AMERICAN UNIVERSITY OF BEIRUT
FACULTY OF ARTS AND SCIENCES

Machine Learning-CMPS 261

Final Project

Higgs-Boson and Background Processes Classification

Advisor: Dr. Mohammad Kobeisy

BEIRUT, APRIL 2023

Member list

No.	Full name	Student ID
1	Tia Charaf	202102143
2	Ali Chehab	202000137
3	Mahdi Al Zein	202153123

Contents

1 Introduction

1.1 General Information

In our current description of Nature, every particle is a wave in a field. The most familiar example of this is light: light is simultaneously a wave in the electromagnetic field and a stream of particles called photons. Meaning Light is of a dual nature.

In the Higgs boson's case, the field came first. The Higgs field was proposed in 1964 as a new kind of field that fills the entire Universe and gives mass to all elementary particles. The Higgs boson is a wave in that field. Its discovery confirms the existence of the Higgs field.

The Higgs boson can't be "discovered" by finding it somewhere but has to be created in a particle collision. Once created, it transforms – or "decays" – into other particles that can be detected in particle detectors.

Physicists look for traces of these particles in data collected by the detectors. The challenge is that these particles are also produced in many other processes, plus the Higgs boson only appears in about one in a billion LHC(see figure 1) collisions. But careful statistical analysis of enormous amounts of data uncovered the particle's faint signal in 2012.[1]

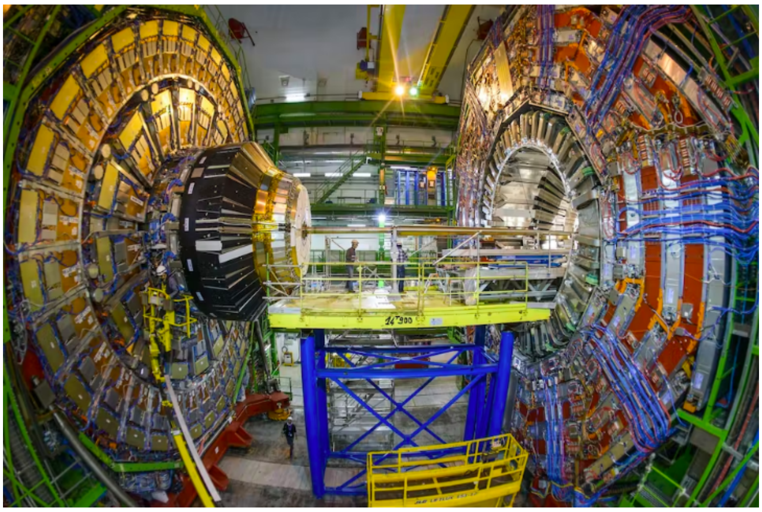


Figure 1: Large Hadron Collider

1.2 How the LHC work

As shown in Figure 1, the Large Hadron Collider is a ring which accelerates the protons to reach about the speed of light then colliding them head on. The easiest element to get a proton from is the Hydrogen which consists of only one proton orbited by an electron. Therefore, the inserted material is a Hydrogen gas.The LHC "Run 1" (2010-2013) provided enough data to test the Standard Model to new levels of precision and discover the Higgs boson. This particle was predicted in the 1960s,but it was almost 50 years before the machine powerful enough to discover it. As well as high energy, it needed lots of data: the Higgs boson is a rare thing, and fewer than one in a billion collisions at the LHC produce one. To make matters worse, the rare new particles we are looking for also tend to be very unstable, and decay too quickly to be seen directly. So the job of the experiments is to measure whatever particles do come out of a collision and try to reconstruct what happened, looking for evidence of something unusual. [2]

2 Signal Processes vs Background

Since the Higgs-Boson can't be measured or detected as it decays quickly and to distinguish between signal processes where a Higgs-Boson is present and background processes, a set of 28 features are measured to ensure that the particle is present. The table below shows the set of features and their description to better understand the data we have[5]:

