(p)DNN for HDBS analysis

05/05/2021

Alessandra Palazzo Moved DSIDtoMass.txt		`66a962b` in 3 minutes	**20** commits
.gitignore	added .gitignore and buildDataset.py		16 days ago
Configuration.txt	Minor changes		6 days ago
DSIDtoMass.txt	Moved DSIDtoMass.txt		now
Functions.py	Removed training fraction from build(P)DNN		yesterday
README.md	Moved DSIDtoMass.txt		now
buildDNN.py	Removed training fraction from build(P)DNN		yesterday
buildDataset.py	Moved DSIDtoMass.txt		now
buildPDNN.py	Distinction between DNN and pDNN / New Functions.py / Removed eve…		7 days ago
saveToPkl.py	Minor changes		6 days ago
splitDataset.py	New common functions		6 days ago

Link github: https://github.com/enricojr/LecceDNN
Link Twiki: https://wiki.infn.it/strutture/le/hdbs/the_dnn_kit

Configuration.txt

This text file contains the configuration parameters:

- ntuplePath: path to the flat ntuples produced by the CxAODReader
- dfPath: path to the pandas data frames into which the flat ntuples are converted by the saveToPkl.py file
- modelPath: path to directory in which the results of the (p)DNN will be stored
- inputFiles: list of names of the input .root files that will be converted into pandas dataframes (without absolute path and extension)
- dataType: list of flags that specifies the nature of the corresponding input file (it can be either 'data', 'sig' or 'bkg')
- rootBranchSubSample = list of the variables (root branches) that will actually be converted into pandas dataframes
- InputFeatures(Merged/Resolved) = list of the variables that will be fed to the (p)DNN in the merged/resolved regime

Configuration.txt :

```
[config]
ntuplePath = /nfs/kloe/einstein4/HDBS/newDataFromRob/
dfPath = /nfs/kloe/einstein4/HDBS/NN_InputDataFrames/
modelPath = /nfs/kloe/einstein4/HDBS/NN_Models/
inputFiles = ['Diboson-0', 'Diboson-1', 'Signal', 'stop-0', 'stop-1', 'ttbar-0', 'ttbar-1', 'ttbar-2', 't\
tbar-3', 'ttbar-4', 'ttbar-5', 'Wjets-1', 'Wjets-2', 'Wjets-3', 'Wjets-4', 'Wjets-5', 'Wjets-6', 'Wjets-7\
', 'Wjets-8', 'Wjets-9', 'Zjets-0', 'Zjets-1', 'Zjets-2', 'Zjets-3', 'Zjets-4', 'Zjets-5', 'Zjets-6', 'Zj\
ets-7', 'Zjets-8', 'Zjets-9', 'Zjets-10', 'Zjets-11']
dataType   = ['bkg', 'bkg', 'sig', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', \
'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', 'bkg', \
'bkg', 'bkg', 'bkg', 'bkg']
rootBranchSubSample = ['DSID', 'lep1_m', 'lep1_pt', 'lep1_eta', 'lep1_phi', 'lep2_m', 'lep2_pt', 'lep2_et\
a', 'lep2_phi', 'fatjet_m', 'fatjet_pt', 'fatjet_eta', 'fatjet_phi', 'NJets', 'weight', 'Zcand_m', 'Zcand\
_pt', 'Pass_MergHP_GGF_ZZ_Tag_SR', 'Pass_MergHP_GGF_ZZ_Untag_SR', 'Pass_MergLP_GGF_ZZ_Tag_SR', 'Pass_Merg\
LP_GGF_ZZ_Untag_SR', 'Pass_MergHP_GGF_ZZ_Tag_ZCR', 'Pass_MergHP_GGF_ZZ_Untag_ZCR', 'Pass_MergLP_GGF_ZZ_Ta\
g_ZCR', 'Pass_MergLP_GGF_ZZ_Untag_ZCR', 'Pass_isVBF', 'Wdijet_m', 'Wdijet_pt', 'Wdijet_eta', 'Wdijet_phi'\
, 'Zdijet_m', 'Zdijet_pt', 'Zdijet_eta', 'Zdijet_phi', 'sigWJ1_m', 'sigWJ1_pt', 'sigWJ1_eta', 'sigWJ1_phi\
', 'sigWJ2_m', 'sigWJ2_pt', 'sigWJ2_eta', 'sigWJ2_phi', 'sigZJ1_m', 'sigZJ1_pt', 'sigZJ1_eta', 'sigZJ1_ph\
i', 'sigZJ2_m', 'sigZJ2_pt', 'sigZJ2_eta', 'sigZJ2_phi']
InputFeaturesMerged = ['lep1_m', 'lep1_pt', 'lep1_eta', 'lep1_phi', 'lep2_m', 'lep2_pt', 'lep2_eta', 'lep\
2_phi', 'fatjet_m', 'fatjet_pt', 'fatjet_eta', 'fatjet_phi', 'NJets', 'Zcand_m', 'Zcand_pt', 'DSID']
InputFeaturesResolved = ['lep1_m', 'lep1_pt', 'lep1_eta', 'lep1_phi', 'lep2_m', 'lep2_pt', 'lep2_eta', 'l\
ep2_phi', 'Zcand_m', 'Zcand_pt', 'Wdijet_m', 'Wdijet_pt', 'Wdijet_eta', 'Wdijet_phi', 'Zdijet_m', 'Zdijet\
_pt', 'Zdijet_eta', 'Zdijet_phi', 'sigWJ1_m', 'sigWJ1_pt', 'sigWJ1_eta', 'sigWJ1_phi', 'sigWJ2_m', 'sigWJ\
2_pt', 'sigWJ2_eta', 'sigWJ2_phi', 'sigZJ1_m', 'sigZJ1_pt', 'sigZJ1_eta', 'sigZJ1_phi', 'sigZJ2_m', 'sigZ\
J2_pt', 'sigZJ2_eta', 'sigZJ2_phi', 'NJets', 'DSID']
```

