduction to Generative Adversarial Networks 24](#_Toc534563974)

[Mathematical Theory: From the Perceptron to GANs 25](#_Toc534563975)

[The Perceptron 25](#_Toc534563976)

[Deep Learning 27](#_Toc534563977)

[A Note on Notation: 27](#_Toc534563978)

[Visual Intuition 28](#_Toc534563979)

[Backpropagation 28](#_Toc534563980)

[Hyperparameters 29](#_Toc534563981)

[Generative Neural Networks Explained at the hand of Pseudocode, based on a minimal example from (13) 29](#_Toc534563982)

[Pseudocode 30](#_Toc534563983)

[Autoencoders 32](#_Toc534563984)

[Variational Autoencoders (VAEs) 33](#_Toc534563985)

[Application of Traditional Deep Learning to High Energy Physics Problems 33](#_Toc534563986)

[Application of Generative Models to High Energy Physics Problems 33](#_Toc534563987)

[Motivation 35](#_Toc534563988)

[High Energy Physics Simulations 35](#_Toc534563989)

[Research Question 37](#_Toc534563990)

[Primary Research Question 37](#_Toc534563991)

[Secondary Research Question 37](#_Toc534563992)

[Auxiliary Research Questions 37](#_Toc534563993)

[Hypothesis 39](#_Toc534563994)

[Aims & Objectives 40](#_Toc534563995)

[Methods 42](#_Toc534563996)

[Data Extraction from WLCG via AliEn 42](#_Toc534563997)

[Deep Learning for Particle Identification (PID) 43](#_Toc534563998)

[Deep Learning towards Geant4 Parameter Optimization based on Environmental Conditions 43](#_Toc534563999)

[HEP Event Simulations 43](#_Toc534564000)

[Monte Carlo Simulation Data Generation using Geant4 with Parameters Relating to Environmental Conditions Optimized 43](#_Toc534564001)

[Generative Adversarial Networks in Keras 43](#_Toc534564002)

[Variational Autoencoders 43](#_Toc534564003)

[Productionalization of Machine Learning-Based HEP Event Simulations 43](#_Toc534564004)

[Results 44](#_Toc534564005)

[Discussion 45](#_Toc534564006)

[Conclusion and Future Work 46](#_Toc534564007)

[Bibliography 47](#_Toc534564008)

[Appendix A 49](#_Toc534564009)

[Appendix B 52](#_Toc534564010)

[Appendix C 53](#_Toc534564011)

[Set up session for Deep Learning: 53](#_Toc534564012)

[Data Wrangling for Deep Learning: 53](#_Toc534564013)

[DEEP LEARNING: 56](#_Toc534564014)

[Tuning: 60](#_Toc534564015)

[Tanh with Dropout 60](#_Toc534564016)

[Rectifier with Dropout: 62](#_Toc534564017)

[Maxout with Dropout: 63](#_Toc534564018)

[Recitfier with Dropout with 2 additional hidden layers of 500 neurons each: 69](#_Toc534564019)

[Ensemble: 70](#_Toc534564020)

[Random Forests: 71](#_Toc534564021)

[K-means clustering 72](#_Toc534564022)

[Support Vector Machines 72](#_Toc534564023)

[Final prediction 76](#_Toc534564024)

[Appendix D 77](#_Toc534564025)

[Pad Response Function 77](#_Toc534564026)

[Remove entries with missing pad data: 77](#_Toc534564027)

[Get a Look-Up Table for Tracks within Events: 77](#_Toc534564028)

[Appendix E 79](#_Toc534564029)

[Set seed for reproducibility: 79](#_Toc534564030)

[Read in all the json files, created from python dictionaries 79](#_Toc534564031)

[Threshold pad data 79](#_Toc534564032)

[Descriptive statistics 81](#_Toc534564033)

[Particle Identification: 86](#_Toc534564034)

[Majority Class Classifier: 86](#_Toc534564035)

[Confusion matrix: 86](#_Toc534564036)

[Simple linear regression: 87](#_Toc534564037)

[Pion Efficiency: 89](#_Toc534564038)

[Electron Efficiency: 89](#_Toc534564039)

[Optimization of linear model prediction cut-off point for electron classification: 89](#_Toc534564040)

[Optimized Cut-off linear model accuracy assessment: 90](#_Toc534564041)

[Pion Efficiency: 90](#_Toc534564042)

[Electron Efficiency: 90](#_Toc534564043)

[DataRobot: Automated ML 91](#_Toc534564044)

[Deep Learning 94](#_Toc534564045)

[Roll out pad data 94](#_Toc534564046)

[Unbalanced class compensation: 95](#_Toc534564047)

[Truly balance classes 98](#_Toc534564048)

[Even Deeper Learning 102](#_Toc534564049)

[Model 3 Analysis 103](#_Toc534564050)

[Possible Causes: 103](#_Toc534564051)

[Undersampling of pions, instead of oversampling of electrons 103](#_Toc534564052)

[Using the second neural network, draw a sample of pions of the same n as the sample of electrons, prefering those pions are most similar to electrons based on nusupervised learning methods 103](#_Toc534564053)

[Fuzzy Clustering 107](#_Toc534564054)

[Fuzzy C-means 107](#_Toc534564055)

[Model based clustering 108](#_Toc534564056)

[H2O 113](#_Toc534564057)

[Convolutional Neural Networks 120](#_Toc534564058)

# Github Repository

A Github repository containing all files relating to this Masters Dissertation exists at:

<https://github.com/PsycheShaman/MSc-thesis>

# Dedication

This dissertation is dedicated to my mother, Elizabeth Suzanna Bloem Viljoen, who has always inspired me to follow my higher passions, despite the myriad difficulties that life makes us face; and to search fearlessly and incessantly for the deeper truths underlying our everyday world.

# Acknowledgements

Firstly, I would like to thank my father, Christiaan Gerhardus Viljoen, for all the support – material, emotional and financial – he has selflessly provided to me throughout my life, and particularly towards my higher education journey. You have no idea how much appreciation I have for all the sacrifices you have made for me, and all the advice you have given me.

Secondly, I want to thank my aunt, Professor Emma Ruttkamp-Bloem, for all the mentoring she has provided to me in navigating the world of academia, and for the inspiration that her own academic career instils in me.

Thirdly, I want to thank Dr Thomas Dietel for providing me with this immense opportunity to be part of the largest scientific experiment in human history, and for the rigorous scientific guidance that he has, and continues to provide to me.

Lastly, I would like to thank my larger family, on both my father’s and mother’s side, for providing the loving and stable environment that makes any place we assemble Home.

# Plagiarism Declaration

**TURNITIN PLAGIARISM DECLARATION**

.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MSc Candidate Full Name | | | Christiaan Gerhardus Viljoen | | |
| Student Number | | | VLJCHR004 | | |
| Faculty | | | Science | | |
| Department | | | Statistical Sciences | | |
| Supervisor | | | Dr Thomas Dietel | | |
| Co-Supervisor | | | N/A | | |
| Thesis Title | | | The Application of Machine Learning Techniques towards the Optimization of High Energy Physics Event Simulations in the ALICE TRD at CERN | | |
| Note: Please complete and sign the declaration. | | | | | |
| DECLARATION BY STUDENT | | | | | |
| I, | Christiaan Gerhardus Viljoen | | | | (Student’s name) |
| hereby declare that this thesis/dissertation has been submitted to the Turnitin module (or equivalent similarity and originality checking software).  Plagiarism is to use another’s work and pretend that it is one’s own and I know that plagiarism is wrong. I confirm that I have discussed and resolved any concerns emanating from the Turnitin report with my supervisor. | | | | | |
| Signed | |  | | Date |  |
|  | | | | | |
| DECLARATION BY SUPERVISOR | | | | | |
| I, | Dr Thomas Dietel | | | | (Supervisor’s name) |
| hereby declare that this thesis/dissertation has been submitted to the Turnitin module (or equivalent similarity and originality checking software). Plagiarism is to use another’s work and pretend that it is one’s own. I confirm that I have discussed and resolved any concerns emanating from the Turnitin report with my student. | | | | | |
| Signed | |  | | Date |  |

# Abstract

# Background

## A Large Ion Collider Experiment (ALICE)

### Introduction to the ALICE Experiment

A Large Ion Collider Experiment (ALICE) is a large scale collaborative experiment dedicated to studying all collisions involving heavy ions at the Large Hadron Collider (LHC) at CERN (European Organization for Nuclear Research) (1).

In central high energy collisions between heavy ions (i.e. where the centres of colliding nuclei overlap sufficiently), a newly discovered deconfined state of strongly interacting matter, the Quark Gluon Plasma (QGP) can be created in small amounts (1). It is thought that this state of matter was dominant during the first 10-6 s of the Universe’s existence (2). Studying the QGP allows us to explore fundamental research avenues such as Cosmology, the Evolution of our Universe, and one of the fundamental forces in the standard model that is the hardest to probe: the strong nuclear force (2).

ALICE is the first experiment in history capable of producing the QGP in a laboratory setting; and as such, it is equipped to infer a variety of physical variables relating to the QGP, by analysing data from electrons produced during many of the physical processes that occur in the wake of heavy ion collisions, e.g. open heavy-flavour hadron decays, virtual photons, etc (1). Robust electron identification is therefore a crucial part of studying the QGP, and accurately-tuned detector triggers are required to ensure the collection of sufficient amounts of data to guide inferences regarding the statistical distributions of the abovementioned measurables (1).

## The ALICE Detector Set-Up

The ALICE apparatus’ dimensions are and weighs around 10 000 tonnes and consists of 17 detector systems, categorized into three sections: central barrel detectors that are immersed in a 0.5 Tesla magnetic field and are used to detect particles produced at midrapidity (the TRD, whose output will be simulated in this project, is located within the central barrel), forward detectors which are important in triggering, investigating multiplicity and event characterization and the MUON spectrometer, which measures muons of high transverse momentum , quarkonium and light vector mesons (3).

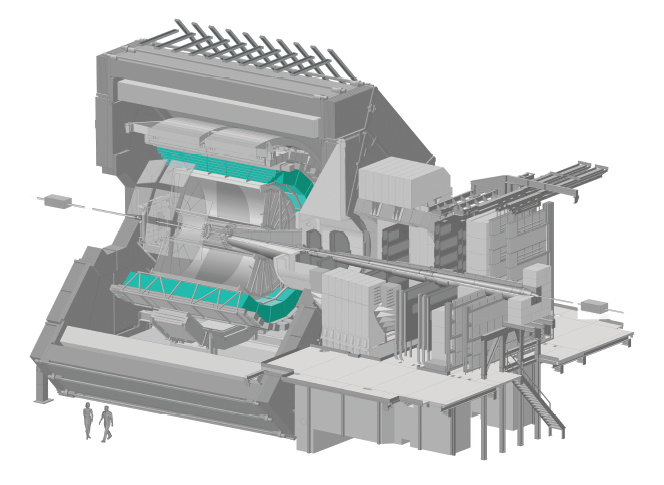
### The ALICE Transition Radiation Detector (TRD)

#### Introduction to the TRD

The main purpose of the ALICE Transition Radiation Detector (TRD) is the identification of electrons, as well as the operation of event triggers that determine whether data from a specific collision should be kept, based on measurements such as collision centrality, amongst others. As an added benefit, the TRD informs the ALICE central barrel’s calibration, and the data it produces is used extensively during track reconstruction and particle identification (1).

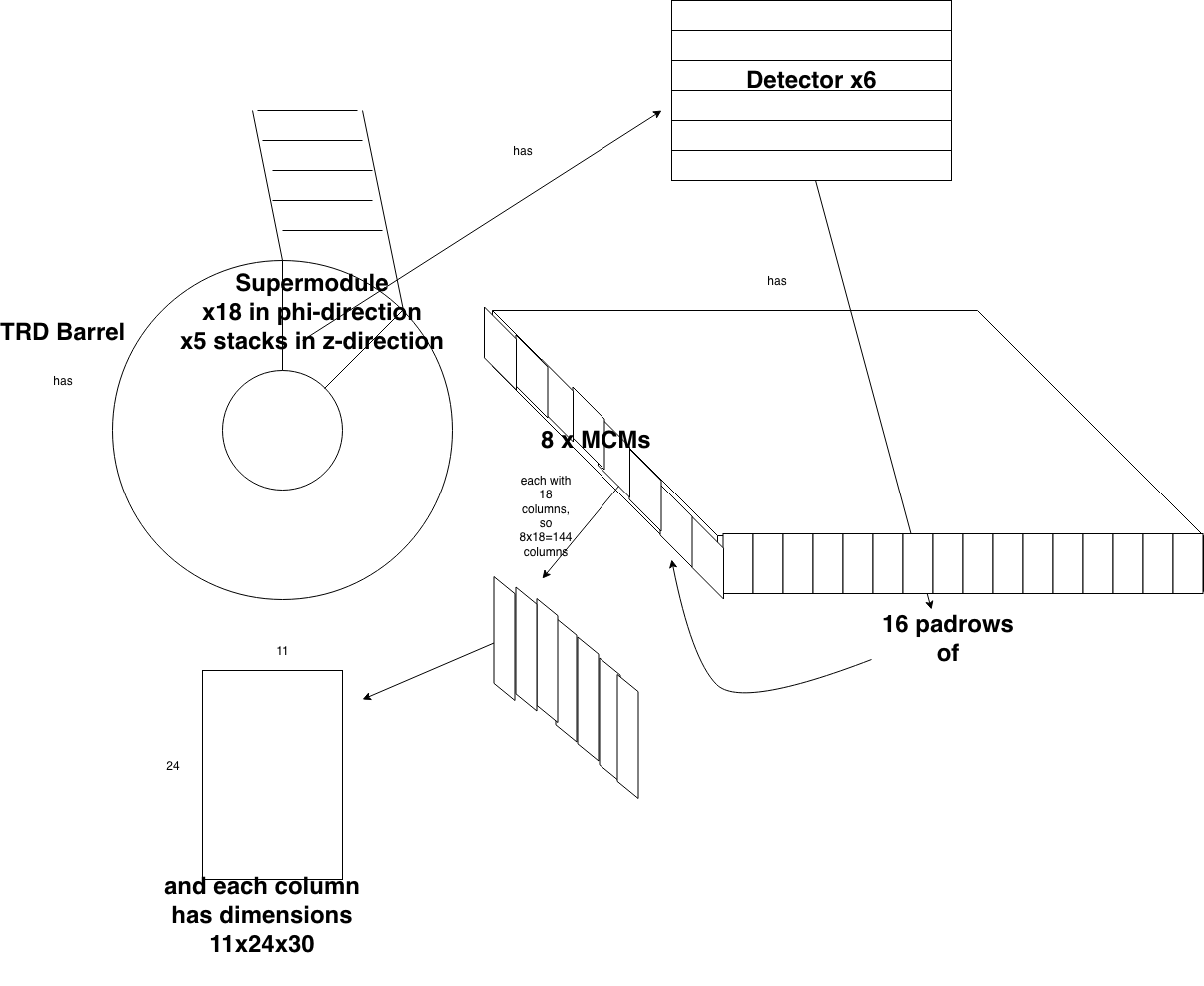
#### TRD Physical Properties

The TRD is located within the ALICE central barrel, it’s inner boundary sits at a radial distance of 2.90 m from the beam axis. It contains 522 chambers, each of which contains a fiber/ foam radiator, a 3 cm drift region and a multi-wire proportional chamber (MWPC) filled with Xenon-; these chambers are arranged in a six-layer configuration (1). The position of the TRD within the ALICE central barrel is shown in *Figure 1*.



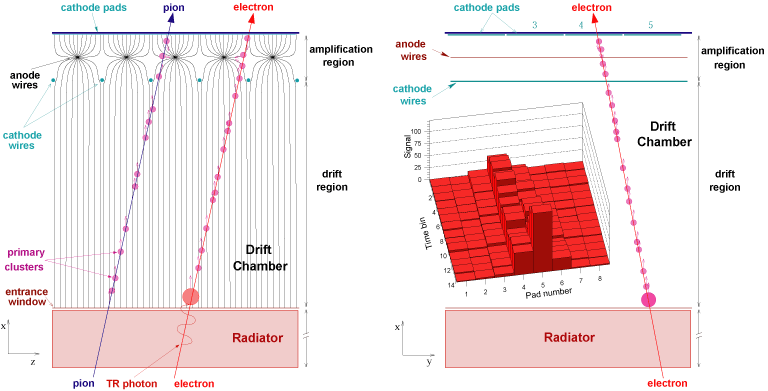
**Figure 1: The ALICE TRD, Highlighted in Cyan, within the ALICE main detector (4)**

//todo: add description of TRD internal geometry from TRD TDR



#### TRD Measurement Mechanism

As the name suggests, transition radiation occurs when a particle transits across a dielectric boundary, this radiation is often measured in particle detectors to inform track reconstruction. Multiple boundaries are typically required to increase radiation yield, and since highly relativistic particles emit transition radiation that extends into the X-ray domain, the TRD utilizes gases with high proton-number (Z) to absorb this radiation, resulting in a high yield of energy deposition relative to the energy lost via ionization (1).



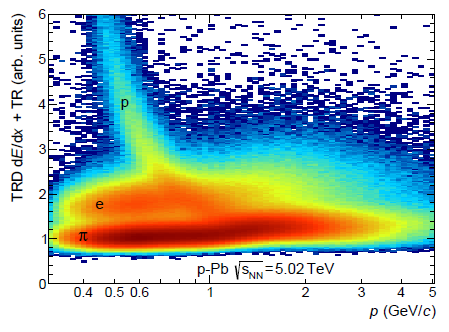
**Figure 2: A schematic representation of the components in an MWPC Module (5)**

The drift time of gas particles within the MWPC provides fine-grained positional information about where the particle tracklet passed through the radiator. The detected signal takes the form of charged gas molecules (ionized via interaction with transition radiation photons and amplified through a chain of interactions between gas molecules), finally being absorbed by a negatively charged wire (anode) (4), this process is depicted in *Figure 2*.

#### Identifying Electrons

The production of Transition Radiation (TR), as well as a higher specific ionization energy loss , are two features that enable accurate differentiation of electrons from other charged particles; and the temporal data provided by the TRD further enhances the specificity and sensitivity of the electron identification process (1).

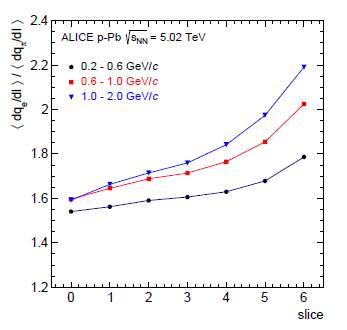
On average, a single TR photon with an energy range of 1-30 is produced by an electron with momentum higher than 1 .



**Figure 3: sum of TRD signal (Transition radiation plus energy loss due to ionization), as a function of momentum, for Protons (), Electrons () and Pions () (3)**

During particle identification (PID), the signal from each chamber is sliced into 7 conceptual sections, moving inward from the outermost readout, each section integrates the signal captured in ±5 mm of detector thickness, see ***Figure 4*** for an indication of how the electron signal (scaled by the pion signal) decays in sections further out from the detector entrance, since most TR is absorbed in the sections closer to the detector entrance (3).

The main analytical methods used to identify electrons are TPC and TOF particle identification and topological cuts (3).



**Figure 4: The average signal ratio (signal units ), as a function of slice number. larger slice numbers achieve a higher electron signal, since they are closest to the entrance of the detector, where most TR is absorbed (3)**

#### Current TRD Accuracy

The TRD detector performance is quantified by pion efficiency, which is the proportion of pions misclassified as electrons (i.e. the False Positive Rate). The pion rejection factor is the inverse of this, i.e. (3).

Researchers are also interested in the electron efficiency (i.e. the True Positive Rate), since the process of eliminating pions will inevitably result in some elimination of electrons misclassified as pions; so ideally one would want to maximize electron efficiency, whilst simultaneously maximizing the pion rejection factor.

Currently, at a momentum of around 1 GeV/*c*, a pion rejection factor of 410 is achievable in p-Pb (proton-Lead) collisions, with resolution improving by about 40% when TRD data is included in track reconstruction (1), with electron efficiency achieved in this momentum range being around 90% [citation needed, inferred from a plot in this paper (3)].

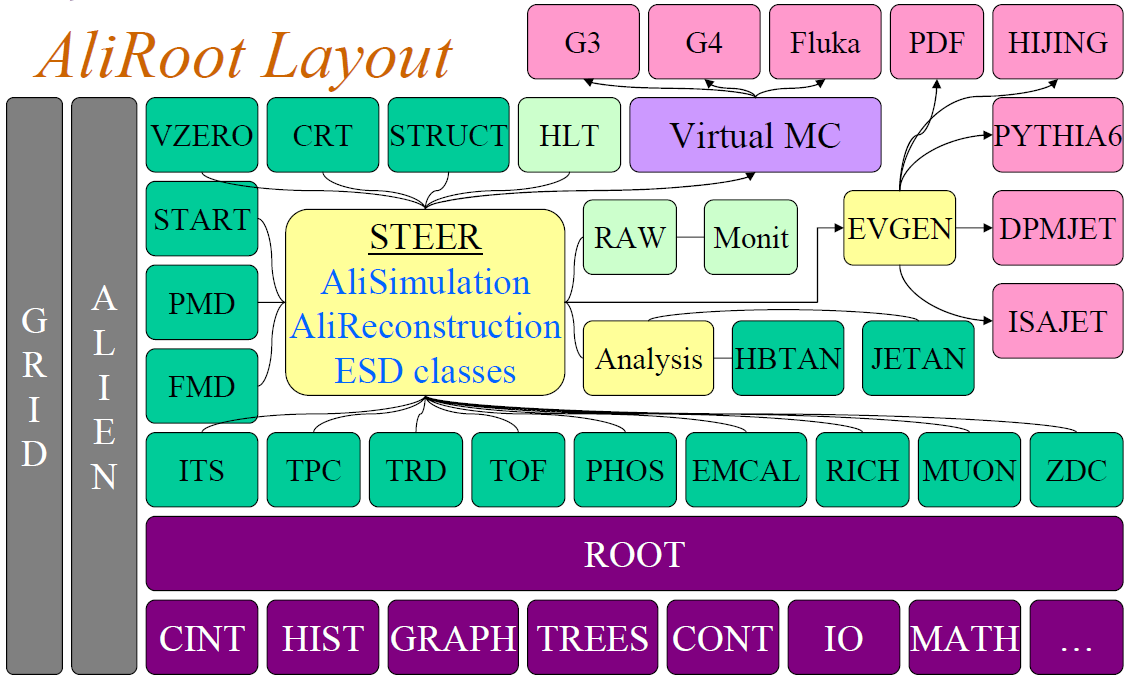
## Existing Particle Physics Software

### Data Analysis Framework: ROOT

ROOT is an open-source data analysis framework developed by the High Energy Physics (HEP) community; it is an object oriented platform, written in C++ and allows for storage of classes in a compressed binary file format (.root), as well as an object container class (TTree), which facilitates statistical analysis of large datasets via vertical storage methods, and enabling analysis across files on local disks, shared file systems, and the Worldwide LHC Computing Grid (WLCG) (6).