Feature	Description
lepton pT	The transverse momentum of the lepton, where they can be electrons or tauons
lepton eta	The pseudorapidity eta of the lepton
lepton phi	The azimuth angle phi of the lepton
Missing energy magnitude	Energy that is not detected in a particle detector
Missing energy phi	Energy that is not detected in a particle detector
jet 1 pT	The transverse momentum of the first jet group
jet 1 eta	The pseudorapidity eta of the first jet group
jet 1 phi	The azimuth angle phi of the first jet group
jet 1 b-tag	jet consistent with b-quarks
jet 2 pT	The transverse momentum of the second jet group
jet 2 eta	The pseudorapidity eta of the second jet group
jet 2 phi	The azimuth angle phi of the second jet group
jet 2 b-tag	Jet consistent with b-quarks
jet 3 pT	The transverse momentum of the third jet group
jet 3 eta	The pseudorapidity eta of the third jet group
jet 3 phi	The azimuth angle phi of the third jet group
jet 3 b-tag	Jet consistent with b-quarks
jet 4 pT	The transverse momentum of the fourth jet group
jet 4 eta	The pseudorapidity eta of the fourth jet group
jet 4 phi	The azimuth angle phi of the fourth jet group
jet 4 b-tag	Jet consistent with b-quarks
$\mathrm{M}_j j$	The transverse momentum of the fourth jet group
$\mathrm{M}_j jj$	The pseudorapidity eta of the fourth jet group
$\mathrm{M}_l v$	The pseudorapidity eta of the fourth jet group
$\mathrm{M}_j lv$	The pseudorapidity eta of the fourth jet group
$\mathrm{M}_b b$	The pseudorapidity eta of the fourth jet group
$\mathrm{M}_w bb$	The pseudorapidity eta of the fourth jet group
$\mathrm{M}_w wbb$	The pseudorapidity eta of the fourth jet group
Event	Signal Or Background Processes

The data we have consists of 600,000 instances, devided into 80% training instances (480,000) and the other

20% testing (120,000). Several methods are used to distinguish between the signal and background processes, such as neural networks(deep learning), and gradient boosting classifier. This is explained in details with the comparison between the different models.

3 Neural Networks

3.1 Cleaning the Data

The First step after getting the data is to make sure that the data we have is cleaned. This means that the y-actual is either 0 or 1 and not any other value. After cleaning the data the next step is to make sure it's evenly distributed. The model used for an imbalanced data will spend most of the time training on background processes and not learn enough from signal processes. If the data given is not evenly distributed we can manipulate the data to have approximately equal number of instances for each class.For the Higgs-Boson we know that there is a huge set of data points for the background process and infinitesimal amount,in comparison to the other class, of signal processes. By finding the skewness of the data using scpiy,the value returned was -0.11...$\approx$ 0, so the data set given is symmetrical.The previous condition was already satisfied with the data we have. The following Histogram represents the Higgs-Boson vs Background cases (2):

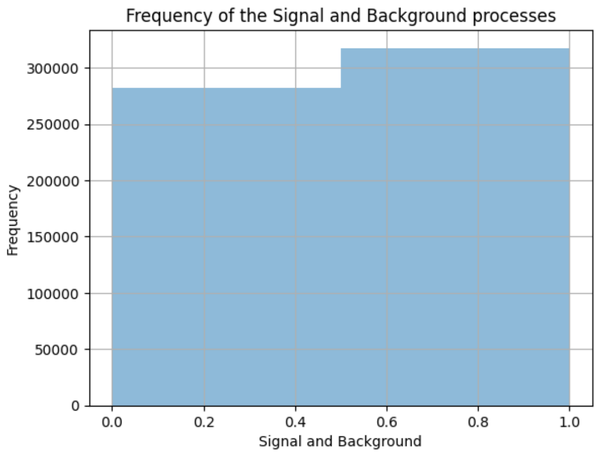


Figure 2: Frequency of the Signal and Background processes

3.2 Neural Networks Layers

For a classification problem, the neural network mainly consist of an input layer of dimensions 600,000 $\times$ 28, hidden layers , where their dimensions depend on how many layers and number of neurons , and lastly the output layer of dimensions 1×1. The Neural Network below clarifies the problem (3):

The green circles represent the scaling neurons, the blue circles are called the perceptron or hidden neurons, and the yellow neuron is the probabilistic neuron. **Note that the links of the neural network are not complete since the figure would not be clear.**