DSIDtoMass.txt

This text file contains the map used to convert DSID variables into mass values.

Functions.py

This file contains useful functions that will be called by the other .py scripts.

Step 1) saveToPkl.py

This script takes the input .root files and converts them into .pkl files.

Step 2) buildDataset.py

This script takes the .pkl files created in the previous step. A new signal/background flag (isSignal) is associated to each one of them (1/0). The "DSID" value of signal events is converted into the corresponding mass value, according to the map stored in DSIDtoMass.txt. For each event only relevant variables (defined in Configuration.txt) are selected and combined into one dataframe, in which the events are shuffled. 3 input flags can be specified:

- analysis (-a): type of analysis ('merged' or 'resolved')
- channel (-c): the channel considered ('ggF' or 'VBF')
- preselection cuts (-p): string that will be translated to python commad to filter the inital dataframes according to it (e.g. 'lep1_pt > 0 and lep1_eta > 0')

Only the first two flags are mandatory.

To build the 4 datasets:

- ggF : Pass_isVBF == False
- VBF : Pass_isVBF == True
- Merged : OR di tutte le variabili
  Pass_(Merg/Res)(HP/LP)_(GGF/VBF)_(ZZ/WZ)_(Tag/Untag)_(SR/CR/...) == True
- Resolved : AND di tutte le variabili
  Pass_(Merg/Res)(HP/LP)_(GGF/VBF)_(ZZ/WZ)_(Tag/Untag)_(SR/CR/...) == False

Step 3) splitDataset.py

This script splits the dataframe produced at the previous step into one train and one test sample and saves them into two separate .pkl files. 3 input flags must be specified:

- analysis (-a): type of analysis ('merged' or 'resolved')
- channel (-c): the channel considered ('ggF' or 'VBF')
- training fraction (-t): relative size of the training sample, between 0 and 1

All these flags are mandatory but only the first two must be specified by the user (the last one can also assume its default value).