### ALIROOT

AliROOT is an extension of ROOT, used specifically by the ALICE collaboration. It is an integrated ecosystem consisting of event generators (PYTHIA6, HIJING, DPMJET and ISAJET), Virtual Monte Carlo simulators (Geant3, Geant4 and Fluka) and various steering classes for simulation and reconstruction; all of which sits upon the basic ROOT architecture, and which communicates via the ALICE Environment Grid Framework (AliEn) with the WLCG. This architecture is depicted in *Figure3* (7)*.*



**Figure 5: ALIROOT Ecosystem (7).**

#### Integration with other Programming Languages

PyROOT and ROOT R are interfaces that allow seamless integration between ROOT and Python as well as R, respectively. Features include calling functions from any Python or R package within ROOT, translating datatypes between languages, as well as access to the R graphical system from within ROOT (8).

#### ROOT File Structure & I/O

ROOT files (extension ".root") are binary storage files, comparable to UNIX file directories, an unlimited level of subdirectories containing files and other subdirectories can be contained in a ROOT file. Additionally, storage is implemented in a machine-independent manner.

Within ROOT, once loaded, a ROOT file exists as a TFile class, and is generally opened as follows:

root[] TFile f("demo.root");

The Map() method can be used to view the structure of the file contents, including the byte index of specific objects, their individual sizes in bytes, Classes and Class descriptions. The first 100 bytes contain the file header.

File recovery can be implemented by scanning and reimplementing the information in the header. This method can obviously only revocer the structure of a file, not its contents.

#### The Logical Root File & Direct Access

Specific objects can be accessed using a list of TKey objects, which serve as indices to specific objects.We can get a list of TKeys, as follows:

root[] f.GetListOfKeys()->Print();

We can then load a specific object into memory using its TKey, for example if our TFile contains a single histogram (called "myHistogram" in this example), we will see something like this:

TKey Name = myHistogram, Title = histo nr:1, Cycle = 1

We can then load this histogram into memory by calling:

root[] TH1F \*myHistogram = (TH1F\*)f.Get("myHistogram")

#### The Current Directory

Upon initializing Root, the current directory is the ROOT session, i.e.:

root[] gDirectory->pwd()

Rint:/

But, since a .root file is in essence a UNIX directory: once it is loaded into memory, it becomes the current directory, i.e.:

root[] TFile f1("AFile1.root");

root[] gDirectory->pwd()

AFile1.root:/

#### The Tree Object:

The TTree Class was specifically designed to optimize storage (i.e. in a more compressed format, and allowing faster access) of a large amount of objects of the same class.

### Getting Event data from TFile:

The following was implemented at the ROOT command prompt:

1. Load the AliESDs.root file:

root [0] TFile f("AliESDs.root");

1. Get the list of keys:

root [1] f.GetListOfKeys()->Print();

Collection name='THashList', class='THashList', size=3

TKey Name = ProcessID0, Title = 149f40a6-c715-11e6-9717-f93db9bcbeef, Cycle = 1

TKey Name = esdTree, Title = Tree with ESD objects, Cycle = 1

TKey Name = HLTesdTree, Title = Tree with HLT ESD objects, Cycle = 1

1. Make sure we're in the AliESDs directory:

root [2] gDirectory->pwd()

AliESDs.root:/

1. List the contents of objects that are currently in memory:

root [3] gDirectory->ls("-m")

TFile\*\* AliESDs.root

TFile\* AliESDs.root

1. Load the ESD Tree into memory:

root [4] TTree \*esdTree = (TTree\*)f.Get("esdTree");

1. Open a ROOT Object Browser to view the structure of the ESD Tree

root [3] new TBrowser

(TBrowser \*) 0x8f00d80

1. This opens a GUI (see ***Figure 12***), which allows us to visually inspect the structure of any TObject. We are interested in seeing what the index of the particle tracklets are in the ESD Tree. We see that the TRD tracklets are at position 23 in this Tree, so we do the following to get the unique ID and the Label:

A screenshot of a social media post

Description generated with very high confidence

**Figure 12: Screenshot of the Root Object Browser**

root [4] esdTree->SetScanField(23)

1. Then, we want to write this to a text file, to save the information we want:

root [5] .> tracklet.txt

root [] esdTree->Scan("fHCId:fLabel")

root [] .>

1. This writes the following output to “tracklet.txt”:

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Row \* Instance \* fHCId \* fLabel \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* 0 \* 0 \* 1 \* 1 \*

\* 0 \* 1 \* 0 \* -1 \*

\* 0 \* 2 \* 3 \* -1 \*

\* 0 \* 3 \* 3 \* -1 \*

\* 0 \* 4 \* 2 \* -1 \*

\* 0 \* 5 \* 2 \* -1 \*

\* 0 \* 6 \* 2 \* 1 \*

\* 0 \* 7 \* 7 \* -1 \*

\* 0 \* 8 \* 7 \* -1 \*

\* 0 \* 9 \* 8 \* -1 \*

\* 0 \* 10 \* 8 \* -1 \*

\* 0 \* 11 \* 10 \* -1 \*

\* 0 \* 12 \* 13 \* -1 \*

\* 0 \* 13 \* 26 \* -1 \*

\* 0 \* 14 \* 26 \* -1 \*

\* 0 \* 15 \* 26 \* -1 \*

\* 0 \* 16 \* 26 \* -1 \*

\* 0 \* 17 \* 26 \* -1 \*

\* 0 \* 18 \* 26 \* 1 \*

\* 0 \* 19 \* 28 \* 1 \*

\* 0 \* 20 \* 28 \* -1 \*

\* 0 \* 21 \* 32 \* 1 \*

\* 0 \* 22 \* 41 \* 1 \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

1. Tracing Back one step in the hierarchy, to look at TRD tracks, we do:

root [6] esdTree->SetScanField(22)

root [7] .> track.txt

## Generative Adversarial Networks

### Introduction to Generative Adversarial Networks

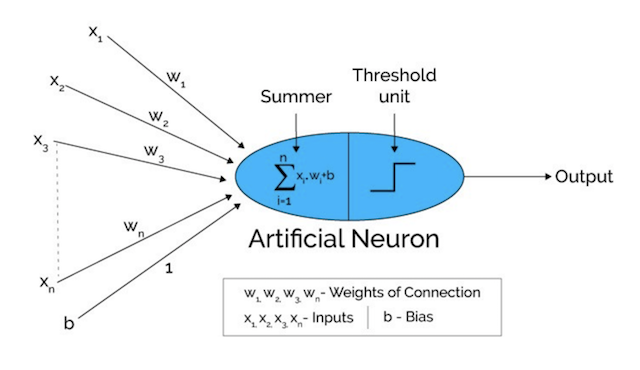
In 2014, Goodfellow et. al. (9) proposed a novel implementation of Deep Learning, involving an adversarial mini-max game between two Artificial Neural Networks (ANNs), which they called Generative Adversarial Nets (GANs). A GAN model consists of two elements: a generative model G, and a discriminative model D. During training, G aims to maximize the loss function of D by accurately capturing the underlying distribution of the training data, and D estimates the probability of whether an observation fed to it is “real”, i.e. it is from the actual data distribution or “fake”, i.e. it was generated by G (9).

If both models are ANNs, they can be trained simultaneously via backpropagation, until a unique solution can theoretically be reached, where G accurately models the data distribution and D outputs 0.5 everywhere (9).

### Mathematical Theory: From the Perceptron to GANs

#### The Perceptron

The field of Deep Learning is based on the idea of a “perceptron” proposed by Rosenblatt in 1958 (10), originally envisioned as a way to mathematically model neural information processing in the brain, it was not very successful in capturing the complexity inherent in neurobiological information processing, but very useful for learning non-linear patterns in data. ***Figure 4*** informs the mathematical theory that follows.



**Figure 6: A Schematic Representation of a Single Perceptron (11)**

In the figure above an observation , consisting of a vector of input features , … , is mapped to an estimated outcome variable: , by:

1. Multiplying each element of the input vector with the transpose of a corresponding vector of weights, i.e. of the same length as X, i.e. , …,
2. Summing the individual products of all the input features with their corresponding weights, and adding a bias term[[1]](#footnote-1):

Passing the result of step 2 (i.e. Z) through an “Activation Function” , where the function , is typically one of:

* 1. Rectified Linear Unit (ReLU):
  2. Sigmoid function:
  3. Hyperbolic Tangent (Tanh):

The result of is an estimate () for the outcome variable . And a simple cost function, such as the root mean squared error:

RMSE = ,

can be minimized via gradient descent to find the optimal values for and .

The original motivation for the use of the activation function is inspired by the way a biological neuron is triggered: it receives inputs from various other neurons, sums up their inputs and fires once a threshold is reached. Practically, in a deep learning set up non-linearities in the data can be captured by these activation functions, particularly when perceptrons are chained into multiple fully-connected layers, as is the case in Deep Learning, which we’ll discuss next.

#### Deep Learning

Deep Learning extends the concept of a single perceptron, by streaming a matrix of training observations (each of which is a vector of features, e.g. ) through multiple layers of perceptrons stacked together into what is called a hidden layer architecture.

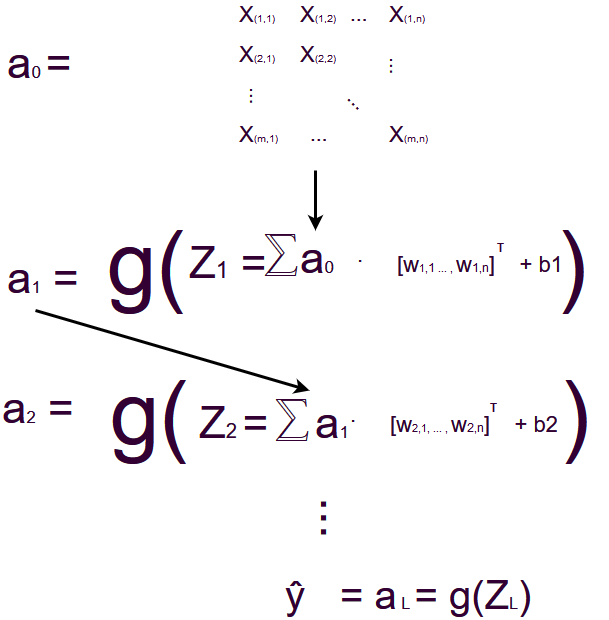
The output of each layer’s activation function becomes the input vector fed to the next layer, i.e. for = hidden layers:

Where is the result of the first activation function, i.e.

##### A Note on Notation:

The initial matrix of inputs can be denoted as , and the output of each subsequent layer’s activation function in the nested equation can also be written in notation.

#### Visual Intuition



**Figure 7: A visual Representation of the mathematics Explained Above (12)**

#### Backpropagation

Backpropagation is an iterative process of adjusting weights and biases, by minimizing a cost function , which can be understood as the error in the estimate tied to a specific value of weights and biases.

Each element of the weights and biases are typically updated proportionately to the partial derivative of the cost function with respect to that specific element, multiplied by a learning rate , i.e.

For the element of the weight matrix , is updated as:

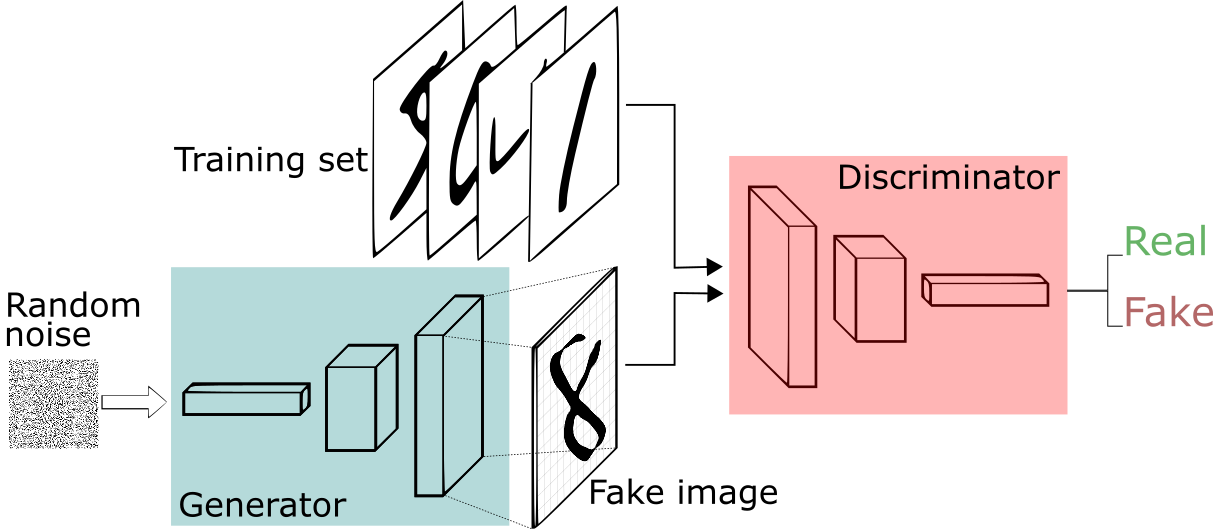
Similarly, for the element of the bias vector b, is updated as:

The training dataset is fed through the network multiple times (called epochs), hopefully allowing the cost function to reach a minimum.

#### Hyperparameters

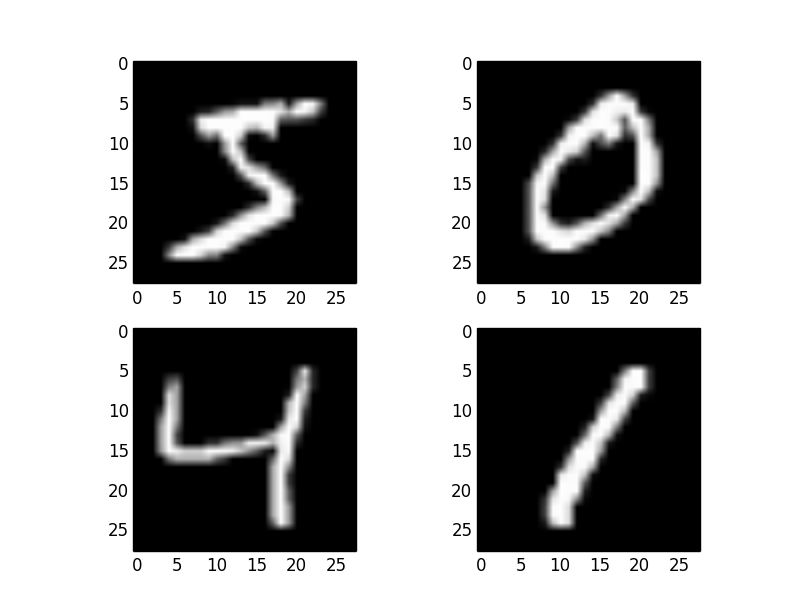
To arrive at an optimal neural network for the defined task, there are multiple hyperparameters that can be optimized. Some of these have already been mentioned in part, i.e. the number of epochs, the choice of activation function, the number of hidden layers and the number of perceptrons within each hidden layer, the distribution of weights to initialize the network with, the learning rate ,etc.; but there are many more, including regularization, dropout, etc., which will be covered in the implementation section of this thesis.

#### Generative Neural Networks Explained at the hand of Pseudocode, based on a minimal example from (13)



**Figure 8: Schematic Representation of a Generative Adverserial Network (14)**

Suppose we have a set of training images, each of which consists of 256 pixels, arranged in 28 rows and 28 columns; such as the well-known MNIST database of handwritten images (15), of which a sample is shown in ***Figure 7*** below:



**Figure 9: a sample of four handwritten digits from The MNIST Dataset (16)**

Our goal is to build a GAN architecture, using these images. GANs is explained in the following lines of pseudocode, with comments indicated in bold text, preceded by a double forward-slash (“//”):

##### Pseudocode

**//DEFINE THE SHAPE OF THE INPUT (AND THEREFORE OUTPUT) IMAGES:**

IMAGE.SHAPE = [ROWS=28,COLUMNS=28]

**//GENERATE A VECTOR RANDOM GAUSSIAN NOISE TO USE AS INPUT TO THE GENERATOR NETWORK:**

A\_0 = RANDOM.NOISE(SIZE=100)

**//DEFINE THE HIDDEN ARCHITECTURE OF THE GENERATOR FUNCTION:**

GENERATOR = SEQUENTIAL\_MODEL[

**//Z\_1 IS THE INPUT LAYER, WITH THE SAME DIMENSIONS AS THE RANDOM NOISE INPUT VECTOR**

Z\_1 = LAYER[NODES=256, INPUT = A\_0]

**//A\_1 IS THE RESULT OF APPLYING A “LEAKY RELU” ACTIVATION FUNCTION TO:**

**//Z\_1 = SUM(W\_T \* A\_0 + BIAS)**

A\_1 = LEAKY\_RELU(ALPHA = 0.2,

**//LEAKY RELU ACTIVATION IS AN INDICATOR FUNCTION,**

**DEFINED AS FOLLOWS:**

FUNCTION = IF Z\_1 > 0 THEN Z\_1 ELSE A\*Z\_1 END,

INPUT = Z\_1)

**//BATCH NORMALIZATION IS A WAY OF SCALING THE TRAINING SAMPLE FEAUTURES, BY REDUCING COVARIANCE SHIFT, AND AS A RESULT, SPEEDING UP LEARNING AND PREVENTING VANISHING OR EXPLODING GRADIENTS DURING BACKPROPAGATION TRAINING**

A\_1 = BATCH\_NORMALIZATION(A\_1)

**//A FURTHER TWO HIDDEN LAYERS ARE ADDED, EACH TIME DOUBLING THE AMOUNT OF PERCEPTRONS PER LAYER, AND APPLYING AN ACTIVATION FUNCTION, AS WELL AS BATCH NORMALIZATION TO EACH LAYER**

Z\_2 = LAYER[NODES=512, INPUT = A\_1]

A\_2 = LEAKY\_RELU(ALPHA = 0.2,

FUNCTION = IF Z\_2 > 0 THEN Z\_2 ELSE ALPHA\*Z\_2 END,

INPUT = Z\_2)

A\_2 = BATCH\_NORMALIZATION(A\_2)

Z\_3 = LAYER[NODES=1024, INPUT = A\_2]

A\_3 = LEAKY\_RELU(ALPHA = 0.2,

FUNCTION = IF Z\_3 > 0 THEN Z\_3 ELSE ALPHA\*Z\_3 END

INPUT = Z\_3)

A\_3 = BATCH\_NORMALIZATION(A\_3)

**//FINALLY, THE SECOND HIDDEN LAYER IS RESHAPED TO THE ORIGINAL DIMENSIONS (28\*28), AND A TANH ACTIVATION FUNCTION (AS DEFINED EARLIER IN THE MATHEMATICAL THEORY SECTION OF THIS THESIS) IS APPLIED TO IT**

Y\_HAT = TANH(A\_3.RESHAPE(IMAGE.SHAPE))

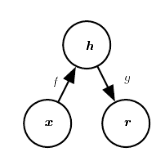
]

DISCRIMINATOR = ANOTHER NEURAL NETWORK OR OTHER FORM OF DISCRIMINATOR

For a defined number of epochs, a sample of both real (training example) and fake (generated by G) images are streamed through the Discriminator Network (D). The loss function of the D informs backpropagation through both networks.

## Autoencoders

Autoencoders are an implementation of artificial neural networks (ANNs) that map from a full representation of an input to a compressed representation (encoding), and then maps back from the compressed representation to the original input (decoding). It contains a hidden layer which learns a latent representation of the input, and the network as a whole consists of an encoder function and a decoder function that results in a reconstruction (17), see ***Figure 10***, below.



**Figure 10: The General structure of an autoencoder, illustrating how the hidden representation layer (h) is inferred from the input data by f(x); and how the input data is subsequently reconstructed by g(h) (17)**

Autoencoders are restricted so that they do not copy the input data exactly, to enable them to learn useful representations of the input data, i.e. the mapping functions are not deterministic, but contain a degree of stochasticity in their mapping protocol (17).

Since AEs are an extension of feedforward neural networks, they can similarly be trained using back-propagation based on mini-batch gradient descent, but unlike traditional feedforward ANNs, training can also be accomplished using recirculation (a method that compares activations in the encoding function to those in the reconstructing function) (17).

It is important to note that the output of the decoder function is generally not the end-goal of implementing an AE, but that arriving at a useful structure of , which encodes important features from the input data, is what motivates the usage of this algorithm. To this end, can be constrained to a smaller feature-space compared to , i.e. an undercomplete representation, which hopefully captures the most important latent variables that determine variability in .

Training an AE involves minimizing a loss function , which penalizes proportionally to the dissimilarity of its output to the original input (17).