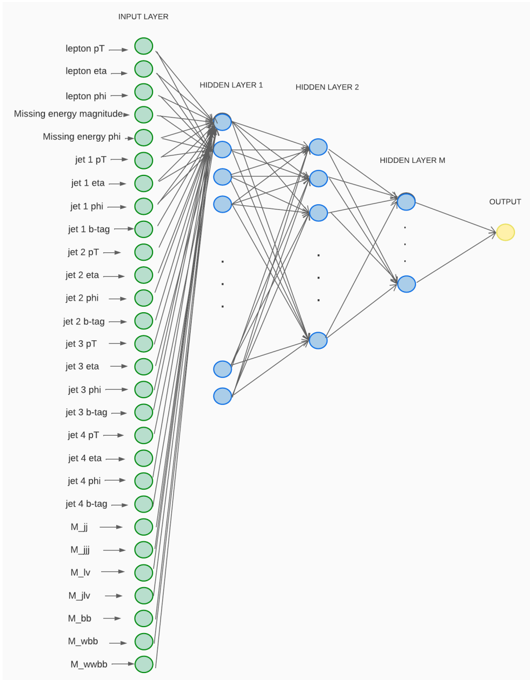


Figure 3: Example on how the neural network looks like for one example

3.3 Training Strategy

3.3.1 Model 1

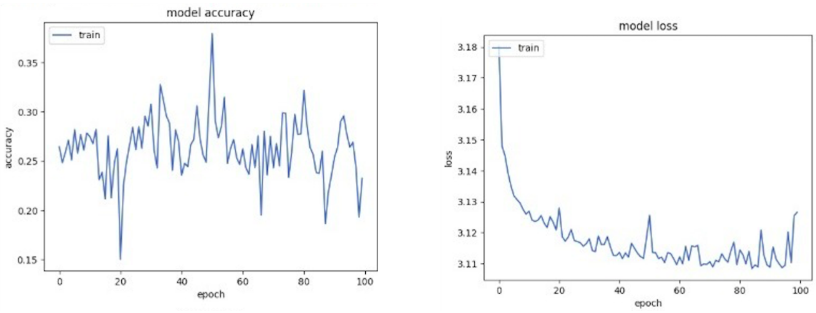
Since the number of features is small (only 28 features), we shouldn't use a huge number of layers and neurons in order to avoid over-fitting. Therefore, the first neural network model we trained our data on is made up of 3 hidden layers, the first layer consist of 15 neurons, second layer of 3 neurons and third layer of 1 neuron. The activation function used is 'relu' in the first 2 layers since it is much more efficient and time saving, however the third hidden layer consist of 'sigmoid' activation function, for a binary output either 0(if <0.5) or 1(>0.5). After running the model and doing some changes, the value of the error we got was too high so as the loss. The Output we got: loss: 0.6276 - accuracy: 0.6611, which are both unacceptable.

3.3.2 Model 2

Now, the first modification in our model was changing the number of layers and change the activation functions. A second modification is to change the number of neurons in the layers. We decided on the following neural network:

- First Hidden Layer , 40 neurons , Activation function: 'Sigmoid'
- Second Hidden Layer , 3 neurons , Activation function : 'tanh'

Results we got:



(a) Testing Accuracy (b) Testing Loss

Figure 4: Accuracy and Loss of the second model

As seen from the above figures, decreasing the number of layers lead to a loss of 3.164 and an accuracy of 0.4265 with significant fluctuations during training. This means that the model was too simple and we shall try the opposite approach. **Note that similar graphs were shown when we used 2 layer neural network with the following combinations of the activation functions: linear/sigmoid, sigmoid/tanh, and linear/tanh**

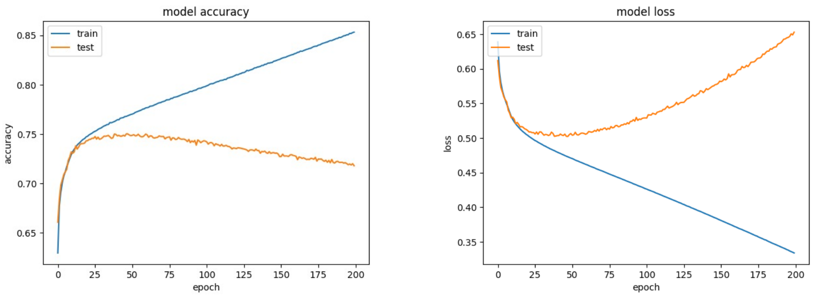
3.3.3 Model 3

To fix the under fitting problem we had earlier, and referring to another source [3] we chose a model that consists of :

- First Hidden Layer , 300 neurons , Activation Function:'tanh'
- Second Hidden Layer , 300 neurons , Activation Function:'tanh'
- Third Hidden Layer , 300 neurons , Activation Function:'tanh'
- Fourth Hidden Layer , 300 neurons , Activation Function:'tanh'
- Final Hidden Layer , 1 neuron , Activation Function:'sigmoid'

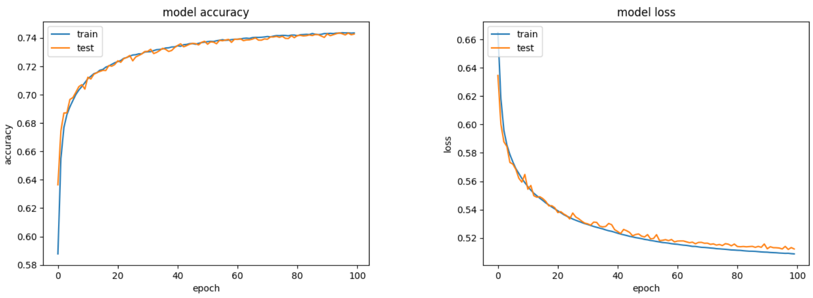
The model may be over fitting considering all the neurons in every layer, the results we got 5:

As expected the model is clearly over-fitting, the training accuracy and testing accuracy were both increasing till the val_accuracy reached its highest of 0.7501 at epoch 47. Then it started declining while the training accuracy was still increasing, meaning the model over-fits the data we have. From the loss graph also, we can see that both losses were decreasing till the testing error reached a minimum value of 0.5023 and then shifted meaning the number of misclassified points is increasing.



(a) Testing Accuracy (b) Testing Loss

Figure 5: Accuracy and Loss of the second model



(a) Testing Accuracy (b) Testing Loss

Figure 6: Accuracy and Loss of the final model

3.3.4 Final Model

After trial and error, the total number of models we had was 7, the accuracies were not increasing by much however the graphs were relatively noisy referring to the last graph we got (see figure 6).Then we finally chose to the most effective model :

- First Hidden Layer, 28 neurons , Activation function : 'relu'
- Second Hidden Layer , 28 neurons , Activation function : 'relu'
- Third Hidden Layer , 28 neurons , Activation function : 'relu'
- Fourth Hidden Layer , 28 neurons , Activation function : 'relu'
- Fifth Hidden Layer , 28 neurons , Activation function : 'relu'
- Final Hidden Layer , 1 neurons , Activation function : 'Sigmoid'

Results we got:

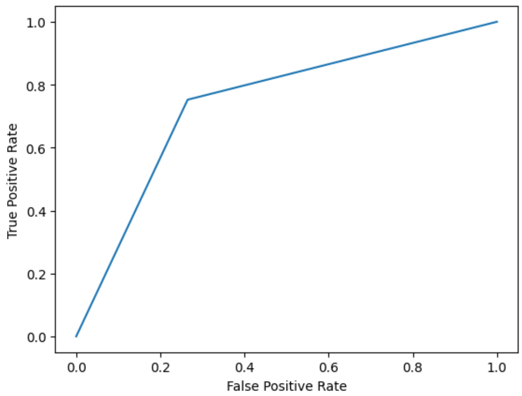


Figure 7: ROC curve

As seen in the graphs below , and when comparing to figures (5), the model accuracy has increased significantly to reach a maximum value of 0.7435 where the graphs are much smoother, meaning the models predictions are consistent when we increase the number iterations of all the training data in a cycle. To validate the model's ability to generalize after training it we can use the ROC curve , to illustrate the performance of the classifiers.The ROC curve shows the trade-off between sensitivity (or TPR) and specificity ($1 - \text{FPR}$). Classifiers that give curves closer to the top-left corner indicate a better performance. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. Our primary metric for comparison is the area under the ROC curve (AUC), with larger AUC values indicating higher classification accuracy across a range of threshold choices. In our case the AUC= 0.7437... This metric is insightful, as it is directly connected to classification accuracy, which is the quantity optimized for in training.