### Variational Autoencoders (VAEs)

## Application of Traditional Deep Learning to High Energy Physics Problems

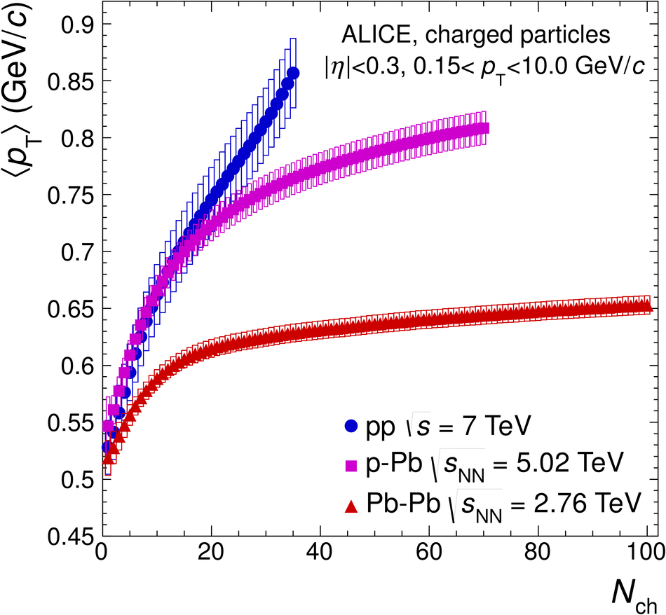
## Application of Generative Models to High Energy Physics Problems

# Motivation

## High Energy Physics Simulations

Detector response simulations is an integral part of all HEP experiments and high accuracy is attainable with first principle methods in current simulation environments such as GEANT4, but at a high computational cost (18).

This computational cost is particularly relevant to the ALICE experiment, where a much higher charged particle multiplicity is achieved, relative to average transverse momentum . In other words, many more particles are produced in heavy ion collisions () compared to proton-proton () collisions, which increases simulation time significantly (19) ←placeholder reference, since I am just guessing here... , see ***Figure 8*** below



**Figure 11: Average Charged Particle Multiplicity vs Average Transverse Momentum in p-p, p-Pb and Pb-Pb collisions at the LHC (20)**

# Research Question

## Primary Research Question

To what level of accuracy can generative machine learning algorithms, such as Generative Adversarial Networks and Variational Autoencoders be developed and trained, in order to simulate data generated by the ALICE TRD during High Energy Physics event simulations?

## Secondary Research Question

Can trained weights be sensibly extracted from individual neurons, or from hidden layers, of a Generative Neural Network, in order to infer:

* the latent variables discovered by the algorithm,
* their relative importance in outputting a detector simulation,
* and to what extent these latent variables map back to observable physical variables and processes, currently used in Monte Carlo simulations, based upon first principles

## Auxiliary Research Questions

1. Can an ensemble of Machine Learning Algorithms be employed successfully to optimize parameters in the existing Monte Carlo event simulation environment, Geant4?
2. Could the Discriminative Neural Network be replaced by a different classification algorithm, such as a support vector machine or an ensembled method, in order to arrive at a more accurate GAN architecture?
3. Could the random gaussian noise-source (z), used by G to simulate detector data, be replaced with a more sensible set of features from the original collision conditions, such as center of mass energy, environmental conditions, etc.

# Hypothesis

It is the hypothesis of this dissertation that Generative Machine Learning Algorithms, such as GANs, can be successfully be applied to HEP event simulations, at a lower computational cost than traditional methods currently being used, and that these algorithms can output data which is indistinguishable from actual data collected from the TRD at CERN.

# Aims & Objectives

1. **To build a highly accurate Neural Network that is able to classify particles passing through the TRD as: electrons, positrons, pions, etc.**
   1. By using h2o.ai (21) within the R statistical software environment (22)
2. **To optimize parameters for Monte Carlo event simulations within Geant4, in order to more accurately account for environmental conditions in the TRD at run-time, e.g. ambient temperature, atmospheric pressure, etc.,**
   1. using an ensembled approach of machine learning (ML) algorithms within h2o.ai (21) ecosystem
3. **To simulate Particle-Detector Interaction data,**
   1. By modelling the output data generated during High Energy Physics Collisions in the ALICE TRD,
   2. that is of sufficient quality so as to be indistinguishable from data generated by current Monte Carlo simulations, such as that generated by Geant4
   3. and that is indistinguishable from data taken from real collision events within the ALICE TRD at CERN
4. **To build an efficient “Proof of Concept” Generative Adversarial Network architecture to this end (point 3.),**
   1. By utilizing existing packages for Deep Learning, e.g. Keras for proof of concept, within the R statistical software environment,
   2. Using data from:
      1. real HEP experiments at ALICE,
         1. Which will be obtained from the WLCG storage system using AliEn, and parsed (using AliRoot) into a data format (.csv/ .json) that can be read into R
      2. Simulated event data from Geant4, with parameters tuned to emulate the effect of environmental variables as mentioned in point 2., above.
5. **To explore variational autoencoders (VAEs) as an alternative methodology for event simulations**
6. **To productionalize the most accurate ML simulations of event data (GANs, VAEs, an ensemble of the two, or something completely different)**
   1. Either by:
      1. recoding the chosen algorithm in C++, based upon first principles from linear algebra outlined in the Mathematical Theory section in the Background of this document; and utilizing the existing ROOT package for ML, the Toolkit for Multivariate Data Analysis (TMVA) to support the implementation of this
      2. Interfacing with ROOT from within R, using ROOT R, and setting up a RESTful API service, using the plumbeR package, to minimize additional dependencies at run time

# Methods

## Data Extraction from WLCG via AliEn

Using ROOT macros previously developed by collaborators at UCT Department of Physics <https://github.com/tdietel/trdML> ~ a data extraction task was executed across specific runs at the LHC, which created text files containing python dictionaries with collision event data stored, with the following basic structure:

0: {'Event': 2,

'V0TrackID': 35,

'track': 11,

'pdgCode': 211,

'det0': 312,

'row0': 5,

'col0': 55,

'layer0': [[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ],

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ],

[10, 15, 22, 18, 15, 16, 16, 16, 15, 12, 11, 15, 11, 11, 11, 10, 12, 11, 12, 11, 9, 10, 12, 10, ],

[8, 26, 44, 40, 34, 35, 59, 51, 39, 34, 33, 38, 39, 38, 31, 28, 24, 22, 27, 24, 24, 20, 18, 19, ],

[11, 10, 13, 14, 14, 17, 21, 19, 17, 16, 19, 26, 31, 28, 21, 22, 27, 30, 34, 35, 39, 31, 31, 31, ],

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ],

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ],

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ],

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ],

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ],

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ],

],},

## Deep Learning for Particle Identification (PID)

## Deep Learning towards Geant4 Parameter Optimization based on Environmental Conditions

## HEP Event Simulations

### Monte Carlo Simulation Data Generation using Geant4 with Parameters Relating to Environmental Conditions Optimized

### Generative Adversarial Networks in Keras

### Variational Autoencoders

## Productionalization of Machine Learning-Based HEP Event Simulations

# Results

# Discussion

# Conclusion and Future Work

# Bibliography

1. *The ALICE Transition Radiation Detector: Construction, operation and performance.* **Collaboration, ALICE.** 2018, Nuclear Inst. and Methods in Physics Research, Vol. 881, pp. 88-127.

2. **Reed, Rosi.** What is the QGP? *Quantum Diaries.* [Online] [Cited: 3 October 2018.] https://www.quantumdiaries.org/2011/11/23/what-is-the-qgp/.

3. *Performance of the ALICE Experiment.* **ALICE Collaboration.** s.l. : CERN-PH-EP, 2014, Vol. 031.

4. **More details on the ALICE TRD. *ALICE - CERN.* [Online] CERN. [Cited: 3 October 2018.] http://alice.web.cern.ch/detectors/more-details-alice-trd.**

**5. The Transition Radiation Detector. *ALICE.* [Online] [Cited: 4 October 2018.] http://aliceinfo.cern.ch/Public/Objects/Chapter2/DetectorComponents/trd\_operation.gif.**

**6. *ROOT -- A C++ framework for petabyte data storage, statistical analysis and visualization.* Antcheva, I., et al. 12, 2009, Computer Physics Communications, Vol. 180, pp. 2499-2512.**

**7. ALICE Collaboration. AliRoot Tutorial. *CERN Indico.* [Online] [Cited: 4 October 2018.] https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=2ahUKEwi748CmwezdAhVSasAKHUVkDwsQFjABegQICBAC&url=https%3A%2F%2Findico.cern.ch%2Fevent%2F6986%2Fcontribution%2F0%2Fattachments%2F1029652%2F1466277%2Ftutorial.pdf&usg=AOvVaw.**

**8. Integration with other languages. *ROOT.* [Online] [Cited: 4 October 2018.] https://root.cern.ch/integration-other-languages.**

**9. *Generative Adverserial Nets.* Goodfellow, Ian J, et al. 2014, stat.ML.**

**10. *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.* Rosenblatt, F. 6, 1958, The Psychological Review, Vol. 65.**

**11. From Perceptron to Deep Neural Nets. *Becoming Human AI.* [Online] [Cited: 4 October 2018.] https://becominghuman.ai/from-perceptron-to-deep-neural-nets-504b8ff616e.**

**12. Viljoen, Christiaan Gerhardus. *Created on draw.io.***

**13. Keras GAN. *Github.* [Online] [Cited: 4 October 2018.] https://github.com/eriklindernoren/Keras-GAN/blob/master/gan/gan.py.**

**14. A Beginner's Guide to Generative Adverserial Networks (GANs). *Skymind.* [Online] [Cited: 4 October 2018.] https://skymind.ai/wiki/generative-adversarial-network-gan.**

**15. LeCunn, Yan, Cortes, Corinna and Burges, Christopher. The MNIST Database of Handwritten Images. *Yan Lecunn.* [Online] http://yann.lecun.com/exdb/mnist/.**

**16. Handwritten Digit Recognition using Convolutional Neural Networks in Python with Keras. *Machine Learning Mastery.* [Online] https://machinelearningmastery.com/handwritten-digit-recognition-using-convolutional-neural-networks-python-keras/.**

**17. Goodfellow, Ian, Bengio, Yoshua and Courville, Aaron. *Deep Learning.* s.l. : The MIT Press, 2016. ISBN 9780262035613.**

**18. *Generative Models for Fast Cluster Simulations in the TPC for the ALICE Experiment.* Deja, Kamil, Trzcinski, Tomasz and Graczykowski, Lukasz.**

**20. *Multiplicity dependence of the average transverse momentum in pp, p-Pb, and Pb-Pb collisions at the LHC.* 2013, Phys. Lett., pp. 371-380.**

**21. H2O.ai.**

**22. R Core Team. R: A language and environment for statistical Computing. Vienna : s.n.**

**23. Karazeev, Anton. Generative Adversarial Networks (GANs): Engine and Applications. *Stats and Bots.* [Online] [Cited: 6 October 2018.] https://blog.statsbot.co/generative-adversarial-networks-gans-engine-and-applications-f96291965b47.**

# Appendix A

#!/usr/bin/env python3

# -\*- coding: utf-8 -\*-

"""

Created on Sun Dec 2 11:55:27 2018

@author: gerhard

"""

#in order to get a list of python dictionaries on the grid,

#the following commands were run in a terminal:

'''

@my\_pc\_terminal > ssh -Y -l \*myusername\* hep01.phy.uct.ac.za

@hep01\_terminal > alienv enter AliPhysics::latest

@hep01\_terminal\_with\_aliphysics\_initialized > aliensh

@alien\_shell\_terminal > find . pythonDict.txt

'''

#since the alien shell is not a bash terminal, it is not possible (or at least not straight-forward) to simply pipe the stdout to a file,

#therefore: the output of this alien command was manually copied and saved to a text file on my local machine, called 'dict\_loc.txt'

import os

#here I tested the viability of calling bash commands via python, but automated password submission didn't work

#import subprocess as shell

#test\_shell = str(shell.check\_output("ssh-pass -p password shh -Y -l gviljoen hep01.phy.uct.ac.za"))

#test\_shell = test\_shell.replace('b','').replace("'","").replace('\\n','')

#print(test\_shell)

#I therefore created a file from which I could copy and paste the contents into an alienshell:

os.chdir("/Users/gerhard/MSc-thesis/")

bash\_command\_1 = "initialize\_aliroot"

bash\_command\_2 = "alien\_token\_init cviljoen"

#bash\_command\_3 = "aliensh"

bash\_commands = [bash\_command\_1,bash\_command\_2]

with open("dict\_loc.txt","r") as f:

for line in f:

line = str(line)

new\_filename = line.replace('/alice/cern.ch/user/c/cviljoen/myWorkDir/myOutDir/', '').replace('/', '').replace('pythonDict.txt','').replace('\n','')

new\_filename = ''.join(new\_filename.split())

new\_filename = new\_filename+".txt"

location = "/home/gviljoen/Thesis\_New\_Data/"+new\_filename

find\_me = line

new\_bash = "cp " + find\_me + " file:" + location

new\_bash = new\_bash.replace(' \n','')

bash\_commands.append(new\_bash)

with open("bashcommands.txt", "a") as bashcommandfile:

for com in bash\_commands:

com = str(com+"\n")

bashcommandfile.write(com)

bashcommandfile.close()

#then, it was needed to append a '}' to the end of each file:

for filename in os.listdir("/Users/gerhard/MSc-thesis/SemiFullData"): # filename is a string

if filename.endswith(".txt"): # notice the indent

appendFile = open(filename, 'a') # file object, notice 'a' mode

appendString = "}" # could be done out of the loop if constant

appendFile.write(appendString)

appendFile.close()

# Appendix B

# -\*- coding: utf-8 -\*-

"""

This script extracts Python Dictionaries with Tracklet information

"""

import os

from ast import literal\_eval

import json

os.chdir("/Users/gerhard/MSc-thesis/SemiFullData")

dat\_files = os.listdir("/Users/gerhard/MSc-thesis/SemiFullData")

for i in range(0,len(dat\_files)):

if dat\_files[i].endswith(".txt"):

if len(dat\_files[i])==16:

print(dat\_files[i])

d = open(dat\_files[i])

d = d.read()

d = literal\_eval(d)

jayson = json.dumps(d,indent=4,sort\_keys=True)

name1="dat"

name2=".json"

name=name1+str(i)+name2

outfile = open(name,"w")

outfile.write(jayson)

# Appendix C

## Set up session for Deep Learning:

Set working directory:

#setwd("~/Documents/R")

Clear R environment objects:

rm(list=ls())

Read in JSON dump of Chris’s python dictionary and convert to three dataframes:

require(jsonlite)  
  
dat <- fromJSON("data/text.txt")  
  
zero <- dat[[1]]  
one <- dat[[2]]  
two <- dat[[3]]  
three <- dat[[4]]

Union these dataframes into one:

all.dat <- data.frame(  
 rbind(  
 zero,  
 one,  
 two,  
 three  
 )  
)

# Data Wrangling for Deep Learning:

require(dplyr)  
library(tidyr)  
  
#isolate the layer columns  
layers <- data.frame(all.dat[,c(8,12,16,20,24,28)])  
  
#choose an example that is not null  
fix <- layers[6,1]  
  
#fill it with zeros, retaining its shape  
for(i in 1:nrow(fix[[1]])){  
 for(j in 1:ncol(fix[[1]])){  
 fix[[1]][i,j] <- 0  
 }  
}  
  
#replace any null values with this matrix of zeroes  
nullToNA <- function(x) {  
 x[sapply(x, is.null)] <- fix  
 return(x)  
}  
  
layers <- as.matrix(layers)  
layers <- nullToNA(layers)  
  
layers <- cbind(all.dat[,1:3],layers)  
  
#unnest each layer individually:  
#//TODO: create a function for this  
  
#LAYER 0  
  
a <- rep(NA,264)  
  
for(i in layers$layer0){  
 add <- as.vector(unlist(i))  
 a <- rbind(a,add)  
}  
  
a <- a[-1,]  
  
layer0 <- cbind(all.dat[,c(1:7,9:11,13:15,17:19,21:23,25:27)],a)  
  
#LAYER 1  
  
a <- rep(NA,264)  
  
for(i in layers$layer1){  
 add <- as.vector(unlist(i))  
 a <- rbind(a,add)  
}  
  
a <- a[-1,]  
  
layer0 <- cbind(layer0,a)  
  
#LAYER 2  
  
a <- rep(NA,264)  
  
for(i in layers$layer2){  
 add <- as.vector(unlist(i))  
 a <- rbind(a,add)  
}  
  
a <- a[-1,]  
  
layer0 <- cbind(layer0,a)  
  
#LAYER 3  
  
a <- rep(NA,264)  
  
for(i in layers$layer3){  
 add <- as.vector(unlist(i))  
 a <- rbind(a,add)  
}  
  
a <- a[-1,]  
  
layer0 <- cbind(layer0,a)  
  
#LAYER 4  
  
a <- rep(NA,264)  
  
for(i in layers$layer4){  
 add <- as.vector(unlist(i))  
 a <- rbind(a,add)  
}  
  
a <- a[-1,]  
  
layer0 <- cbind(layer0,a)  
  
#LAYER 5  
  
a <- rep(NA,264)  
  
for(i in layers$layer5){  
 add <- as.vector(unlist(i))  
 a <- rbind(a,add)  
}  
  
a <- a[-1,]  
  
layer0 <- cbind(layer0,a)  
  
#make all missing values 0  
layer0[is.na(layer0)] <- 0  
  
my.dat <- layer0[,4:100]  
  
#make the outcome variable categorical  
my.dat$pdgCode <- as.factor(my.dat$pdgCode)  
  
#save the manipulated dataset and clear the R environment  
save(my.dat,file="NNdata.RData")  
rm(list=ls())

Scale the wrangled data and remove unnecessary elements:

#load the manipulated data set  
load("NNdata.RData")  
#scale numerical predictor variables  
my.dat[,-1] <- scale(my.dat[,-1])  
my.dat <- my.dat %>%  
 subset(pdgCode!=-11) %>%  
 subset(pdgCode!=11)

# DEEP LEARNING:

require(h2o)  
h2o.init(max\_mem\_size = "28G",nthreads = -1)

#h2o.no\_progress()

Upload data to H2O cluster, split into training (60%), validation (20%) and test (20%) sets:

dat.hex <- as.h2o(my.dat,"dat.hex")

splitz <- h2o.splitFrame(dat.hex,ratios=c(0.6,0.2),  
 destination\_frames = c("train.hex","valid.hex","test.hex"))  
  
train.hex <- splitz[[1]]  
valid.hex <- splitz[[2]]  
test.hex <- splitz[[3]]

Build first neural network:

* 2 hidden layers of 200 neurons each
* 100 epochs
* 10-fold cross validation

nn\_1 <- h2o.deeplearning(  
 x= 2:97,  
 y= 1,  
 model\_id = "nn\_1",  
 training\_frame = train.hex,  
 validation\_frame = valid.hex,  
 hidden = c(200,200),  
 nfolds=10,  
 standardize = F,  
 epochs=100,  
 fast\_mode = F,  
 sparse = T  
)

Build a differenct architecture:

* One hidden layer of 500 neurons
* 100 epochs
* 10-fold cross validation

nn\_2 <- h2o.deeplearning(  
 x= 2:97,  
 y= 1,  
 model\_id = "nn\_2",  
 training\_frame = train.hex,  
 validation\_frame = valid.hex,  
 hidden = 500,  
 nfolds=10,  
 standardize = F,  
 fast\_mode = F,  
 sparse = T,  
 epochs=100  
)

Build yet another differenct architecture:

* 10 hidden layers
* 100 epochs
* 10-fold cross validation

nn\_3 <- h2o.deeplearning(  
 x= 2:97,  
 y= 1,  
 model\_id = "nn\_3",  
 training\_frame = train.hex,  
 validation\_frame = valid.hex,  
 hidden = c(32,32,32,32,32,32,32,32,32,32),  
 nfolds=10,  
 standardize = F,  
 fast\_mode = F,  
 sparse = T,  
 epochs=100  
)

Find most probable architecture:

h2o.mean\_per\_class\_error(nn\_1)

## [1] 0.2241379

h2o.mean\_per\_class\_error(nn\_2)

## [1] 0.2155172

h2o.mean\_per\_class\_error(nn\_3)

## [1] 0.2241379

h2o.confusionMatrix(nn\_1)

## Confusion Matrix (vertical: actual; across: predicted) for max f1 @ threshold = 0.0230669511644981:  
## -211 211 Error Rate  
## -211 32 26 0.448276 =26/58  
## 211 0 66 0.000000 =0/66  
## Totals 32 92 0.209677 =26/124

h2o.confusionMatrix(nn\_2)

## Confusion Matrix (vertical: actual; across: predicted) for max f1 @ threshold = 0.120827021588128:  
## -211 211 Error Rate  
## -211 33 25 0.431034 =25/58  
## 211 0 66 0.000000 =0/66  
## Totals 33 91 0.201613 =25/124

h2o.confusionMatrix(nn\_3)

## Confusion Matrix (vertical: actual; across: predicted) for max f1 @ threshold = 0.566688777517043:  
## -211 211 Error Rate  
## -211 32 26 0.448276 =26/58  
## 211 0 66 0.000000 =0/66  
## Totals 32 92 0.209677 =26/124

h2o.confusionMatrix(nn\_1,newdata = test.hex)

## Confusion Matrix (vertical: actual; across: predicted) for max f1 @ threshold = 0.416494562377328:  
## -211 211 Error Rate  
## -211 6 10 0.625000 =10/16  
## 211 1 20 0.047619 =1/21  
## Totals 7 30 0.297297 =11/37

h2o.confusionMatrix(nn\_2,newdata = test.hex)

## Confusion Matrix (vertical: actual; across: predicted) for max f1 @ threshold = 0.120827021588128:  
## -211 211 Error Rate  
## -211 4 12 0.750000 =12/16  
## 211 0 21 0.000000 =0/21  
## Totals 4 33 0.324324 =12/37

h2o.confusionMatrix(nn\_3,newdata = test.hex)

## Confusion Matrix (vertical: actual; across: predicted) for max f1 @ threshold = 0.566688777517043:  
## -211 211 Error Rate  
## -211 6 10 0.625000 =10/16  
## 211 0 21 0.000000 =0/21  
## Totals 6 31 0.270270 =10/37

plot(nn\_1)



plot(nn\_2)



plot(nn\_3)



# Tuning:

The above information suggests fewer hidden layers, with more neurons is more suited to this problem.

We build a fourth neural network architecture: \* 2 hidden layers of 500 neurons each \* 200 epochs

And test this with various activation functions:

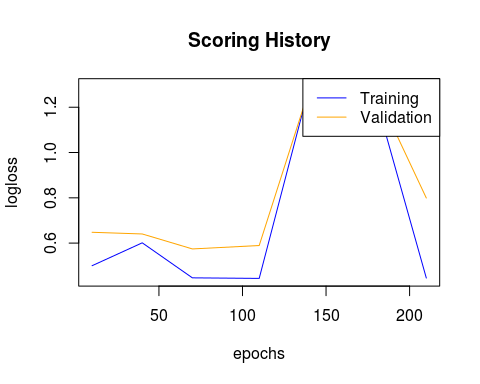
## Tanh with Dropout

nn\_4 <- h2o.deeplearning(  
 x= 2:97,  
 y= 1,  
 model\_id = "nn\_4",  
 training\_frame = train.hex,  
 validation\_frame = valid.hex,  
 activation = "TanhWithDropout",  
 hidden = c(500,500),  
 nfolds=10,  
 standardize = F,  
 fast\_mode = F,  
 sparse = T,  
 epochs=200,  
 balance\_classes = T  
)

h2o.performance(nn\_4,test.hex)

## H2OBinomialMetrics: deeplearning  
##   
## MSE: 0.3263933  
## RMSE: 0.5713084  
## LogLoss: 1.019133  
## Mean Per-Class Error: 0.34375  
## AUC: 0.6845238  
## Gini: 0.3690476  
##   
## Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:  
## -211 211 Error Rate  
## -211 5 11 0.687500 =11/16  
## 211 0 21 0.000000 =0/21  
## Totals 5 32 0.297297 =11/37  
##   
## Maximum Metrics: Maximum metrics at their respective thresholds  
## metric threshold value idx  
## 1 max f1 0.187928 0.792453 10  
## 2 max f2 0.187928 0.905172 10  
## 3 max f0point5 0.187928 0.704698 10  
## 4 max accuracy 0.187928 0.702703 10  
## 5 max precision 0.998246 1.000000 0  
## 6 max recall 0.187928 1.000000 10  
## 7 max specificity 0.998246 1.000000 0  
## 8 max absolute\_mcc 0.187928 0.452856 10  
## 9 max min\_per\_class\_accuracy 0.627600 0.333333 9  
## 10 max mean\_per\_class\_accuracy 0.187928 0.656250 10  
##   
## Gains/Lift Table: Extract with `h2o.gainsLift(<model>, <data>)` or `h2o.gainsLift(<model>, valid=<T/F>, xval=<T/F>)`

plot(nn\_4)



## Rectifier with Dropout:

nn\_5 <- h2o.deeplearning(  
 x= 2:97,  
 y= 1,  
 model\_id = "nn\_5",  
 training\_frame = train.hex,  
 validation\_frame = valid.hex,  
 activation = "RectifierWithDropout",  
 hidden = c(500,500),  
 nfolds=10,  
 standardize = F,  
 fast\_mode = F,  
 sparse = T,  
 epochs=200,  
 balance\_classes = T  
)

h2o.performance(nn\_5,test.hex)

## H2OBinomialMetrics: deeplearning  
##   
## MSE: 0.193127  
## RMSE: 0.4394622  
## LogLoss: 2.171004  
## Mean Per-Class Error: 0.3125  
## AUC: 0.7261905  
## Gini: 0.452381  
##   
## Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:  
## -211 211 Error Rate  
## -211 6 10 0.625000 =10/16  
## 211 0 21 0.000000 =0/21  
## Totals 6 31 0.270270 =10/37  
##   
## Maximum Metrics: Maximum metrics at their respective thresholds  
## metric threshold value idx  
## 1 max f1 0.586526 0.807692 7  
## 2 max f2 0.586526 0.913043 7  
## 3 max f0point5 0.586526 0.724138 7  
## 4 max accuracy 0.586526 0.729730 7  
## 5 max precision 1.000000 0.833333 3  
## 6 max recall 0.586526 1.000000 7  
## 7 max specificity 1.000000 0.937500 0  
## 8 max absolute\_mcc 0.586526 0.504016 7  
## 9 max min\_per\_class\_accuracy 0.586526 0.375000 7  
## 10 max mean\_per\_class\_accuracy 0.586526 0.687500 7  
##   
## Gains/Lift Table: Extract with `h2o.gainsLift(<model>, <data>)` or `h2o.gainsLift(<model>, valid=<T/F>, xval=<T/F>)`

plot(nn\_5)



# Maxout with Dropout:

nn\_6 <- h2o.deeplearning(  
 x= 2:97,  
 y= 1,  
 model\_id = "nn\_6",  
 training\_frame = train.hex,  
 validation\_frame = valid.hex,  
 activation = "MaxoutWithDropout",  
 hidden = c(500,500),  
 nfolds=10,  
 standardize = F,  
 fast\_mode = F,  
 sparse = T,  
 epochs=200,  
 balance\_classes = T  
)

h2o.performance(nn\_6,test.hex)

## H2OBinomialMetrics: deeplearning  
##   
## MSE: 0.4449376  
## RMSE: 0.6670364  
## LogLoss: 2.888669  
## Mean Per-Class Error: 0.3988095  
## AUC: 0.610119  
## Gini: 0.2202381  
##   
## Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:  
## -211 211 Error Rate  
## -211 4 12 0.750000 =12/16  
## 211 1 20 0.047619 =1/21  
## Totals 5 32 0.351351 =13/37  
##   
## Maximum Metrics: Maximum metrics at their respective thresholds  
## metric threshold value idx  
## 1 max f1 0.019517 0.754717 10  
## 2 max f2 0.000000 0.867769 15  
## 3 max f0point5 0.019517 0.671141 10  
## 4 max accuracy 0.019517 0.648649 10  
## 5 max precision 1.000000 1.000000 0  
## 6 max recall 0.000000 1.000000 15  
## 7 max specificity 1.000000 1.000000 0  
## 8 max absolute\_mcc 0.999996 0.303895 3  
## 9 max min\_per\_class\_accuracy 0.998602 0.285714 7  
## 10 max mean\_per\_class\_accuracy 0.019517 0.601190 10  
##   
## Gains/Lift Table: Extract with `h2o.gainsLift(<model>, <data>)` or `h2o.gainsLift(<model>, valid=<T/F>, xval=<T/F>)`

plot(nn\_6)



# h2o.saveModel(nn\_1,"nn1.RData")  
# h2o.saveModel(nn\_2,"nn2.RData")  
# h2o.saveModel(nn\_3,"nn3.RData")  
# h2o.saveModel(nn\_4,"nn4.RData")  
# h2o.saveModel(nn\_5,"nn5.RData")  
# h2o.saveModel(nn\_6,"nn6.RData")

h2o.performance(nn\_1,dat.hex)

## H2OBinomialMetrics: deeplearning  
##   
## MSE: 0.1625316  
## RMSE: 0.4031521  
## LogLoss: 0.8764427  
## Mean Per-Class Error: 0.2624625  
## AUC: 0.8447447  
## Gini: 0.6894895  
##   
## Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:  
## -211 211 Error Rate  
## -211 46 44 0.488889 =44/90  
## 211 4 107 0.036036 =4/111  
## Totals 50 151 0.238806 =48/201  
##   
## Maximum Metrics: Maximum metrics at their respective thresholds  
## metric threshold value idx  
## 1 max f1 0.023067 0.816794 44  
## 2 max f2 0.000234 0.908333 49  
## 3 max f0point5 0.542199 0.795107 41  
## 4 max accuracy 0.416495 0.761194 42  
## 5 max precision 0.542199 0.962963 41  
## 6 max recall 0.000000 1.000000 72  
## 7 max specificity 1.000000 0.988889 0  
## 8 max absolute\_mcc 0.000234 0.548460 49  
## 9 max min\_per\_class\_accuracy 0.416495 0.522222 42  
## 10 max mean\_per\_class\_accuracy 0.416495 0.738589 42  
##   
## Gains/Lift Table: Extract with `h2o.gainsLift(<model>, <data>)` or `h2o.gainsLift(<model>, valid=<T/F>, xval=<T/F>)`

h2o.performance(nn\_2,dat.hex)

## H2OBinomialMetrics: deeplearning  
##   
## MSE: 0.2387207  
## RMSE: 0.4885905  
## LogLoss: 0.9372256  
## Mean Per-Class Error: 0.2711712  
## AUC: 0.8333333  
## Gini: 0.6666667  
##   
## Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:  
## -211 211 Error Rate  
## -211 42 48 0.533333 =48/90  
## 211 1 110 0.009009 =1/111  
## Totals 43 158 0.243781 =49/201  
##   
## Maximum Metrics: Maximum metrics at their respective thresholds  
## metric threshold value idx  
## 1 max f1 0.120827 0.817844 59  
## 2 max f2 0.120827 0.913621 59  
## 3 max f0point5 0.862533 0.796460 52  
## 4 max accuracy 0.120827 0.756219 59  
## 5 max precision 0.985880 0.959184 44  
## 6 max recall 0.000173 1.000000 85  
## 7 max specificity 1.000000 0.988889 0  
## 8 max absolute\_mcc 0.120827 0.554959 59  
## 9 max min\_per\_class\_accuracy 0.551938 0.504505 56  
## 10 max mean\_per\_class\_accuracy 0.120827 0.728829 59  
##   
## Gains/Lift Table: Extract with `h2o.gainsLift(<model>, <data>)` or `h2o.gainsLift(<model>, valid=<T/F>, xval=<T/F>)`

h2o.performance(nn\_3,dat.hex)

## H2OBinomialMetrics: deeplearning  
##   
## MSE: 0.1475263  
## RMSE: 0.3840915  
## LogLoss: 0.658837  
## Mean Per-Class Error: 0.2701201  
## AUC: 0.8379379  
## Gini: 0.6758759  
##   
## Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:  
## -211 211 Error Rate  
## -211 43 47 0.522222 =47/90  
## 211 2 109 0.018018 =2/111  
## Totals 45 156 0.243781 =49/201  
##   
## Maximum Metrics: Maximum metrics at their respective thresholds  
## metric threshold value idx  
## 1 max f1 0.566689 0.816479 61  
## 2 max f2 0.566689 0.908333 61  
## 3 max f0point5 0.954867 0.805970 55  
## 4 max accuracy 0.566689 0.756219 61  
## 5 max precision 0.999615 0.981132 52  
## 6 max recall 0.000000 1.000000 89  
## 7 max specificity 1.000000 0.988889 0  
## 8 max absolute\_mcc 0.566689 0.548460 61  
## 9 max min\_per\_class\_accuracy 0.903940 0.495495 57  
## 10 max mean\_per\_class\_accuracy 0.954867 0.732132 55  
##   
## Gains/Lift Table: Extract with `h2o.gainsLift(<model>, <data>)` or `h2o.gainsLift(<model>, valid=<T/F>, xval=<T/F>)`

h2o.performance(nn\_4,dat.hex)

## H2OBinomialMetrics: deeplearning  
##   
## MSE: 0.2196784  
## RMSE: 0.4686986  
## LogLoss: 0.6327696  
## Mean Per-Class Error: 0.2711712  
## AUC: 0.8318318  
## Gini: 0.6636637  
##   
## Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:  
## -211 211 Error Rate  
## -211 42 48 0.533333 =48/90  
## 211 1 110 0.009009 =1/111  
## Totals 43 158 0.243781 =49/201  
##   
## Maximum Metrics: Maximum metrics at their respective thresholds  
## metric threshold value idx  
## 1 max f1 0.155375 0.817844 63  
## 2 max f2 0.155375 0.913621 63  
## 3 max f0point5 0.746435 0.785498 54  
## 4 max accuracy 0.155375 0.756219 63  
## 5 max precision 0.998246 1.000000 0  
## 6 max recall 0.009154 1.000000 84  
## 7 max specificity 0.998246 1.000000 0  
## 8 max absolute\_mcc 0.155375 0.554959 63  
## 9 max min\_per\_class\_accuracy 0.226987 0.495495 61  
## 10 max mean\_per\_class\_accuracy 0.155375 0.728829 63  
##   
## Gains/Lift Table: Extract with `h2o.gainsLift(<model>, <data>)` or `h2o.gainsLift(<model>, valid=<T/F>, xval=<T/F>)`

h2o.performance(nn\_5,dat.hex)

## H2OBinomialMetrics: deeplearning  
##   
## MSE: 0.1436784  
## RMSE: 0.3790493  
## LogLoss: 0.7531213  
## Mean Per-Class Error: 0.2656156  
## AUC: 0.8436436  
## Gini: 0.6872873  
##   
## Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:  
## -211 211 Error Rate  
## -211 43 47 0.522222 =47/90  
## 211 1 110 0.009009 =1/111  
## Totals 44 157 0.238806 =48/201  
##   
## Maximum Metrics: Maximum metrics at their respective thresholds  
## metric threshold value idx  
## 1 max f1 0.084465 0.820896 46  
## 2 max f2 0.084465 0.915141 46  
## 3 max f0point5 0.956244 0.801749 41  
## 4 max accuracy 0.586526 0.761194 44  
## 5 max precision 1.000000 0.960000 8  
## 6 max recall 0.005610 1.000000 52  
## 7 max specificity 1.000000 0.988889 0  
## 8 max absolute\_mcc 0.084465 0.563722 46  
## 9 max min\_per\_class\_accuracy 0.956244 0.495495 41  
## 10 max mean\_per\_class\_accuracy 0.586526 0.735435 44  
##   
## Gains/Lift Table: Extract with `h2o.gainsLift(<model>, <data>)` or `h2o.gainsLift(<model>, valid=<T/F>, xval=<T/F>)`

h2o.performance(nn\_6,dat.hex)

## H2OBinomialMetrics: deeplearning  
##   
## MSE: 0.2847582  
## RMSE: 0.5336274  
## LogLoss: 1.411247  
## Mean Per-Class Error: 0.2746246  
## AUC: 0.832032  
## Gini: 0.6640641  
##   
## Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:  
## -211 211 Error Rate  
## -211 43 47 0.522222 =47/90  
## 211 3 108 0.027027 =3/111  
## Totals 46 155 0.248756 =50/201  
##   
## Maximum Metrics: Maximum metrics at their respective thresholds  
## metric threshold value idx  
## 1 max f1 0.019517 0.812030 59  
## 2 max f2 0.019517 0.901503 59  
## 3 max f0point5 0.514351 0.805970 54  
## 4 max accuracy 0.019517 0.751244 59  
## 5 max precision 1.000000 1.000000 0  
## 6 max recall 0.000000 1.000000 104  
## 7 max specificity 1.000000 1.000000 0  
## 8 max absolute\_mcc 0.019517 0.533549 59  
## 9 max min\_per\_class\_accuracy 0.514351 0.486486 54  
## 10 max mean\_per\_class\_accuracy 0.514351 0.732132 54  
##   
## Gains/Lift Table: Extract with `h2o.gainsLift(<model>, <data>)` or `h2o.gainsLift(<model>, valid=<T/F>, xval=<T/F>)`

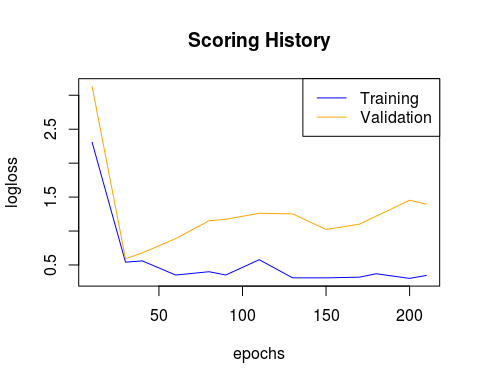
# Recitfier with Dropout with 2 additional hidden layers of 500 neurons each:

nn\_7 <- h2o.deeplearning(  
 x= 2:97,  
 y= 1,  
 model\_id = "nn\_7",  
 training\_frame = train.hex,  
 validation\_frame = valid.hex,  
 activation = "RectifierWithDropout",  
 hidden = c(500,500,500,500),  
 nfolds=10,  
 standardize = F,  
 fast\_mode = F,  
 sparse = T,  
 epochs=200,  
 balance\_classes = T,  
 l1=1e-06  
)

# h2o.saveModel(nn\_7,"nn7.RData")  
  
h2o.performance(nn\_7,dat.hex)

## H2OBinomialMetrics: deeplearning  
##   
## MSE: 0.1582569  
## RMSE: 0.3978151  
## LogLoss: 0.9577213  
## Mean Per-Class Error: 0.2711712  
## AUC: 0.8334334  
## Gini: 0.6668669  
##   
## Confusion Matrix (vertical: actual; across: predicted) for F1-optimal threshold:  
## -211 211 Error Rate  
## -211 42 48 0.533333 =48/90  
## 211 1 110 0.009009 =1/111  
## Totals 43 158 0.243781 =49/201  
##   
## Maximum Metrics: Maximum metrics at their respective thresholds  
## metric threshold value idx  
## 1 max f1 0.695753 0.817844 51  
## 2 max f2 0.695753 0.913621 51  
## 3 max f0point5 1.000000 0.796460 44  
## 4 max accuracy 0.695753 0.756219 51  
## 5 max precision 1.000000 0.951220 28  
## 6 max recall 0.000113 1.000000 60  
## 7 max specificity 1.000000 0.977778 0  
## 8 max absolute\_mcc 0.695753 0.554959 51  
## 9 max min\_per\_class\_accuracy 0.934974 0.504505 48  
## 10 max mean\_per\_class\_accuracy 0.695753 0.728829 51  
##   
## Gains/Lift Table: Extract with `h2o.gainsLift(<model>, <data>)` or `h2o.gainsLift(<model>, valid=<T/F>, xval=<T/F>)`

plot(nn\_7)



# Ensemble:

Add the predictions from all 7 neural networks as features to dataset:

# nn\_1 <- h2o.loadModel("nn\_1")  
# nn\_2 <- h2o.loadModel("nn\_2")  
# nn\_3 <- h2o.loadModel("nn\_3")  
# nn\_4 <- h2o.loadModel("nn\_4")  
# nn\_5 <- h2o.loadModel("nn\_5")  
# nn\_6 <- h2o.loadModel("nn\_6")  
# nn\_7 <- h2o.loadModel("nn\_7")  
#   
p1 <- as.data.frame(h2o.predict(nn\_1,dat.hex))

p2 <- as.data.frame(h2o.predict(nn\_2,dat.hex))

p3 <- as.data.frame(h2o.predict(nn\_3,dat.hex))

p4 <- as.data.frame(h2o.predict(nn\_4,dat.hex))

p5 <- as.data.frame(h2o.predict(nn\_5,dat.hex))

p6 <- as.data.frame(h2o.predict(nn\_6,dat.hex))

p7 <- as.data.frame(h2o.predict(nn\_7,dat.hex))

p1 <- p1[,3]  
p2 <- p2[,3]  
p3 <- p3[,3]  
p4 <- p4[,3]  
p5 <- p5[,3]  
p6 <- p6[,3]  
p7 <- p7[,3]  
  
p <- data.frame(cbind(p1,p2,p3,p4,p5,p6,p7))  
  
p <- scale(p)  
  
my.dat <- data.frame(cbind(my.dat,p))

h2o.shutdown(prompt=F)

## [1] TRUE

# h2o.init(max\_mem\_size = "28G",nthreads = -1)  
# h2o.removeAll()  
# #h2o.no\_progress()  
#   
# dat.hex <- as.h2o(my.dat,"dat.hex")  
#   
# splitz <- h2o.splitFrame(dat.hex,ratios=c(0.6,0.2),  
# destination\_frames = c("train.hex","valid.hex","test.hex"))  
#   
# train.hex <- splitz[[1]]  
# valid.hex <- splitz[[2]]  
# test.hex <- splitz[[3]]

## Random Forests:

Predict using randomforests, and add its prediction as a feature:

# rf1 <- h2o.randomForest(y=1,  
# x=2:111,  
# training\_frame=train.hex,  
# validation\_frame=valid.hex,  
# nfolds=30,  
# ntrees=200)  
#   
# h2o.performance(rf1,dat.hex)  
#   
# p8 <- as.data.frame(h2o.predict(rf1,dat.hex))  
#   
# plot(rf1)  
#   
# p8 <- p8[,3]  
# my.dat <- data.frame(my.dat,p8)  
# #   
# h2o.shutdown(prompt = F)

## K-means clustering

Perform k-means clustering, with k=10, and add cluster group as a variable:

rm(list=ls())  
load("mydat.RData")  
k <- kmeans(x=my.dat[,-1],centers=10)  
k <- k$cluster  
k <- as.factor(k)  
  
require(dummies)  
  
k <- dummy(k)  
  
my.dat <- data.frame(cbind(my.dat,k))  
  
my.dat$k1 <- as.factor(my.dat$k1)  
my.dat$k2 <- as.factor(my.dat$k2)  
my.dat$k3 <- as.factor(my.dat$k3)  
my.dat$k4 <- as.factor(my.dat$k4)  
my.dat$k5 <- as.factor(my.dat$k5)  
my.dat$k6 <- as.factor(my.dat$k6)  
my.dat$k7 <- as.factor(my.dat$k7)  
my.dat$k8 <- as.factor(my.dat$k8)  
my.dat$k9 <- as.factor(my.dat$k9)  
my.dat$k10 <- as.factor(my.dat$k10)  
save(my.dat,file="mydat.RData")

rm(list=ls())  
load("mydat.RData")

## Support Vector Machines

Predict using various kernels and add as features:

Linear Kernel:

require(e1071)  
sv <- svm(pdgCode~.,data=my.dat,scale=F,kernel="linear")  
svm.p <- data.frame(predict(sv,my.dat))

Polynomial Kernel:

sv <- svm(pdgCode~.,data=my.dat,scale=F,kernel="polynomial")  
svm.p <- data.frame(cbind(svm.p,predict(sv,my.dat)))

Radial Kernel

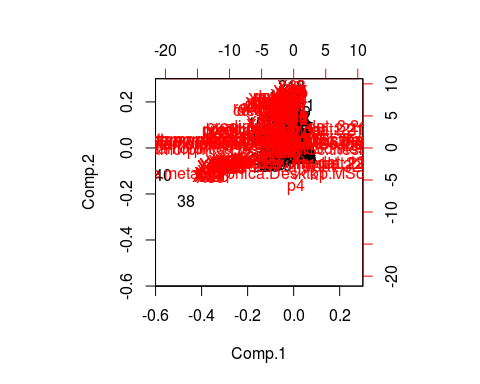
require(e1071)  
sv <- svm(pdgCode~.,data=my.dat,scale=F,kernel="radial")  
svm.p <- data.frame(cbind(svm.p,predict(sv,my.dat)))