Moving on to the confusion matrix, the rows represent the targets (or real values) and the columns correspond to the outputs (or predictive values). The diagonal cells show the correctly classified cases, and the off-diagonal cells show the misclassified cases.

	Predicted Positive(Higgs-Boson)	Predicted Negative(Background)
Real positive (Higgs Boson)	207619	74809
Real negative (Backrgound)	78664	238903

Note that the results were based on the dataset we have but our model was also used on external dataset.[4].

3.3.5 Time Analysis

The run time of all the models varied slightly starting from 40 minutes up till an hour and 15 minutes approximately. To help speed up the process, we set jit_compile=True, which decreased the runtime of our final model to about 24 minutes .

4 Gradient Boost Classifier

4.1 Accuracy for Learning rates 0.01,0.03

The Learning rates 0.01 and 0.03 of the gradient boosting classifier models were so small so the training progressed very slowly as we are making very tiny updates, so as expected the accuracies are not optimized, with the consecutive values : 0.7119..., and 0.7246...

4.2 Accuracy for Learning rate 0.1,0.3

The Learning rates 0.1 and 0.3 of the gradient boosting classifier models were ideal as we are making significant updates without diverging, so the accuracy of the learning rate 0.3 was optimized with the following value :0.7338... , and the accuracy of the learning rate 0.1 was so close to that value 0.7322...

4.3 Accuracy for Learning rate 1,3

The Learning rates 1 and 3 of the gradient boosting classifier models were large so there was a risk of undesired diverging behaviour in the loss function, so the accuracies are not optimized, with the consecutive values : 0.7119..., and 0.5544...

4.4 Time Analysis

All Models took approximately the same amount of time to run which is roughly 10 to 14 minutes.

4.5 Comparison between Neural Networks and Gradient Boost Classifier

The two classifiers had similar results where the difference between the optimal models' accuracies is 0.0097 which is a small value. Regarding the Runtime of both models, the Gradient Boost Classifier took remarkably less time to run (about 14 minutes)compared to the Neural Network Model(about 1 hour).

5 Conclusion

The classification of the Higgs boson with machine learning algorithms is a challenging but very important task in particle physics. The discovery of the Higgs boson has changed our understanding of the basic building blocks of the universe and opened up new opportunities to explore the nature of dark matter and dark energy.

Scientists have used different machine learning techniques such as neural networks, decision trees, and gradient boosting classifiers to classify the Higgs boson from background events in experiments like the Large Hadron Collider. These techniques have shown impressive performance in accuracy, speed, and precision, which allows physicists to extract useful information from the huge amount of data generated by these experiments.

One of the main difficulties in classifying the Higgs boson is the fact that signal events are a small fraction of the total events, which makes it difficult to get enough data for accurate results. However, this was not an issue in our dataset.

To measure the performance of machine learning algorithms, scientists use metrics such as accuracy, precision, recall, and ROC curve. These metrics help to evaluate the algorithm's ability to correctly identify signal events while minimizing the number of false positives. After running a large variety of models, none were able to achieve over 76% accuracy on test data. This leads us to believe that we have yet to find a proper model or that it is difficult to detect Higgs Boson with a neural network.

Apart from the discovery of the Higgs boson, machine learning algorithms are also being used to search for new particles and phenomena that can help explain the mysteries of the universe, such as the search for dark matter particles.

In conclusion, the classification of the Higgs boson using machine learning algorithms is a great example of how computer science and physics can be combined successfully. The use of advanced machine learning techniques has enabled physicists to extract valuable information from the vast amount of data generated by particle collider experiments, leading to groundbreaking discoveries such as the Higgs boson. As machine learning continues to evolve, we can expect even more exciting discoveries and insights into the fundamental nature of the universe.

References

[1] CERN. The higgs boson.

[2] Gavin Hesketh. Explainer: how does an experiment at the large hadron collider work? *The Conversation*, 2015.

[3] D. Whiteson P. Baldi, P. Sadowsk. Searching for exotic particles in high-energy physics with deep learning. 2014.

[4] UCI Machine Learning Repository. Higgs data set.

[5] UNKOWN. Detection of the higgs boson using machine learning. *Neural Designer*, 2020.