Sigmoid Kernel:

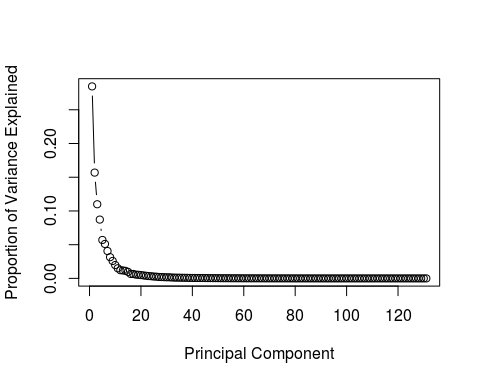
require(e1071)  
sv <- svm(pdgCode~.,data=my.dat,scale=F,kernel="sigmoid")  
svm.p <- data.frame(cbind(svm.p,predict(sv,my.dat)))

require(dummies)  
svm.p <- as.data.frame(svm.p)  
svm.p <- dummy.data.frame(svm.p)  
my.dat <- data.frame(cbind(my.dat,svm.p))

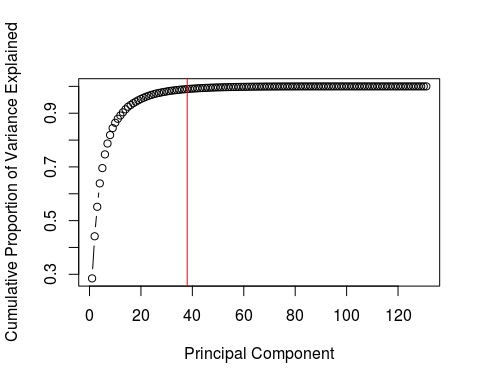
my.dat$k1 <- as.numeric(my.dat$k1)  
my.dat$k2 <- as.numeric(my.dat$k2)  
my.dat$k3 <- as.numeric(my.dat$k3)  
my.dat$k4 <- as.numeric(my.dat$k4)  
my.dat$k5 <- as.numeric(my.dat$k5)  
my.dat$k6 <- as.numeric(my.dat$k6)  
my.dat$k7 <- as.numeric(my.dat$k7)  
my.dat$k8 <- as.numeric(my.dat$k8)  
my.dat$k9 <- as.numeric(my.dat$k9)  
my.dat$k10 <- as.numeric(my.dat$k10)  
  
pc <- princomp(my.dat[,-1])  
biplot(pc)



std\_dev <- pc$sdev  
  
pr\_var <- std\_dev^2  
  
prop\_varex <- pr\_var/sum(pr\_var)  
  
plot(prop\_varex, xlab = "Principal Component",  
 ylab = "Proportion of Variance Explained",  
 type = "b")



plot(cumsum(prop\_varex), xlab = "Principal Component",  
 ylab = "Cumulative Proportion of Variance Explained",  
 type = "b")  
  
abline(v=min(which(cumsum(prop\_varex)>=.99)),col="red")



min(which(cumsum(prop\_varex)>=.99))

## [1] 38

new.dat <- data.frame(cbind(as.character(my.dat$pdgCode),pc$scores[,1:38]))  
new.dat$V1 <- as.factor(new.dat$V1)  
names(new.dat)[1] <- "pdgCode"  
  
for(i in 2:ncol(new.dat)){  
 new.dat[,i] <- as.numeric(as.character(new.dat[,i]))  
}

# Final prediction

Build a random forest that uses all the new features, as well as the original dataset to classify particles as pdgcode 211 or pdgcode -211:

require(randomForest)  
  
rf <- randomForest(x=new.dat[1:100,2:39],y=new.dat$pdgCode[1:100])  
  
pp <- predict(rf,newdata = new.dat)  
  
pp <- data.frame(cbind(as.factor(new.dat$pdgCode),as.factor(pp)))  
pp[,3] <- pp[,1]==pp[,2]  
  
length(which(pp[,3])==T)/nrow(pp)

## [1] 0.7562189

# Appendix D

require(bigints)

# Pad Response Function

The surface charge density induced on the metal surface of the cathode pad plane, by a charge at a distance from the pad plane, is given by:

Integrating this charge density over the metal surface, gives the total charge induced on the cathode pad plane, as follows:

Using this approach of integration over each pad

# Remove entries with missing pad data:

## [1] "Removed 72305 entries"

# Get a Look-Up Table for Tracks within Events:

* Use EventID, trackID, v0ID, and PDG code to get the indices in the full event list that pertain to a specific track:

events <- sapply(j,`[[`,"Event")  
events <- as.numeric(events)  
  
tracks <- sapply(j, `[[`,"track")  
tracks <- as.numeric(tracks)  
  
v0s <- sapply(j, `[[`,"V0TrackID")  
v0s <- as.numeric(v0s)  
  
pdg <- sapply(j, `[[`,"pdgCode")  
pdg <- as.numeric(pdg)

* Seperate out the detector traces for these events:

detectors <- sapply(j, `[[`,"det0")  
  
columns <- sapply(j, `[[`,"col0")  
  
rows <- sapply(j,`[[`,"row0")  
  
pads <- sapply(j,`[[`,"layer0")

* Create the Look-Up Tables for electrons and pions:

# look.up.table <- data.frame(cbind(events,tracks,v0s,pdg))  
# look.up.table$index <- 1:nrow(look.up.table)  
  
look.up.table <- data.frame(cbind(pdg,1:length(pdg)))  
names(look.up.table) <- c("pdg","index")  
  
electrons <- look.up.table[which(look.up.table$pdg %in% pdg.elec),]  
pions <- look.up.table[which(look.up.table$pdg %in% pdg.pion),]  
  
electrons <- unique(electrons)  
pions <- unique(pions)

* Create a unique identifier for a specific track, by concatenating its eventID and trackID, and get the number of detector hits for that unique ID:

# Appendix E

require(tufte)

### Set seed for reproducibility:

set.seed(123456789)

# Read in all the json files, created from python dictionaries

rm(list=ls())  
  
require(jsonlite)  
require(readtext)  
  
#PDG codes  
pdg.elec <- c(11,-11)  
pdg.pion <- c(-211,211)  
  
files <- list.files(path="~/Thesis data/SemiFullData", pattern="\*json", full.names=T, recursive=FALSE)  
  
j <- fromJSON(files[1])  
  
for(i in 2:length(files)){  
  
 f <- fromJSON(files[i])  
 j <- c(j,f)  
}  
  
length(j)  
  
save(j,file="~/Thesis data/SemiFullData/fulljson.rdata")

# Threshold pad data

Load full json record of all extracted events:

rm(list=ls())  
load("~/Thesis data/SemiFullData/fulljson.rdata")

Extract pad data

pads <- sapply(j,`[[`,"layer0")

Replace all pads that have data, with only those bin-rows which are all non-zero

no.zero.row.pads <- list()  
  
for(i in 1:length(pads)){  
 if(is.null(pads[[i]])){  
 no.zero.row.pads[[i]] <- NULL  
 }else if(typeof(pads[[i]])=="list"){  
 no.zero.row.pads[[i]] <- NULL  
 }else{  
 this.pad <- as.matrix(pads[[i]])  
 r <- rowSums(this.pad)  
 r <- which(r==0)  
  
 if(length(r)==11){  
 this.pad <- NULL  
 }else{  
 this.pad <- this.pad[-r,]  
 }  
  
  
  
 no.zero.row.pads[[i]] <- this.pad  
 }  
}  
  
save(no.zero.row.pads,file="~/Thesis data/SemiFullData/no\_zero\_row\_pads.rdata")

Create look-up tables to find pad data for both electrons and pions, and add an indicator column for which entries have no pad data.

Also create an indicator of which tracks’ trackIDs are 0, since this may indicate that an error has occurred

tracks <- sapply(j, `[[`,"track")  
tracks <- as.numeric(tracks)  
  
  
pdg <- sapply(j, `[[`,"pdgCode")  
pdg <- as.numeric(pdg)  
  
exclude <- rep(0,length(no.zero.row.pads))  
  
for(i in 1:length(no.zero.row.pads)){  
 if(is.null(no.zero.row.pads[[i]])){  
 exclude[i] <- 1  
 }  
}  
  
exclude2 <- rep(0,length(no.zero.row.pads))  
  
for(i in 1:length(tracks)){  
 if(tracks[i]==0){  
 exclude2[i] <- 1  
 }  
}  
  
  
look.up.table <- data.frame(cbind(pdg,1:length(pdg),exclude,exclude2))  
names(look.up.table) <- c("pdg","index","exclude.null.pads","exclude.track.id.0")

require(dplyr)  
pdg.elec <- c(11,-11)  
pdg.pion <- c(211,-211)  
electrons <- look.up.table[which(look.up.table$pdg %in% pdg.elec),]  
pions <- look.up.table[which(look.up.table$pdg %in% pdg.pion),]  
  
electrons <- unique(electrons)  
pions <- unique(pions)  
  
electrons <- electrons %>%  
 subset(exclude.null.pads==0) %>%  
 subset(exclude.track.id.0==0)  
  
pions <- pions %>%  
 subset(exclude.null.pads==0) %>%  
 subset(exclude.track.id.0==0)  
  
electrons <- as.data.frame(electrons)  
pions <- as.data.frame(pions)  
  
electrons <- electrons$index  
pions <- pions$index  
  
save(electrons,file="~/Thesis data/SemiFullData/electrons.rdata")  
save(pions,file="~/Thesis data/SemiFullData/pions.rdata")  
save(pads,file="~/Thesis data/SemiFullData/pads.rdata")

# Descriptive statistics

Get total energy deposition per pad, for pions and electrons, respectively:

electron.pad.sum <- c()  
  
for(i in electrons){  
 this.pad.sum <- sum(as.numeric(pads[[i]]))  
 electron.pad.sum <- c(electron.pad.sum,this.pad.sum)  
}  
  
pion.pad.sum <- c()  
  
for(i in pions){  
 this.pad.sum <- sum(as.numeric(pads[[i]]))  
 pion.pad.sum <- c(pion.pad.sum,this.pad.sum)  
}

electrons <- data.frame(cbind(electrons,electron.pad.sum))  
pions <- data.frame(cbind(pions,pion.pad.sum))  
  
names(electrons) <- c("index","total.charge.deposit")  
names(pions) <- c("index","total.charge.deposit")

Get 5 number summary statistics, for each pad where 0-sum bin-rows have been removed.

electron.pad.summary <- numeric(6)  
  
for(i in electrons$index){  
 this.pad <- summary(as.numeric(no.zero.row.pads[[i]]))  
 electron.pad.summary <- rbind(electron.pad.summary,this.pad)  
}  
  
electron.pad.summary <- as.data.frame(electron.pad.summary)  
names(electron.pad.summary) <- c("Min","1Q","Median","Mean","3Q","Max")  
  
pion.pad.summary <- numeric(6)  
  
for(i in pions$index){  
 this.pad <- summary(as.numeric(no.zero.row.pads[[i]]))  
 pion.pad.summary <- rbind(pion.pad.summary,this.pad)  
}  
  
pion.pad.summary <- as.data.frame(pion.pad.summary)  
names(pion.pad.summary) <- c("Min","1Q","Median","Mean","3Q","Max")

Clean up:

electron.pad.summary <- electron.pad.summary[-1,]  
pion.pad.summary <- pion.pad.summary[-1,]  
  
electrons <- data.frame(cbind(electrons,electron.pad.summary))  
pions <- data.frame(cbind(pions,pion.pad.summary))  
  
  
save(electrons,file="~/Thesis data/SemiFullData/electrons.rdata")  
save(pions,file="~/Thesis data/SemiFullData/pions.rdata")

Get the number of non-zero bin-rows per detector pad, for electrons and pions:

electron.timebins <- c()  
  
for(i in electrons$index){  
  
 if(is.null(nrow(no.zero.row.pads[[i]]))){  
 this.num.bin <- 0  
 }else{  
 this.num.bin <- nrow(no.zero.row.pads[[i]])  
 }  
  
  
 electron.timebins <- c(electron.timebins,this.num.bin)  
  
}  
  
pion.timebins <- c()  
  
for(i in pions$index){  
  
 if(is.null(nrow(no.zero.row.pads[[i]]))){  
 this.num.bin <- 0  
 }else{  
 this.num.bin <- nrow(no.zero.row.pads[[i]])  
  
  
 }  
  
 pion.timebins <- c(pion.timebins,this.num.bin)  
}  
  
  
electrons$num.bins <- electron.timebins  
pions$num.bins <- pion.timebins  
  
save(electrons,file="~/Thesis data/SemiFullData/electrons.rdata")  
save(pions,file="~/Thesis data/SemiFullData/pions.rdata")

Get the column means for each pad, where non-zero time-bins have been removed:

electron.compressed.bins <- numeric(24)  
  
for(i in electrons$index){  
  
 if(is.null(dim(no.zero.row.pads[[i]]))){  
 this.entry <- as.numeric(no.zero.row.pads[[i]])  
 }else{  
  
this.entry <- as.numeric(colSums(no.zero.row.pads[[i]]))  
  
  
 }  
  
 electron.compressed.bins <- rbind(electron.compressed.bins,this.entry)  
}  
  
pion.compressed.bins <- c()  
  
for(i in pions$index){  
  
 if(is.null(dim(no.zero.row.pads[[i]]))){  
 this.entry <- as.numeric(no.zero.row.pads[[i]])  
 }else{  
  
 this.entry <- as.numeric(colSums(no.zero.row.pads[[i]]))  
  
 }  
 pion.compressed.bins <- rbind(pion.compressed.bins,this.entry)  
}  
  
electron.compressed.bins <- electron.compressed.bins[-1,]  
  
pion.compressed.bins <- pion.compressed.bins[-1,]  
  
electron.compressed.bins <- as.data.frame(electron.compressed.bins)  
pion.compressed.bins <- as.data.frame(pion.compressed.bins)  
  
names <- paste0("c",1:24)  
  
names(electron.compressed.bins) <- names  
names(pion.compressed.bins) <- names  
  
electrons <- data.frame(cbind(electrons,electron.compressed.bins))  
pions <- data.frame(cbind(pions,pion.compressed.bins))  
  
save(electrons,file="~/Thesis data/SemiFullData/electrons.rdata")  
save(pions,file="~/Thesis data/SemiFullData/pions.rdata")  
  
electrons$id <- "electron"  
pions$id <- "pion"  
  
electrons$id <- as.factor(electrons$id)  
pions$id <- as.factor(pions$id)

dat <- data.frame(rbind(electrons,pions))  
save(dat,file="~/Thesis data/SemiFullData/unscaled\_data.rdata")

Scale data

dat[,2:33] <- scale(dat[,2:33])  
save(dat,file="~/Thesis data/SemiFullData/scaled\_data.rdata")  
  
load("~/Thesis data/SemiFullData/scaled\_data.rdata")  
i <- dat$index  
rm(dat)

Remove missing values

load("~/Thesis data/SemiFullData/scaled\_data.rdata")  
dat <- na.omit(dat)

rownames(dat) <- NULL  
  
index <- dat$index  
  
dat <- dat[,-1]  
  
dat$id <- as.character(dat$id)  
  
dat$electron <- ifelse(dat$id=="electron",1,0)  
  
id <- dat$id  
  
dat <- dat[,-1]

Get a equal sample of electrons and pions (50 000 each):

id.index <- data.frame(cbind(id,index))  
  
require(dplyr)  
  
electron.id <- id.index[id.index$id=="electron",]  
  
pion.id <- id.index[id.index$id=="pion",]  
  
electron.train <- base::sample(electron.id$index,50000,replace = F)  
pion.train <- base::sample(pion.id$index, 50000,replace=F)  
  
electron.train <- as.numeric(as.character(electron.train))  
pion.train <- as.numeric(as.character(pion.train))

Since particles pass through pads at different angles, and at different coordinates within a pad, the exact column/ row in the matrix representation of a detector trace is not important.

To this end, the actual pad data (“images”) will not be included in the feature set used for vanilla feed-forward neural networks, to get an early estimation of particle ID based on summary statistics.

At a later stage, kernels/ masks will be employed in a convolutional neural network setup, to detect features in the detector “images” that are diagnostic of the “electron”/ “pion” ID.

Furthermore, the 24 column sums can only be informative insomuch as they are averaged across all columns, since the positional information is not relevant to the PID process:

dat <- data.frame(cbind(index,dat))  
  
c <- dat[,9:32]  
  
c <- rowMeans(c)  
  
dat <- dat[,c(1:8,33,34)]  
  
dat <- data.frame(cbind(c,dat))  
  
dat <- dat[,-10]  
  
save(dat,file="~/Thesis data/SemiFullData/final\_data.rdata")  
rm(c)  
  
train.id <- c(electron.train,pion.train)  
  
train <- dat %>%  
 subset(index %in% train.id)  
  
test <- dat %>%  
 subset(! index %in% train.id)  
  
save(train,file="~/Thesis data/SemiFullData/train1.rdata")  
save(test,file="~/Thesis data/SemiFullData/test1.rdata")

load("~/Thesis data/SemiFullData/train1.rdata")  
load("~/Thesis data/SemiFullData/test1.rdata")

Separate out index from data:

test.id.index <- test[,c(2,10)]  
train.id.index <- train[,c(2,10)]  
  
test <- test[,-2]  
train <- train[,-2]

# Particle Identification:

## Majority Class Classifier:

The vast majority of particles detected in the TRD are pions, in fact in our dataset we have the following ratio:

PID <- c(train$electron,test$electron)  
  
pion.percentage <- 100-sum(PID/length(PID))  
electron.percentage <- sum(PID/length(PID))  
  
t <- data.frame(pion.percentage,electron.percentage)  
require(knitr)  
kable(t)

|  |  |
| --- | --- |
| pion.percentage | electron.percentage |
| 99.89238 | 0.1076221 |

It is clear, then, that our task is not smiply as classification problem, since predicting “pion” for each particle will result in upwards of 99% accuracy, without any effort.

Our task centers around pion efficiency, i.e. not misclassifying a pion as an electron, while not losing too many electrons in the process.

## Confusion matrix:

PID <- data.frame(cbind(PID,rep(0,length(PID))))  
names(PID) <- c("GroundTruth","MajorityClassClassifierPred")  
  
PID$GroundTruth <- as.factor(PID$GroundTruth)  
PID$MajorityClassClassifierPred <- as.factor(PID$MajorityClassClassifierPred)  
  
levels(PID$MajorityClassClassifierPred) <- c("0","1")  
  
kable(table(PID))

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 539247 | 0 |
| 1 | 65034 | 0 |

negs <- c("True Negative","False Negative")  
poss <- c("False Positives","True Positives")  
  
conf.m <- data.frame(cbind(negs,poss))  
names(conf.m) <- c("Predicted PION","Predicted ELECTRON")  
rownames(conf.m) <- c("IS a PION", "IS an ELECTRON")  
  
kable(conf.m)

|  |  |  |
| --- | --- | --- |
|  | Predicted PION | Predicted ELECTRON |
| IS a PION | True Negative | False Positives |
| IS an ELECTRON | False Negative | True Positives |

A more realistic evaluation of our Majority Classifier model accuracy, would be as follows:

We reject 100% of Pions, but we keep 0% of electrons.

Our goal is to find a model that rejects as many pions as possible, but keeps as many electrons as possible as well.

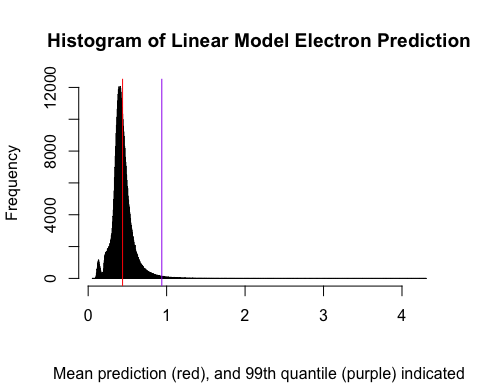
## Simple linear regression:

linmod1 <- lm(electron~.,data=train)  
  
linmod.preds <- predict(linmod1,test)

We plot the distribution of electron predictions, where electrons were labeled as integer 1 and pions as integer 0, and add reference lines for the mean prediction in red and the 99th quantile in purple.

If we go with our prior knowledge that around 99.9% of particles detected in the TRD, the 99th quantile seems like a reasonable cut-off point at this point to get a benchmark for pion rejection.

accuracy <- data.frame(cbind(as.numeric(test$electron),linmod.preds))  
  
names(accuracy) <- c("actual","Linear Model Prediction")  
  
r <- range(accuracy$`Linear Model Prediction`)  
  
q <- quantile(accuracy$`Linear Model Prediction`,0.99)  
  
hist(accuracy$`Linear Model Prediction`,breaks=1000,main="Histogram of Linear Model Electron Prediction",sub="Mean prediction (red), and 99th quantile (purple) indicated",xlab="")  
abline(v=mean(accuracy$`Linear Model Prediction`),col="red")  
abline(v=q,col="purple")



We apply this logic to get a binary response value from our basic linear model, and show the confusion matrix for this prediction:

accuracy$lm1.99quantile.pred <- ifelse(accuracy$`Linear Model Prediction`>=q,1,0)  
  
accuracy$actual <- as.factor(accuracy$actual)  
  
accuracy$lm1.99quantile.pred <- as.factor(accuracy$lm1.99quantile.pred)  
  
kable(table(accuracy$actual,accuracy$lm1.99quantile.pred))

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 484793 | 4454 |
| 1 | 14445 | 589 |

Assessing our model accuracy:

accuracy$actual <- as.numeric(as.character(accuracy$actual))  
accuracy$lm1.99quantile.pred <- as.numeric(as.character(accuracy$lm1.99quantile.pred))  
accuracy$correct <- ifelse(accuracy$actual==accuracy$lm1.99quantile.pred,1,0)  
  
a <- sum(accuracy$correct/nrow(accuracy))

## [1] "We achieve an accuracy score of 0.962522879109071"

### Pion Efficiency:

pion.efficiency <- ifelse(accuracy$actual==0&accuracy$lm1.99quantile.pred==0,1,0)  
p <- length(which(accuracy$actual==0)==T)  
pion.efficiency <- sum(pion.efficiency)/p  
  
pion.efficiency <- 100-pion.efficiency\*100

## [1] "Our majority class classifier incorrectly classified 0.10762% of pions as electrons"

## [1] "Based on predictions from our basic linear model, we incorrectly classify 0.91038% of pions in our test set as electrons"

### Electron Efficiency:

electron.efficiency <- ifelse(accuracy$actual==1&accuracy$lm1.99quantile.pred==1,1,0)  
e <- length(which(accuracy$actual==1)==T)  
electron.efficiency <- sum(electron.efficiency)/e  
  
electron.efficiency <- electron.efficiency\*100

## [1] "Our majority class classifier correctly accepted 0% of electrons in our test set"

## [1] "Based on predictions from our basic linear model, we correctly accept 3.91779% of electrons in our test set"

## Optimization of linear model prediction cut-off point for electron classification:

library(pROC)  
  
#Penalize pion error and electron error equally:  
  
lm.optimization <- function(cut.off){  
 p <- predict(linmod1,test)  
 a <- test$electron  
   
 my.prediction <- ifelse(p>=cut.off,1,0)  
   
 roc\_obj <- roc(a, my.prediction)  
 auc(roc\_obj)  
   
}  
  
optim.q <- optim(par=r[1],fn=lm.optimization,lower=r[1],upper=r[2],method="Brent",control = list(fnscale=-1))  
  
q <- optim.q$par

## [1] "After optimizing the electron prediction cut-off, by maximizing the area under the curve of the receiver operating characteristic, an AUC value of 0.709412331525969 is achieved"

Our new confusion matrix, with a more informed cut-off point, looks as follows:

accuracy$lm1.99quantile.pred <- ifelse(accuracy$`Linear Model Prediction`>=q,1,0)  
  
accuracy$actual <- as.factor(accuracy$actual)  
  
accuracy$lm1.99quantile.pred <- as.factor(accuracy$lm1.99quantile.pred)  
  
kable(table(accuracy$actual,accuracy$lm1.99quantile.pred))

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 367069 | 122178 |
| 1 | 4983 | 10051 |

### Optimized Cut-off linear model accuracy assessment:

accuracy$actual <- as.numeric(as.character(accuracy$actual))  
accuracy$lm1.99quantile.pred <- as.numeric(as.character(accuracy$lm1.99quantile.pred))  
accuracy$correct <- ifelse(accuracy$actual==accuracy$lm1.99quantile.pred,1,0)  
  
a <- sum(accuracy$correct/nrow(accuracy))

## [1] "We achieve an accuracy score of 0.74783701943956"

### Pion Efficiency:

old.pion.efficiency <- pion.efficiency  
  
pion.efficiency <- ifelse(accuracy$actual==0&accuracy$lm1.99quantile.pred==0,1,0)  
p <- length(which(accuracy$actual==0)==T)  
pion.efficiency <- sum(pion.efficiency)/p  
  
pion.efficiency <- 100-pion.efficiency\*100

## [1] "Our majority class classifier incorrectly classified 0.10762% of pions as electrons"

## [1] "Based on predictions from our basic linear model, we incorrectly classify 0.91038% of pions in our test set as electrons"

## [1] "Based on predictions from our cut-off optimized basic linear model, we incorrectly classify 24.97266% of pions in our test set as electrons"

### Electron Efficiency:

old.electron.efficiency <- electron.efficiency  
  
electron.efficiency <- ifelse(accuracy$actual==1&accuracy$lm1.99quantile.pred==1,1,0)  
e <- length(which(accuracy$actual==1)==T)  
electron.efficiency <- sum(electron.efficiency)/e  
  
electron.efficiency <- electron.efficiency\*100

## [1] "Our majority class classifier correctly accepted 0% of electrons in our test set"

## [1] "Based on predictions from our basic linear model, we correctly accept 3.91779% of electrons in our test set"

## [1] "Based on predictions from our cut-off optimized basic linear model, we correctly accept 66.85513% of electrons in our test set as electrons"

# DataRobot: Automated ML

DataRobot is a platform for automated ML, which automates data-preprocessing, feature generation, model building with cross validation, and prediction API deployment.

In order to find models which could be useful in PID, the full training and test set were uploaded onto their platform and “Autopilot” was initialized.

Figure one shows the feature importance results after data upload:

datarobot.train <- rbind(train,test)  
datarobot.train$electron <- ifelse(datarobot.train$electron==1,"electron","pion")  
write.csv(datarobot.train,"~/MSc-train-data.csv")

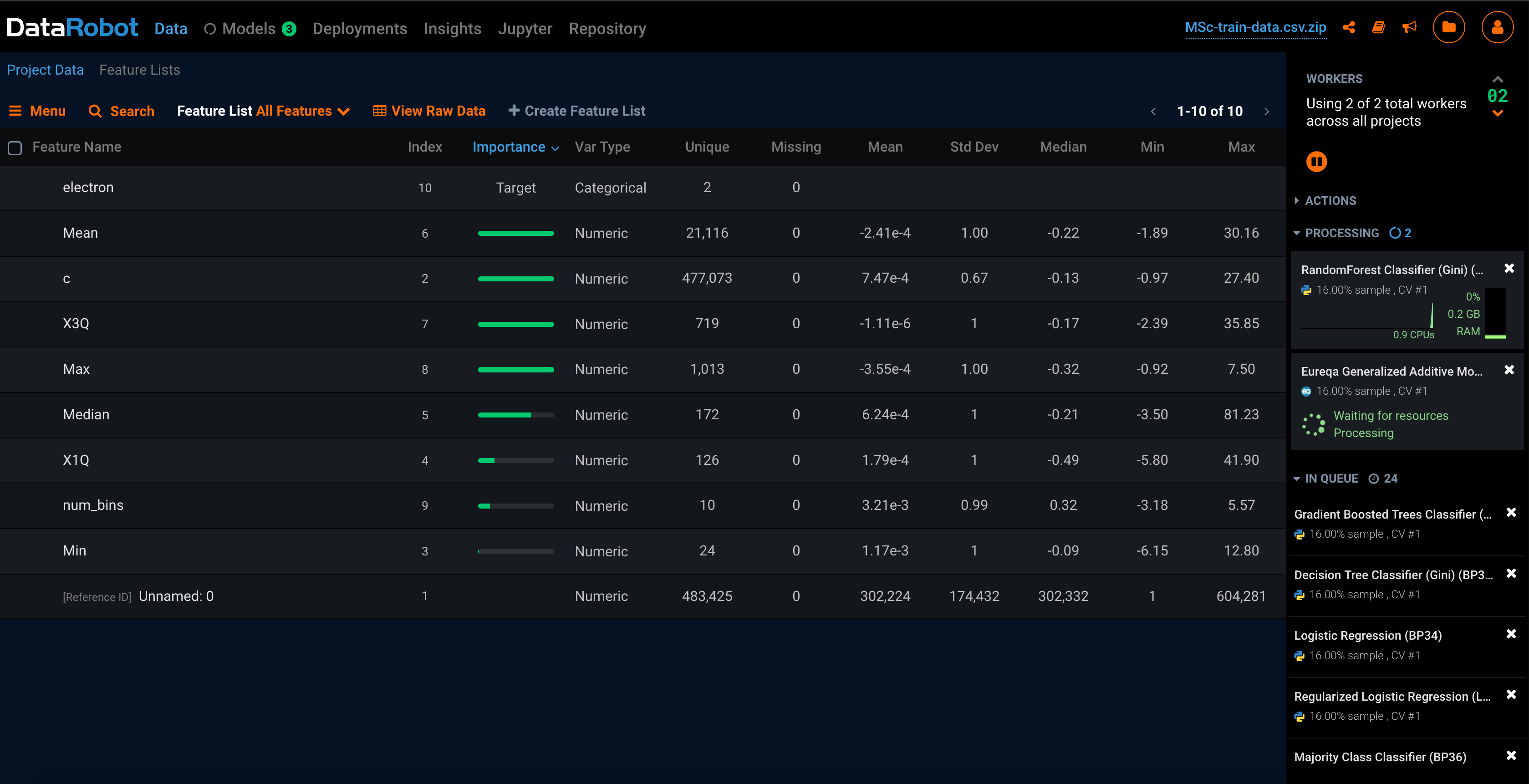


Figure1: Feature Importance

After running models overnight, DataRobot built 58 models, of which the top 5 are shown here:

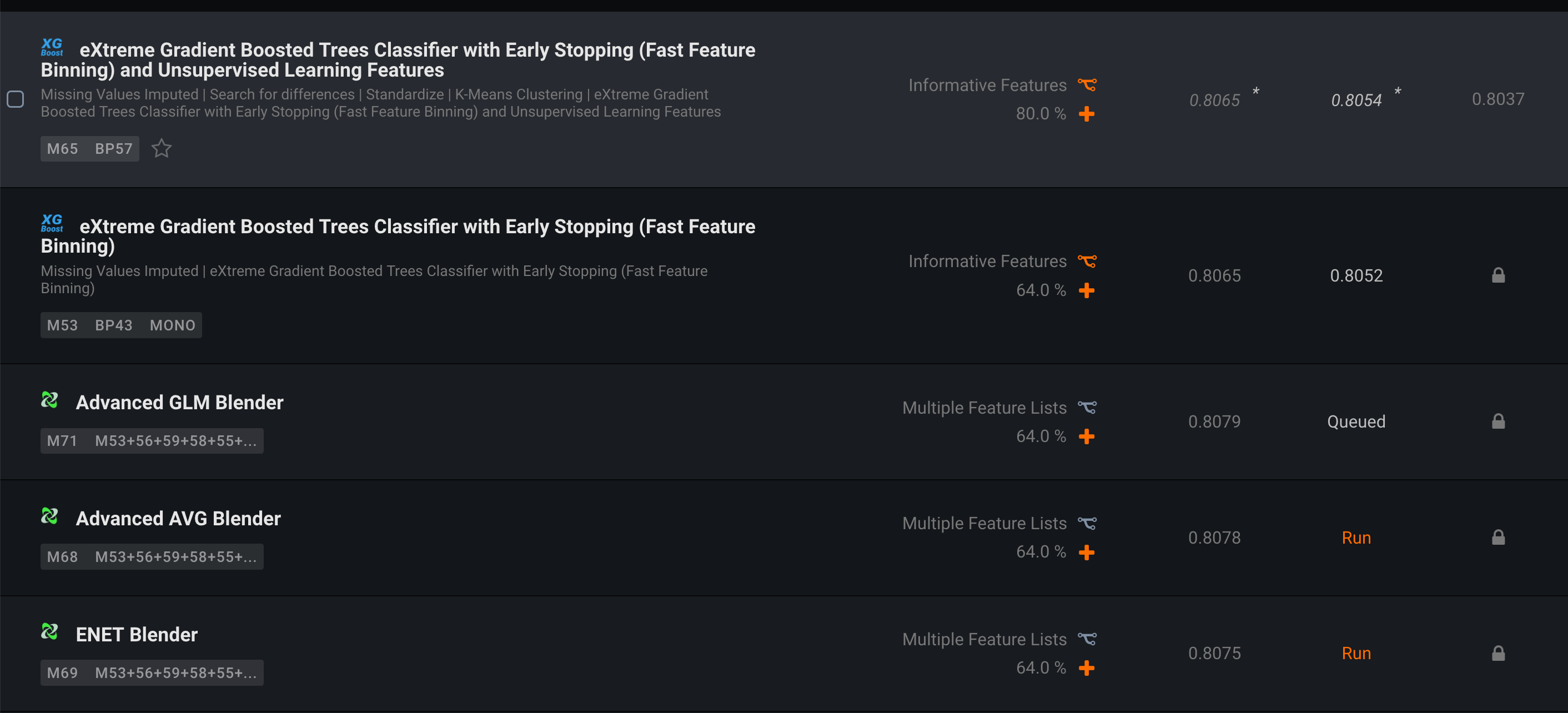


Figure 2: Top 5 DataRobot models

Evaluating the top model’s metrics, shows the following model blueprint:

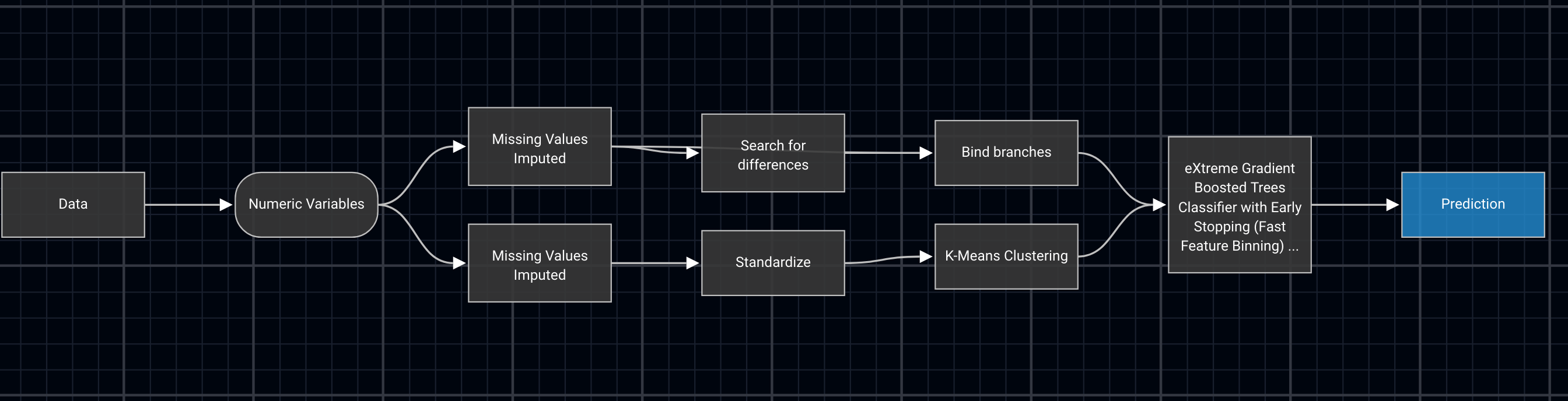


Figure 3: Extreme Gradient Boosted Tree Classifier with Early Stopping (Fast Feature Binning) Bueprint

The ROC curve and Prediction Distribtution plots are as follows:

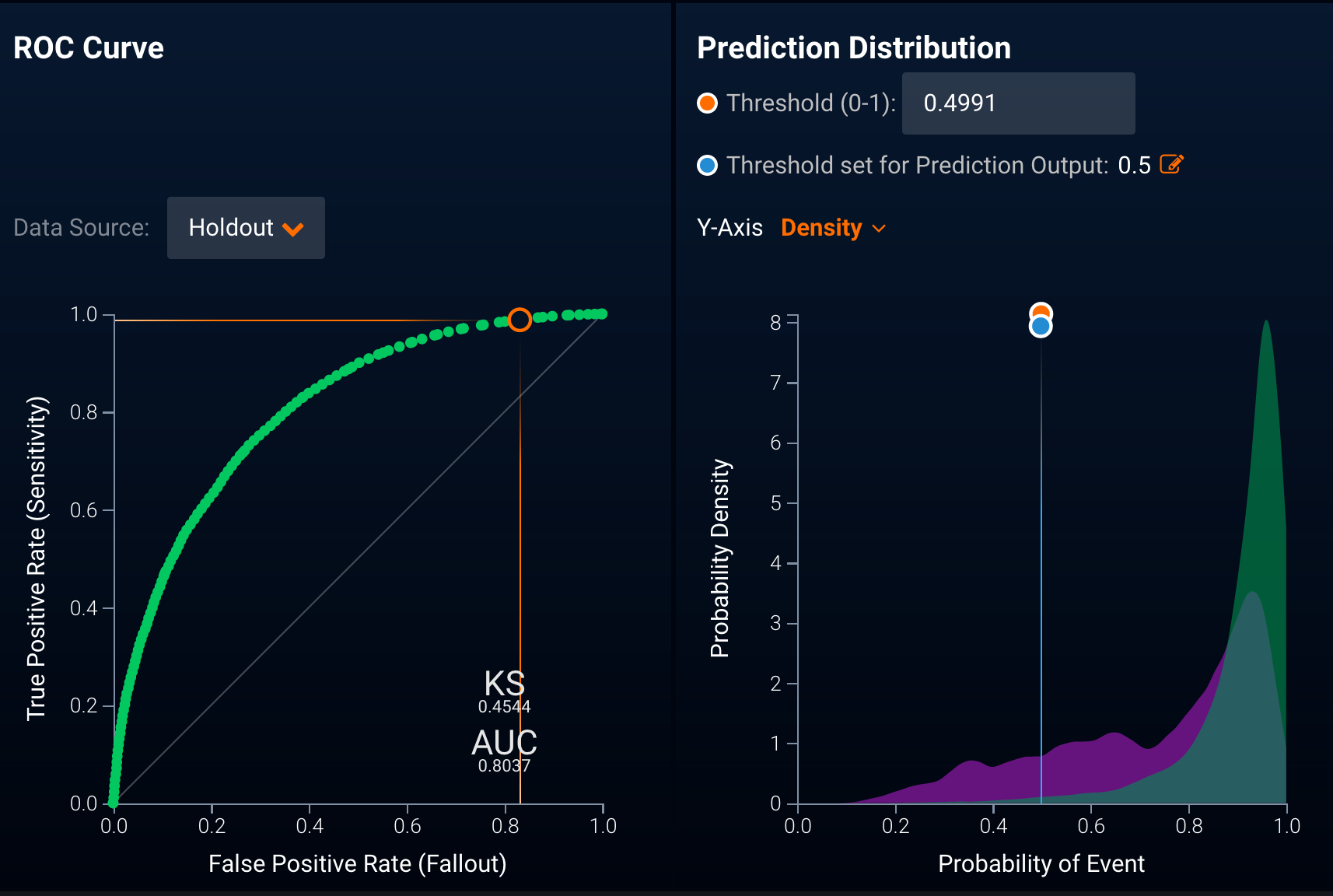


Figure 4: Top Model Evaluation

DataRobot’s Confusion Matrix is flipped relative to what has been shown so far, i.e. IS(pion) +/ IS NOT(pion) -



Figure 5: Top Model Confusion Matrix

## [1] "DataRobot detects 16.9139694010917% of electrons in the holdout set it created from the FULL dataset that was uploaded"

## [1] "DataRobot incorrectly classifies 1.14697400995837% of pions as electrons in the holdout set it created from the FULL dataset that was uploaded"

# Deep Learning

rm(list=ls())  
  
load("~/Thesis data/SemiFullData/fulljson.rdata")  
  
pads <- sapply(j,`[[`,"layer0")  
rm(j)

## Roll out pad data

require(Smisc)  
  
pad.df <- sapply(pads,as.numeric)  
rm(pads)  
  
for(i in 1:length(pad.df)){  
 if(length(pad.df[[i]])!=264){  
 pad.df[[i]] <- as.numeric(rep(0,264))  
 }  
}  
  
pad.df <- do.call(rbind,pad.df)  
  
save(pad.df,file="~/Thesis data/SemiFullData/cleanpads.rdata")

load("~/Thesis data/SemiFullData/fulljson.rdata")  
load("~/Thesis data/SemiFullData/cleanpads.rdata")  
pdg <- sapply(j,`[[`,"pdgCode")  
  
pdg <- as.numeric(pdg)  
  
pdg <- as.data.frame(pdg)  
pdg$index <- rownames(pdg)  
  
pdg$pdg <- ifelse(pdg$pdg %in% c(11,-11),T,F)  
  
names(pdg) <- c("electron","index")  
  
pdg <- pdg[,c(2,1)]  
  
cnn.dat <- data.frame(cbind(pdg,pad.df))  
  
save(cnn.dat,file="~/Thesis data/SemiFullData/cnn.rdata")

load("~/Thesis data/SemiFullData/cnn.rdata")  
  
cnn.dat <- as.matrix(cnn.dat[,-c(1,2)])  
  
zero.pads <- c()  
  
for(i in 1:nrow(cnn.dat)){  
 if(all(cnn.dat[i,]==0)){  
 zero.pads <- c(zero.pads,i)  
 }  
}  
  
cnn.dat <- cnn.dat[-zero.pads,]  
pdg <- pdg[-zero.pads,]  
  
save(cnn.dat,file="~/Thesis data/SemiFullData/cnn.rdata")  
save(pdg,file="~/Thesis data/SemiFullData/pdg.rdata")

set.seed(123456)  
pdg$index <- 1:nrow(pdg)  
  
require(dplyr)  
  
e.ind <- pdg %>%  
 subset(electron==T)  
  
p.ind <- pdg %>%  
 subset(electron==F)  
  
e.holdout <- base::sample(e.ind$index,size=5000,replace=F)  
p.holdout <- base::sample(p.ind$index,size=5000,replace=F)  
  
e.ind <- e.ind %>%  
 subset(!index %in% e.holdout)  
  
p.ind <- p.ind %>%  
 subset(!index %in% p.holdout)  
  
holdout.ind <- as.numeric(c(e.holdout,p.holdout))  
  
holdout.data <- cnn.dat[holdout.ind,]  
  
cnn.dat <- cnn.dat[-holdout.ind,]  
  
pdg.holdout <- pdg[holdout.ind,]  
  
pdg <- pdg[-holdout.ind,]  
  
pdg$index <- 1:nrow(pdg)  
pdg.holdout$index <- 1:nrow(pdg.holdout)  
  
save(cnn.dat,file="~/Thesis data/SemiFullData/cnn.rdata")  
save(holdout.data,file="~/Thesis data/SemiFullData/holdout.rdata")  
save(pdg,file="~/Thesis data/SemiFullData/pdg.rdata")  
save(pdg.holdout,file="~/Thesis data/SemiFullData/pdg.holdout.rdata")

## Unbalanced class compensation:

Resample electrons 10 times:

e.bootstrap.ind <- pdg %>%  
 subset(electron==T)  
  
e.bootstrap.ind <- rbind(e.bootstrap.ind,e.bootstrap.ind,e.bootstrap.ind,e.bootstrap.ind,e.bootstrap.ind,e.bootstrap.ind,e.bootstrap.ind,e.bootstrap.ind,e.bootstrap.ind,e.bootstrap.ind)  
  
pdg <- rbind(pdg,e.bootstrap.ind)  
pdg$index <- 1:nrow(pdg)  
  
e.bootstrap.ind <- e.bootstrap.ind$index  
  
cnn.dat <- rbind(cnn.dat,cnn.dat[e.bootstrap.ind,])  
  
save(cnn.dat,file="~/Thesis data/SemiFullData/cnn.rdata")  
save(pdg,file="~/Thesis data/SemiFullData/pdg.rdata")

install.packages("keras")  
  
library(keras)  
install\_keras()

load("~/Thesis data/SemiFullData/cnn.rdata")  
load("~/Thesis data/SemiFullData/pdg.rdata")  
load("~/Thesis data/SemiFullData/pdg.holdout.rdata")  
load("~/Thesis data/SemiFullData/holdout.rdata")  
require(keras)

x\_train <- array\_reshape(cnn.dat, c(nrow(cnn.dat), 264))  
x\_test <- array\_reshape(holdout.data, c(nrow(holdout.data), 264))  
  
y\_train <- as.matrix(ifelse(pdg$electron, 1,0))  
y\_test <- as.matrix(ifelse(pdg.holdout$electron, 1,0))  
  
scale.factor <- mean(mean(x\_train),mean(x\_test))  
  
x\_train <- x\_train/scale.factor  
x\_test <- x\_test/scale.factor

y\_test = to\_categorical(y\_test)  
y\_train = to\_categorical(y\_train)  
  
dim(x\_train) <- c(nrow(x\_train), ncol(x\_train))  
dim(x\_test) <- c(nrow(x\_test), ncol(x\_test))

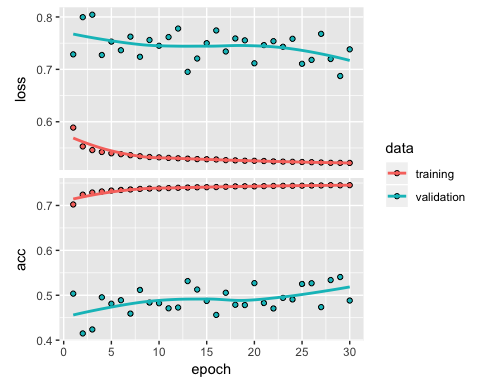
save(x\_test,file="~/Thesis data/SemiFullData/xtest.rdata")  
save(x\_train,file="~/Thesis data/SemiFullData/xtrain.rdata")  
save(y\_train,file="~/Thesis data/SemiFullData/ytrain.rdata")  
save(y\_test,file="~/Thesis data/SemiFullData/ytest.rdata")  
  
  
load("~/Thesis data/SemiFullData/ytest.rdata")  
load("~/Thesis data/SemiFullData/ytrain.rdata")

require(keras)  
rm(model)  
model <- keras\_model\_sequential()   
model %>%   
 layer\_dense(units = 128, activation = "relu", input\_shape = ncol(x\_train)) %>%   
 layer\_dropout(rate = 0.4) %>%   
 layer\_dense(units = 64, activation = "relu") %>%  
 layer\_dropout(rate = 0.3) %>%  
 layer\_dense(units = 2, activation = "softmax")  
  
save(model,file="~/Thesis data/SemiFullData/model.rdata")

## <pointer: 0x0>

model %>% compile(  
 loss = "categorical\_crossentropy",  
 optimizer = optimizer\_adam(),  
 metrics = c("accuracy")  
)

history <- model %>% fit(  
 x\_train, y\_train,   
 epochs = 30, batch\_size = 128,   
 validation\_split = 0.2  
)  
  
save(history,file="~/Thesis data/SemiFullData/history.rdata")



# Truly balance classes

rm(list=ls())  
load("~/Thesis data/SemiFullData/fulljson.rdata")  
load("~/Thesis data/SemiFullData/cleanpads.rdata")  
pdg <- sapply(j,`[[`,"pdgCode")  
  
pdg <- as.numeric(pdg)  
  
pdg <- as.data.frame(pdg)  
pdg$index <- rownames(pdg)  
  
pdg$pdg <- ifelse(pdg$pdg %in% c(11,-11),T,F)  
  
names(pdg) <- c("electron","index")  
  
pdg <- pdg[,c(2,1)]  
  
cnn.dat <- data.frame(cbind(pdg,pad.df))  
  
save(cnn.dat,file="~/Thesis data/SemiFullData/cnn2.rdata")

load("~/Thesis data/SemiFullData/cnn2.rdata")  
  
cnn.dat <- as.matrix(cnn.dat[,-c(1,2)])  
  
zero.pads <- c()  
  
for(i in 1:nrow(cnn.dat)){  
 if(all(cnn.dat[i,]==0)){  
 zero.pads <- c(zero.pads,i)  
 }  
}  
  
cnn.dat <- cnn.dat[-zero.pads,]  
pdg <- pdg[-zero.pads,]  
  
pdg$index <- 1:length(pdg$index)  
  
save(cnn.dat,file="~/Thesis data/SemiFullData/cnn2.rdata")  
save(pdg,file="~/Thesis data/SemiFullData/pdg2.rdata")

load("~/Thesis data/SemiFullData/cnn2.rdata")  
load("~/Thesis data/SemiFullData/pdg2.rdata")  
set.seed(123)  
electron.ind <- pdg %>%  
 subset(electron==T)  
  
pion.ind <- pdg %>%  
 subset(electron!=T)  
  
electron.ind <- as.numeric(electron.ind$index)  
  
pion.ind <- as.numeric(pion.ind$index)  
  
final.test <- c(base::sample(electron.ind,100,replace = F),base::sample(pion.ind,100,replace=F))  
  
electron.ind <- electron.ind %>%  
 subset(!electron.ind %in% final.test)  
  
pion.ind <- pion.ind %>%  
 subset(!pion.ind %in% final.test)  
  
N <- length(pion.ind)  
  
electron.ind <- sample(electron.ind,N, replace = T)  
  
train.id <- c(electron.ind,pion.ind)  
train.id <- sample(train.id)  
test.id <- final.test  
  
train <- cnn.dat[train.id,]  
test <- cnn.dat[test.id,]  
  
save(train,file="~/Thesis data/SemiFullData/train.rdata")  
save(test,file="~/Thesis data/SemiFullData/test.rdata")  
  
pdg.train <- pdg[train.id,]  
pdg.test <- pdg[test.id,]  
  
save(pdg.train,file="~/Thesis data/SemiFullData/y\_train.rdata")  
save(pdg.test,file="~/Thesis data/SemiFullData/y\_test.rdata")

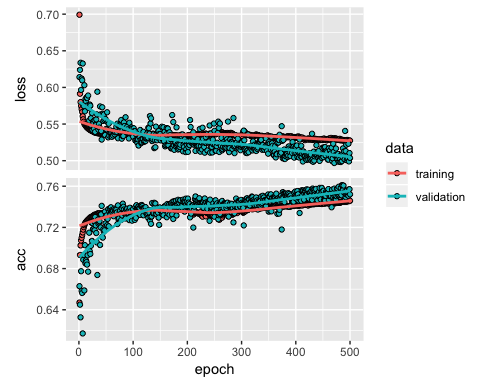
rm(list=ls())  
load("~/Thesis data/SemiFullData/train.rdata")  
load("~/Thesis data/SemiFullData/test.rdata")  
load("~/Thesis data/SemiFullData/y\_train.rdata")  
load("~/Thesis data/SemiFullData/y\_test.rdata")

require(keras)  
rm(model2)  
model2 <- keras\_model\_sequential()   
model2 %>%   
 layer\_dense(units = 256, activation = "relu", input\_shape = ncol(train)) %>%   
 layer\_dropout(rate = 0.4) %>%   
 layer\_dense(units = 128, activation = "relu") %>%  
 layer\_dropout(rate = 0.3) %>%  
 layer\_dense(units = 64, activation = "relu") %>%  
 layer\_dropout(rate = 0.3) %>%  
 layer\_dense(units = 32, activation = "relu") %>%  
 layer\_dropout(rate = 0.2) %>%  
 layer\_dense(units = 16, activation = "relu") %>%  
 layer\_dropout(rate = 0.2) %>%  
 layer\_dense(units = 2, activation = "softmax")  
  
save(model2,file="~/Thesis data/SemiFullData/model2.rdata")

model2 %>% compile(  
 loss = "categorical\_crossentropy",  
 optimizer = optimizer\_rmsprop(),  
 metrics = c("accuracy")  
)

y\_test <- as.matrix(ifelse(pdg.test$electron,1,0))  
y\_test <- to\_categorical(y\_test)  
  
y\_train <- as.matrix(ifelse(pdg.train$electron,1,0))  
y\_train <- to\_categorical(y\_train)

history2 <- model2 %>% fit(  
 train, y\_train,   
 epochs = 500, batch\_size = 1000,   
 validation\_split = 0.2  
)  
  
save(history2,file="~/Thesis data/SemiFullData/history2.rdata")  
save(model2,file="~/Thesis data/SemiFullData/model2.rdata")



load("~/Thesis data/SemiFullData/ytest.rdata")  
p <- model2 %>% predict\_proba(test)  
  
p <- cbind(as.matrix(y\_test),p)  
  
p <- p[,c(2,4)]  
  
p <-data.frame(p)  
  
names(p) <- c("actual","model2.pred")  
  
p$error <- p$actual-p$model2.pred  
  
p$squared.error <- p$error^2  
  
mean(p$squared.error)  
  
p$p0.5 <- ifelse(p$model2.pred>=0.5,1,0)  
p$p0.5 <- as.character(p$p0.5)  
p$actual <- as.character(p$actual)  
  
t1 <- table(p$actual,p$p0.5)  
  
save(t1,file="~/Thesis data/SemiFullData/t1.rdata")

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 84 | 16 |
| 1 | 38 | 62 |

p2 <- model2 %>% predict\_proba(train)  
  
p2 <- cbind(as.matrix(y\_train),p2)  
  
p2 <- p2[,c(2,4)]  
  
p2 <-data.frame(p2)  
  
names(p2) <- c("actual","model2.pred")  
  
p2$error <- p2$actual-p2$model2.pred  
  
p2$squared.error <- p2$error^2  
  
mean(p2$squared.error)  
  
p2$p0.5 <- ifelse(p2$model2.pred>=0.5,1,0)  
p2$p0.5 <- as.character(p2$p0.5)  
p2$actual <- as.character(p2$actual)  
  
t2 <- table(p2$actual,p2$p0.5)  
save(t2,file="~/Thesis data/SemiFullData/t2.rdata")

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 514878 | 77988 |
| 1 | 205381 | 387485 |

## [1] "Our second keras deep learning model correctly detects 65.3579392307874% of electrons in our training set (bootstrapped sample size n = 1 185 732)"

## [1] "Our second keras deep learning model incorrectly classifies 13.1544058859844% of pions in our training set as electrons (bootstrapped sample size n = 1 185 732)"

## [1] "Our second keras deep learning model correctly detects 62% of electrons in our holdout set (sample size n = 100"

## [1] "Our second keras deep learning model incorrectly classifies 16% of pions in our holdout set as electrons (sample size n = 100"

//TODO:

Optimize prediction cut-off point.

# Even Deeper Learning

require(keras)  
#rm(model3)  
model3 <- keras\_model\_sequential()   
model3 %>%   
 layer\_dense(units = 256, activation = "relu", input\_shape = ncol(train)) %>%   
 layer\_dropout(rate = 0.4) %>%   
 layer\_dense(units = 256, activation = "relu", input\_shape = ncol(train)) %>%   
 layer\_dropout(rate = 0.4) %>%   
 layer\_dense(units = 256, activation = "relu", input\_shape = ncol(train)) %>%   
 layer\_dropout(rate = 0.4) %>%   
 layer\_dense(units = 256, activation = "relu", input\_shape = ncol(train)) %>%   
 layer\_dropout(rate = 0.4) %>%   
 layer\_dense(units = 128, activation = "relu") %>%  
 layer\_dropout(rate = 0.3) %>%  
 layer\_dense(units = 64, activation = "relu") %>%  
 layer\_dropout(rate = 0.3) %>%  
 layer\_dense(units = 32, activation = "relu") %>%  
 layer\_dropout(rate = 0.2) %>%  
 layer\_dense(units = 16, activation = "relu") %>%  
 layer\_dropout(rate = 0.2) %>%  
 layer\_dense(units = 2, activation = "softmax")  
  
save(model3,file="~/Thesis data/SemiFullData/model3.rdata")

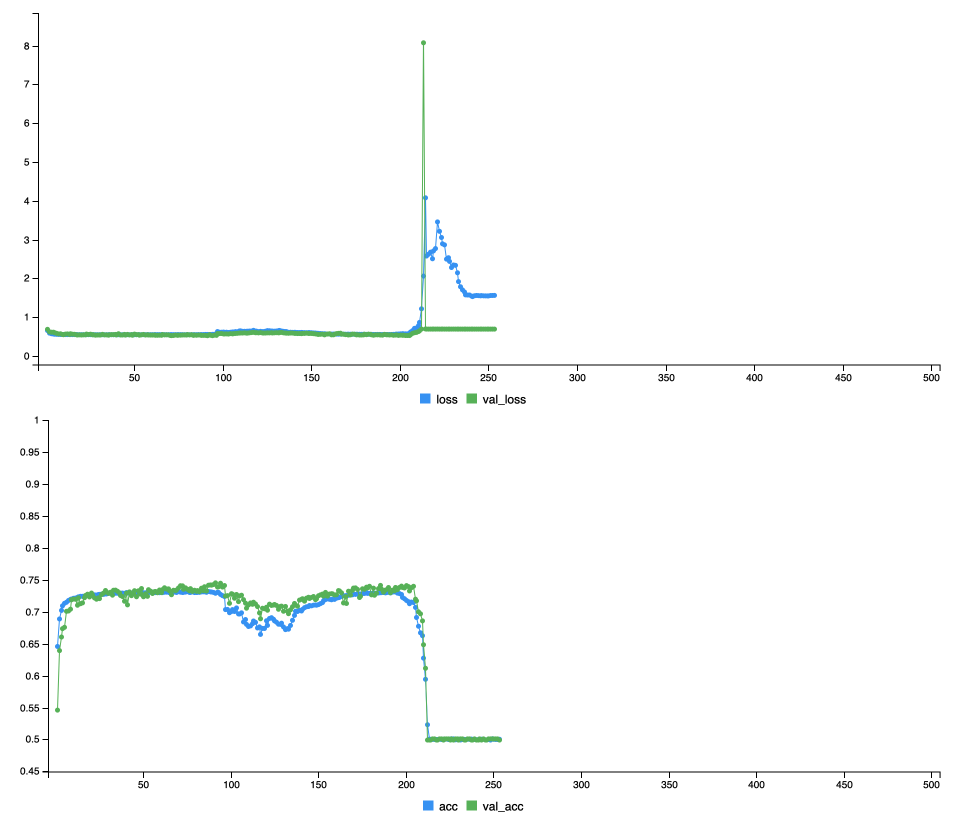
model3 %>% compile(  
 loss = "categorical\_crossentropy",  
 optimizer = optimizer\_rmsprop(),  
 metrics = c("accuracy")  
)

history3 <- model3 %>% fit(  
 train, y\_train,   
 epochs = 500, batch\_size = 1000,   
 validation\_split = 0.2  
)  
  
save(history3,file="~/Thesis data/SemiFullData/history3.rdata")  
save(model3,file="~/Thesis data/SemiFullData/model3.rdata")

## Model 3 Analysis

### Possible Causes:

* Learning rate too high
* Overfitting (too many epochs, oversampling of electrons)
* Log operation in categorical cross-entropy causing unexpected jumps in loss function, which in turn causes backpropagation to adjust weights by too much
* Highly likely that neural network started outputting a single class prediction for the last 40-50 epochs, resulting in accuracy of around 0.5 throughout these training rounds.



# Undersampling of pions, instead of oversampling of electrons

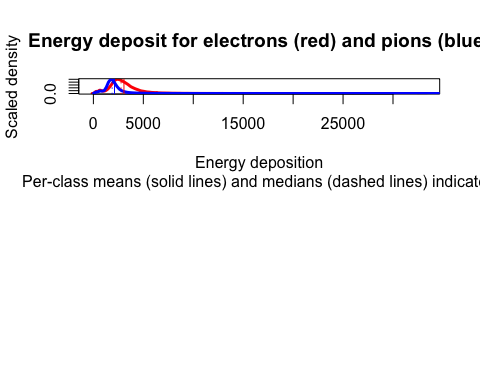
## Using the second neural network, draw a sample of pions of the same n as the sample of electrons, prefering those pions are most similar to electrons based on nusupervised learning methods

rm(list=ls())  
load("~/Thesis data/SemiFullData/fulljson.rdata")  
load("~/Thesis data/SemiFullData/cleanpads.rdata")  
pdg <- sapply(j,`[[`,"pdgCode")  
  
pdg <- as.numeric(pdg)  
  
pdg <- as.data.frame(pdg)  
pdg$index <- rownames(pdg)  
  
pdg$pdg <- ifelse(pdg$pdg %in% c(11,-11),1,0)  
  
names(pdg) <- c("electron","index")  
  
pdg <- pdg[,c(2,1)]  
  
cnn.dat <- data.frame(cbind(pdg,pad.df))

cnn.dat <- as.matrix(cnn.dat[,-c(1,2)])  
  
zero.pads <- c()  
  
for(i in 1:nrow(cnn.dat)){  
 if(all(cnn.dat[i,]==0)){  
 zero.pads <- c(zero.pads,i)  
 }  
}  
  
cnn.dat <- cnn.dat[-zero.pads,]  
pdg <- pdg[-zero.pads,]  
  
pdg$index <- 1:nrow(pdg)  
rm(zero.pads)

set.seed(1234)  
require(dplyr)  
  
elec\_ind <- pdg %>%  
 subset(electron==1) %>%  
 select(index) %>%  
 unlist() %>%  
 as.numeric()  
  
pion\_ind <- pdg %>%  
 subset(electron!=1) %>%  
 select(index) %>%  
 unlist() %>%  
 as.numeric()  
  
e.h <- sample(elec\_ind,500,replace=F)  
p.h <- sample(pion\_ind,500,replace=F)  
  
test.ind <- c(e.h,p.h)  
  
train.ind <- pdg$index[-test.ind] %>%  
 unlist() %>%  
 as.numeric()  
  
test.y <- pdg[test.ind,]  
train.y <- pdg[train.ind,]  
  
train\_electrons <- train.y %>%  
 subset(electron==1) %>%  
 select(index) %>%  
 unlist() %>%  
 as.numeric()  
  
train\_pions <- train.y %>%  
 subset(electron==0) %>%  
 select(index) %>%  
 unlist() %>%  
 as.numeric()  
  
train\_electrons <- cnn.dat[train\_electrons,]  
train\_pions <- cnn.dat[train\_pions,]  
  
save(train\_electrons,file="~/Thesis data/SemiFullData/train\_electrons.rdata")  
save(train\_pions,file="~/Thesis data/SemiFullData/train\_pions.rdata")  
save(test.y,file="~/Thesis data/SemiFullData/test.y.rdata")  
save(train.y,file="~/Thesis data/SemiFullData/train.y.rdata")

load("~/Thesis data/SemiFullData/train\_electrons.rdata")  
load("~/Thesis data/SemiFullData/train\_pions.rdata")  
e.rs <- as.numeric(rowSums(train\_electrons))  
p.rs <- as.numeric(rowSums(train\_pions))  
  
d.e.rs <- density(e.rs)  
d.p.rs <- density(p.rs)  
  
par(mfcol=c(2,1))  
  
sc.f <- length(p.rs)/length(e.rs)  
  
plot(x=d.e.rs$x,y=d.e.rs$y/max(d.e.rs$y),col="red",type="l",lwd=3,main="Energy deposit for electrons (red) and pions (blue)",xlab="Energy deposition",ylab="Scaled density",sub="Per-class means (solid lines) and medians (dashed lines) indicated")  
lines(x=d.p.rs$x,y=d.p.rs$y/max(d.p.rs$y),col="blue",lwd=3)  
abline(v=median(e.rs),col="red",lty="dotted")  
abline(v=median(p.rs),col="blue",lty="dotted")  
abline(v=mean(e.rs),col="red")  
abline(v=mean(p.rs),col="blue")



elec.feat <- data.frame(e.rs)  
pion.feat <- data.frame(p.rs)  
  
rm(e.rs)  
rm(p.rs)  
  
names(elec.feat) <- "energy\_deposit"  
names(pion.feat) <- "energy\_deposit"  
  
range.lo <- 0  
range.hi <- max(range(train\_electrons)[2],range(train\_pions)[2])+100  
  
breaks <- seq(range.lo,range.hi,50)  
  
e.bins <- apply(train\_electrons,1,function(x) hist(x,breaks = breaks, plot=F)$count)  
  
p.bins <- apply(train\_pions,1,function(x) hist(x,breaks = breaks, plot=F)$count)  
  
rm(train\_electrons)  
rm(train\_pions)  
  
e.bins <- t(e.bins)  
p.bins <- t(p.bins)  
  
elec.feat <- data.frame(cbind(elec.feat,e.bins))  
pion.feat <- data.frame(cbind(pion.feat,p.bins))  
  
save(elec.feat,file="~/Thesis data/SemiFullData/elec.feat.rdata")  
save(pion.feat,file="~/Thesis data/SemiFullData/pion.feat.rdata")

# Fuzzy Clustering

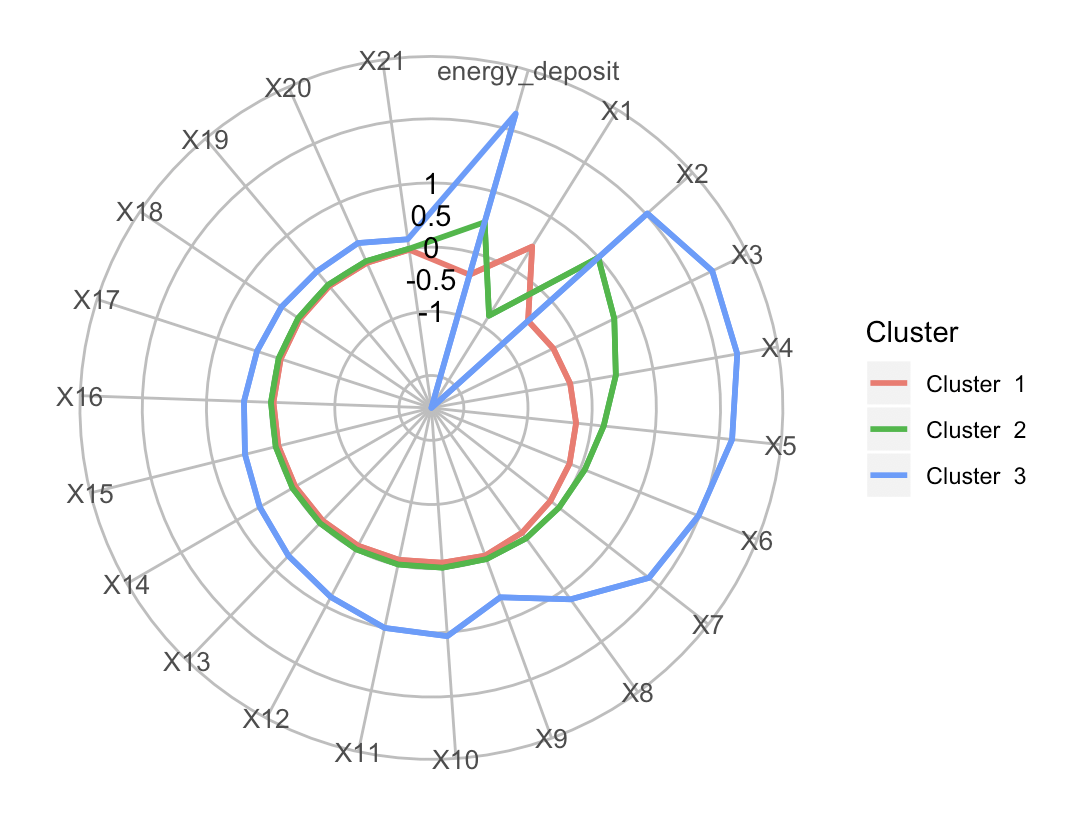
## Fuzzy C-means

Mathematical distance computed with euclidian norm:

load("~/Thesis data/SemiFullData/elec.feat.rdata")  
load("~/Thesis data/SemiFullData/pion.feat.rdata")  
cl.d <- rbind(elec.feat,pion.feat)  
  
all(cl.d[,23]==0)  
  
cl.d <- cl.d[,-23]  
  
cl.d <- scale(cl.d)  
  
rm(elec.feat)  
rm(pion.feat)

require(advclust)  
fuzclust <- fuzzy.CM(X=cl.d,K = 3,m = 2,RandomNumber = 1234)  
  
save(fuzclust,file="~/Thesis data/SemiFullData/fuzclust.rdata")  
  
load("~/Thesis data/SemiFullData/train.y.rdata")  
  
unsupervised.features <- data.frame(cbind(train.y,fuzclust@member))  
  
save(unsupervised.features,file="~/Thesis data/SemiFullData/unsup\_feat1.rdata")

radar.plotting(fuzclust, cl.d)

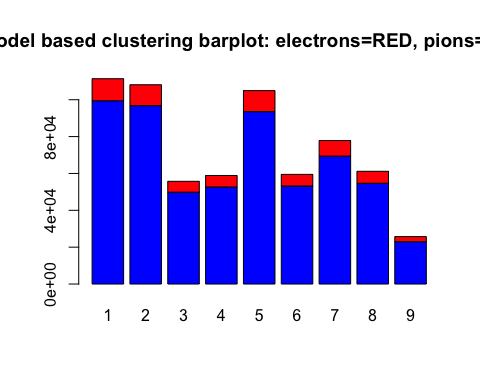


Radarplot for c means clustering

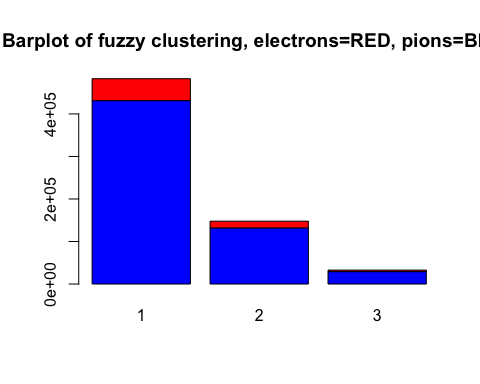
## Model based clustering

require(mclust)  
  
clustered.dat <- Mclust(cl.d)  
  
save(clustered.dat,file="~/Thesis data/SemiFullData/mclust.rdata")  
  
unsupervised.features <- data.frame(cbind(unsupervised.features,clustered.dat$z))  
  
save(unsupervised.features,file="~/Thesis data/SemiFullData/unsup\_feat1.rdata")

load("~/Thesis data/SemiFullData/fuzclust.rdata")  
load("~/Thesis data/SemiFullData/mclust.rdata")  
load("~/Thesis data/SemiFullData/unsup\_feat1.rdata")  
barplot(table(as.character(unsupervised.features$electron),as.character(clustered.dat$classification)),col=c("blue","red"),  
 main="Model based clustering barplot: electrons=RED, pions=BLUE")



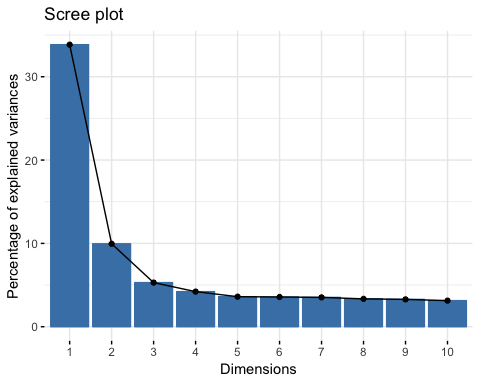
barplot(table(as.character(unsupervised.features$electron), as.character(fuzclust@hard.label)),main="Barplot of fuzzy clustering, electrons=RED, pions=BLUE",col=c("blue","red"))



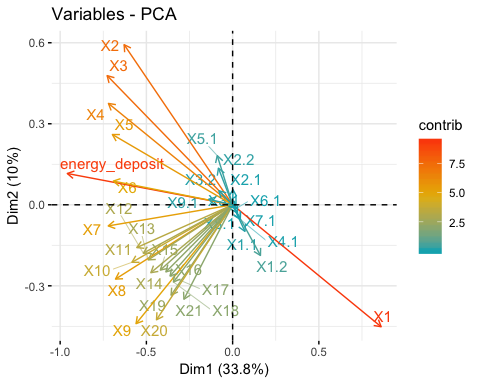
cl.d.2 <- data.frame(cbind(cl.d,unsupervised.features))  
  
  
  
save(cl.d.2,file="~/Thesis data/SemiFullData/cld2.rdata")

load("~/Thesis data/SemiFullData/cld2.rdata")  
  
cl.d.2 <- cl.d.2[,-c(23:24)]  
  
pc <- prcomp(cl.d.2)

require(factoextra)  
fviz\_eig(pc)

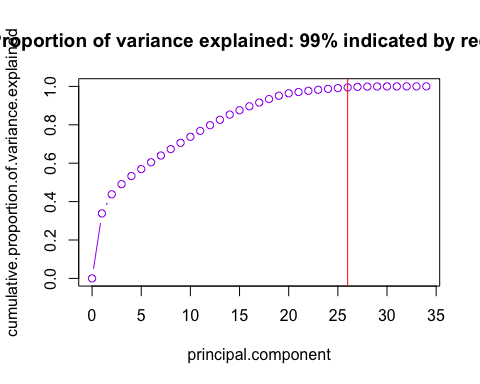


fviz\_pca\_var(pc,  
 col.var = "contrib", # Color by contributions to the PC  
 gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),  
 repel = TRUE # Avoid text overlapping  
 )



find.uncert.pions.x <- pc$x

principal.component <- 0:34  
cumulative.proportion.of.variance.explained <- c(0,cumsum(pc$sdev^2)/sum(pc$sdev^2))  
  
plot(principal.component,cumulative.proportion.of.variance.explained,type="b",col="purple",main="Proportion of variance explained: 99% indicated by red line")  
abline(v=min(which(cumulative.proportion.of.variance.explained>=.99)),col="red")

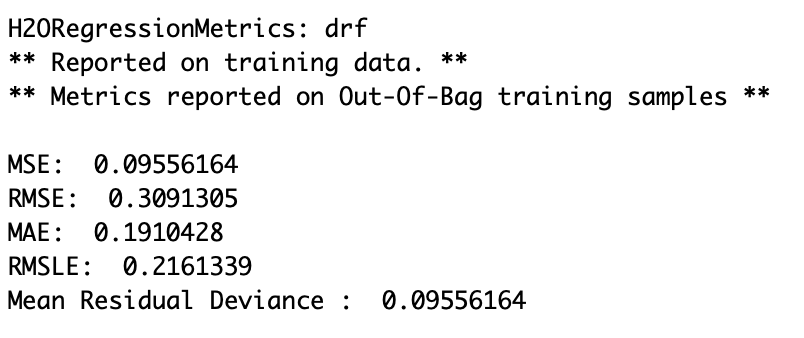


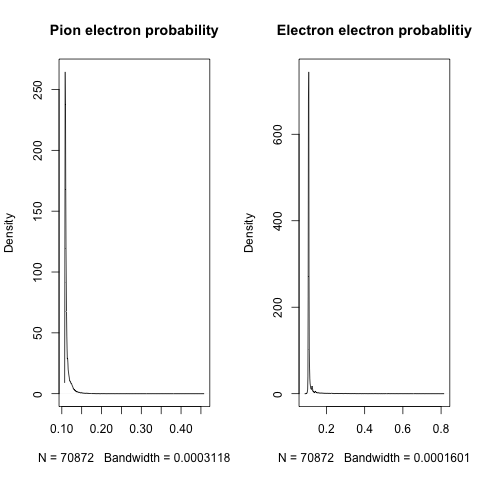
# H2O

require(h2o)  
h2o.init()

load("~/Thesis data/SemiFullData/train.y.rdata")  
  
dat <- data.frame(cbind(find.uncert.pions.x,train.y$electron))  
names(dat)[35] <- "electron"  
  
dat.hex <- as.h2o(dat)  
  
find.uncert.pion <- h2o.randomForest(x=1:34,y=35,dat.hex,nfolds=10)

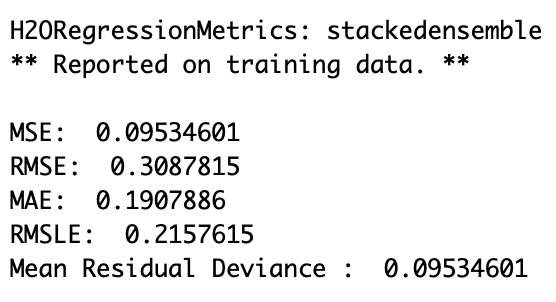
h2o.performance(find.uncert.pion)  
  
p.rf <- h2o.predict(find.uncert.pion,dat.hex)  
  
p.rf <- as.data.frame(p.rf)  
  
p.rf <- data.frame(cbind(train.y,p.rf))  
  
row.names(p.rf) <- NULL  
  
p.rf$index <- 1:nrow(p.rf)  
  
electron.sample.ind <- p.rf$index[train.y$electron==1]  
  
pion.sample.ind <- p.rf[train.y$electron==0,]  
  
pion.sample.ind <- pion.sample.ind[order(pion.sample.ind$predict,decreasing = T),]  
  
N <- length(electron.sample.ind)  
  
pion.sample.ind <- pion.sample.ind$index[1:N]  
  
any(is.na(pion.sample.ind))  
  
  
  
png(filename="~/Thesis data/SemiFullData/probability.dens.png")  
par(mfrow=c(1,2))  
  
d <- p.rf[pion.sample.ind,]  
  
d <- as.numeric(d$predict)  
  
max(d)  
plot(density(d),"Pion electron probability")  
  
d <- p.rf[electron.sample.ind,]  
  
d <- as.numeric(d$predict)  
  
max(d)  
plot(density(d),"Electron electron probablitiy")  
  
dev.off()  
  
save(pion.sample.ind,file="~/Thesis data/SemiFullData/pion.sample.ind.rdata")  
save(electron.sample.ind,file="~/Thesis data/SemiFullData/electron.sample.ind.rdata")

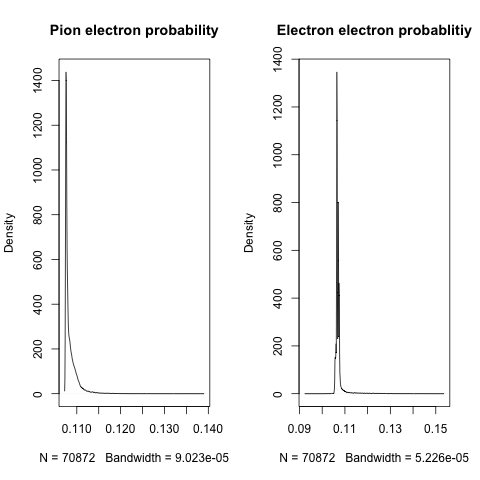




find.uncert.pion.2 <- h2o.automl(x=1:34,y=35,dat.hex,nfolds=10)

h2o.performance(find.uncert.pion.2@leader)  
  
p.rf <- h2o.predict(find.uncert.pion.2@leader,dat.hex)  
  
p.rf <- as.data.frame(p.rf)  
  
p.rf <- data.frame(cbind(train.y,p.rf))  
  
row.names(p.rf) <- NULL  
  
p.rf$index <- 1:nrow(p.rf)  
  
electron.sample.ind <- p.rf$index[train.y$electron==1]  
  
pion.sample.ind <- p.rf[train.y$electron==0,]  
  
pion.sample.ind <- pion.sample.ind[order(pion.sample.ind$predict,decreasing = T),]  
  
N <- length(electron.sample.ind)  
  
pion.sample.ind <- pion.sample.ind$index[1:N]  
  
any(is.na(pion.sample.ind))  
  
  
  
png(filename="~/Thesis data/SemiFullData/probability.dens2.png")  
par(mfrow=c(1,2))  
  
d <- p.rf[pion.sample.ind,]  
  
d <- as.numeric(d$predict)  
  
max(d)  
plot(density(d),"Pion electron probability")  
  
d <- p.rf[electron.sample.ind,]  
  
d <- as.numeric(d$predict)  
  
max(d)  
plot(density(d),"Electron electron probablitiy")  
  
dev.off()  
  
save(pion.sample.ind,file="~/Thesis data/SemiFullData/pion.sample.ind.rdata")  
save(electron.sample.ind,file="~/Thesis data/SemiFullData/electron.sample.ind.rdata")





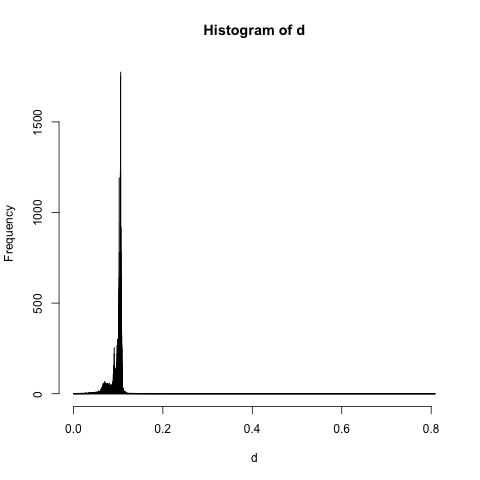
dat$electron <- as.factor(dat$electron)  
  
dat2.hex <- as.h2o(dat)  
  
find.uncert.pion.classification <- h2o.automl(x=1:34,y=35,dat2.hex,nfolds=10)

h2o.performance(find.uncert.pion.classification@leader)  
  
h2o.confusionMatrix(find.uncert.pion.classification@leader)  
  
p.rf <- h2o.predict(find.uncert.pion.classification,dat.hex)  
  
p.rf <- as.data.frame(p.rf$p1)  
  
p.rf <- data.frame(cbind(train.y,p.rf))  
  
row.names(p.rf) <- NULL  
  
p.rf$index <- 1:nrow(p.rf)  
  
electron.sample.ind <- p.rf$index[train.y$electron==1]  
  
pion.sample.ind <- p.rf[train.y$electron==0,]  
  
pion.sample.ind <- pion.sample.ind[order(pion.sample.ind$p1,decreasing = T),]  
  
N <- length(electron.sample.ind)  
  
pion.sample.ind <- pion.sample.ind$index[1:N]  
  
any(is.na(pion.sample.ind))  
  
  
  
png(filename="~/Thesis data/SemiFullData/probability.dens3.png")  
par(mfrow=c(1,2))  
  
d <- p.rf[pion.sample.ind,]  
  
d <- as.numeric(d$p1)  
  
max(d)  
plot(density(d),"Pion electron probability")  
  
d <- p.rf[electron.sample.ind,]  
  
d <- as.numeric(d$p1)  
  
max(d)  
plot(density(d),"Electron electron probablitiy")  
  
dev.off()  
  
save(pion.sample.ind,file="~/Thesis data/SemiFullData/pion.sample.ind.rdata")  
save(electron.sample.ind,file="~/Thesis data/SemiFullData/electron.sample.ind.rdata")

Training on principal components results in very similar predictions for all observations, rather train on features created for clustering algorithms above.

rm(list=ls())  
load("~/Thesis data/SemiFullData/cld2.rdata")  
load("~/Thesis data/SemiFullData/train.y.rdata")  
  
cl.d.2 <- cl.d.2[,-c(23,24)]  
  
dat <- data.frame(cbind(cl.d.2,train.y$electron))  
  
dat$train.y.electron <- ifelse(dat$train.y.electron==1,"electron","pion")  
  
dat$train.y.electron <- as.factor(dat$train.y.electron)  
  
h2o.shutdown(prompt = F)  
  
h2o.init()  
  
h2o.removeAll()  
  
dat.hex <- as.h2o(dat,"dat.hex")  
  
rf.m <- h2o.randomForest(x=1:34,y=35,training\_frame="dat.hex",nfolds=10,  
 ntrees = 100,balance\_classes = T,  
 stopping\_metric = "AUC",stopping\_tolerance = 0.001)  
  
p <- h2o.predict(rf.m,dat.hex)  
  
p <- as.data.frame(p)  
  
a <- train.y  
  
p <- data.frame(cbind(a,p))  
  
h2o.performance(rf.m)  
  
save(p,file="~/Thesis data/SemiFullData/p.rdata")  
  
p$predict <- ifelse(as.character(p$predict)=="electron",1,0)  
  
save(p,file="~/Thesis data/SemiFullData/p.rdata")

load("~/Thesis data/SemiFullData/electron.sample.ind.rdata")  
  
N <- length(electron.sample.ind)  
  
pion.sample.ind <- p[p$electron==0,]  
  
pion.sample.ind <- pion.sample.ind[order(pion.sample.ind$electron.1,decreasing = T),]  
  
pion.sample.ind <- pion.sample.ind$index[1:N]  
  
png(filename="~/Thesis data/SemiFullData/probability.dens4.png")  
par(mfrow=c(1,1))  
  
d <- p[pion.sample.ind,]  
  
d <- as.numeric(d$electron.1)  
  
hist(d,breaks=10000)  
  
dev.off()  
  
save(pion.sample.ind,file="~/Thesis data/SemiFullData/pion.sample.ind.rdata")



# Convolutional Neural Networks

rm(list=ls())  
load("~/Thesis data/SemiFullData/train\_electrons.rdata")  
load("~/Thesis data/SemiFullData/train\_pions.rdata")  
load("~/Thesis data/SemiFullData/train.y.rdata")  
  
x\_train <- rbind(train\_electrons,train\_pions)  
  
load("~/Thesis data/SemiFullData/pion.sample.ind.rdata")  
load("~/Thesis data/SemiFullData/electron.sample.ind.rdata")  
  
final.ind <- c(electron.sample.ind,pion.sample.ind)

x\_train <- cnn.dat[final.ind,]  
y\_train <- as.matrix(pdg$electron[final.ind])  
y\_train <- to\_categorical(y\_train)

require(keras)  
  
batch\_size <- 1000  
num\_classes <- 2  
epochs <- 100  
  
img\_rows <- 11  
img\_cols <- 24  
  
x\_train <- scale(x\_train)  
  
x\_train <- as.array(x\_train)  
  
x\_train <- array\_reshape(x\_train, c(nrow(y\_train), img\_rows, img\_cols,1))  
  
input\_shape <- c(img\_rows, img\_cols, 1)  
  
  
  
model <- keras\_model\_sequential() %>%  
 layer\_conv\_2d(filters = 32, kernel\_size = c(3,3), activation = 'relu',  
 input\_shape = input\_shape) %>%   
 layer\_conv\_2d(filters = 64, kernel\_size = c(3,3), activation = 'relu') %>%   
 layer\_max\_pooling\_2d(pool\_size = c(2, 2)) %>%   
 layer\_dropout(rate = 0.25) %>%   
 layer\_flatten() %>%   
 layer\_dense(units = 128, activation = 'relu') %>%   
 layer\_dropout(rate = 0.5) %>%   
 layer\_dense(units = num\_classes, activation = 'softmax')  
  
summary(model)  
  
# Compile model  
model %>% compile(  
 loss = loss\_categorical\_crossentropy,  
 optimizer = optimizer\_adadelta(),  
 metrics = c('accuracy')  
)  
  
# Train model  
model %>% fit(  
 x\_train, y\_train,  
 batch\_size = batch\_size,  
 epochs = epochs,  
 validation\_split = 0.2  
)  
  
  
  
  
scores <- model %>% evaluate(  
 x\_test, y\_test, verbose = 0  
)  
  
# Output metrics  
cat('Test loss:', scores[[1]], '\n')  
cat('Test accuracy:', scores[[2]], '\n')

1. At this stage, the ReLU model is essentially nothing more than a Multi-variable linear model: y=1x1+2x2+…+nxn+c [↑](#footnote-ref-1)