Step 4) buildDNN.py / buildPDNN.py

These scripts run the (parametric) Deep Neural Network. 7 input flags can be specified:

- analysis (-a): type of analysis ('merged' or 'resolved')
- channel (-c): the channel considered ('ggF' or 'VBF')
- nodes (-n): number of nodes of the (p)DNN
- layers (-l): number of layers of the (p)DNN
- epochs (-e): number of epochs for the training
- validation (-v): fraction of the training data that will actually be used for validation
- dropout (-d): fraction of the neurons that will be dropped during the training

All these flags are mandatory but only the first two must be specficied by the user (the other five can also assume their default values).

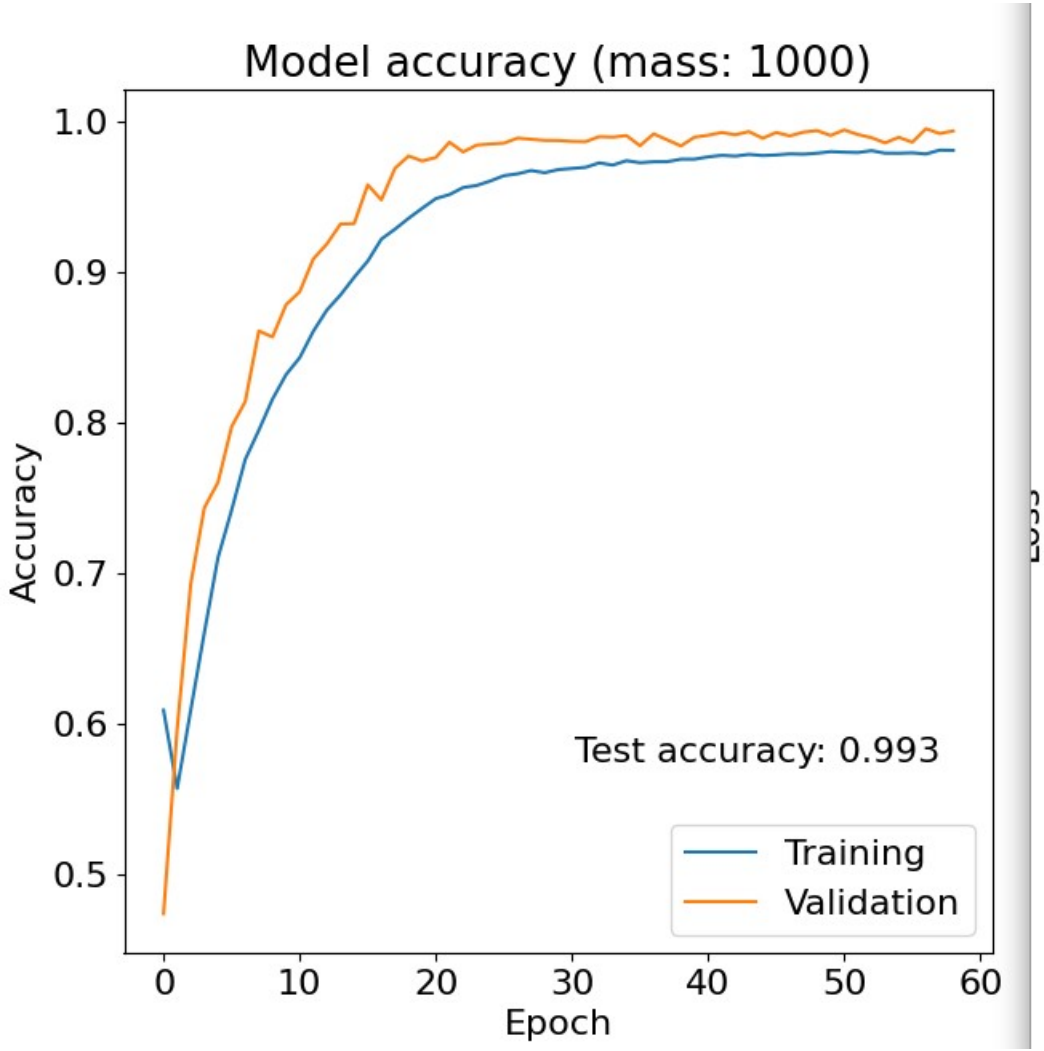
DNN

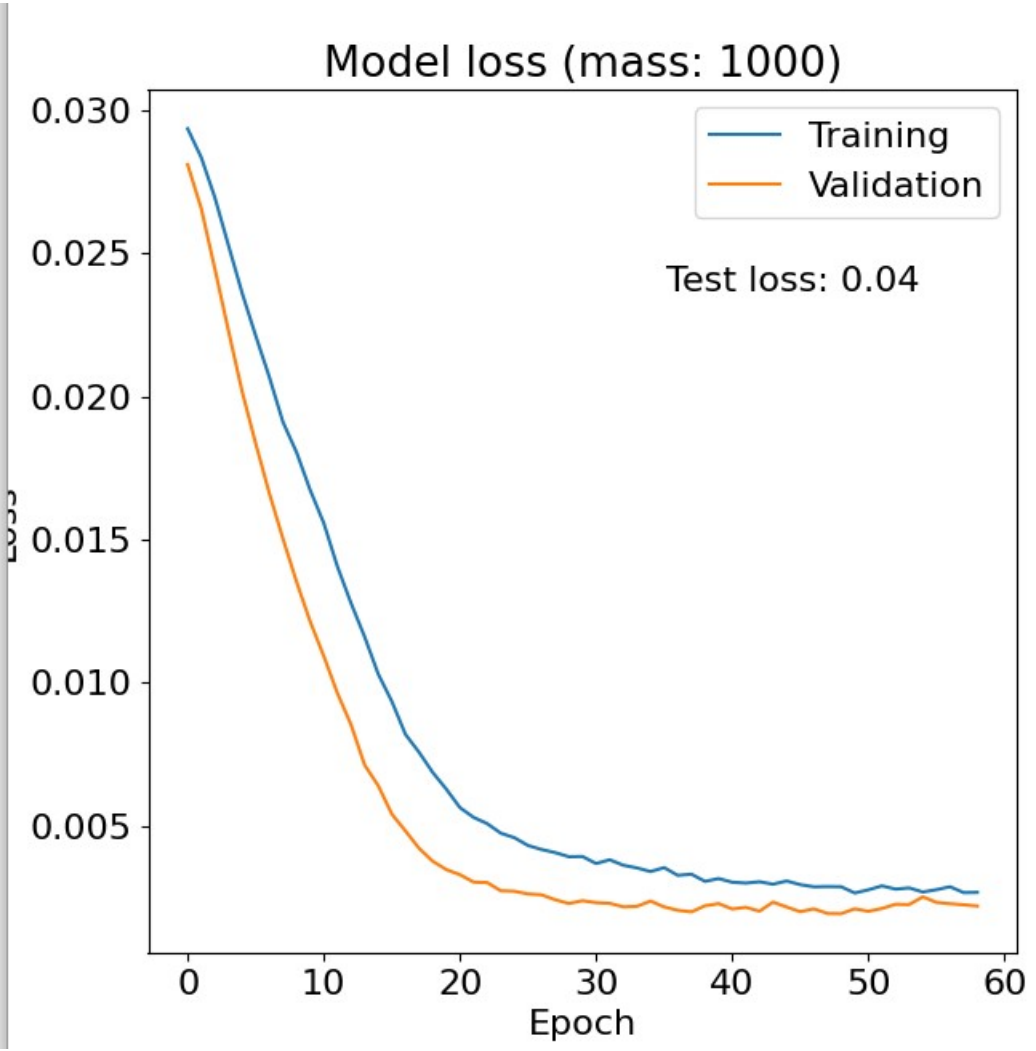
- Loading dei dataframe di input
- Distinzione segnale/background
- Per ogni valore di massa del segnale:

  - Selezione di segnale con la stessa massa
  - Scaling delle variabili e rimozione della massa
  - Calcolo di un peso per il segnale ed uno per il background per compensare la diversa statistica
  - Creazione di un unico dataframe contenente segnale e background
  - Shuffling dei dati
  - Addestramento della rete sul campione di train
  - Plot di accuracy e loss
  - Test della rete sugli eventi del campione di test con massa pari a quella selezionata
  - Plot di ROC, scores e matrice di confusione

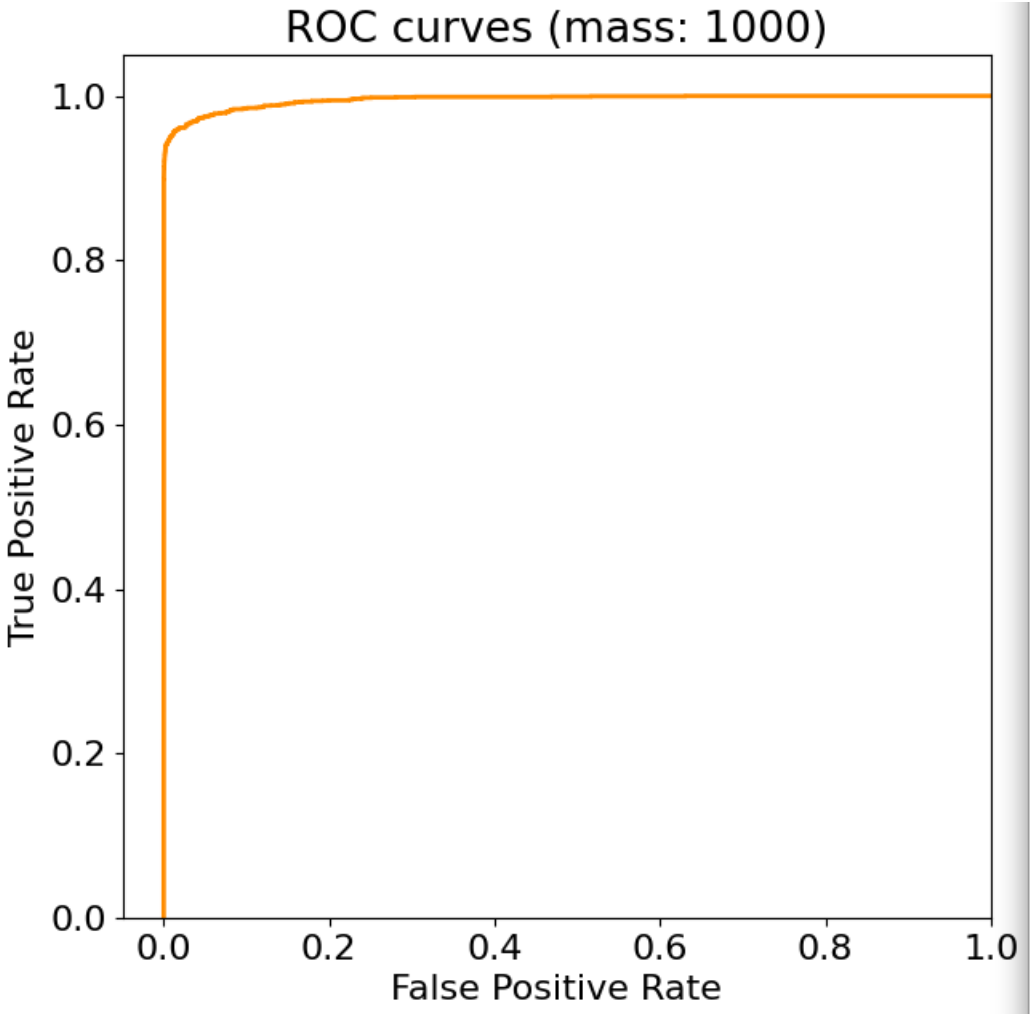
```
[apalazzo@wn-10 ~]$ cd /nfs/kloe/einstein4/HDBS/NN_Models/
[apalazzo@wn-10 NN_Models]$ ll
total 16
drwxrwx---+ 4 apalazzo gruppo1 4096 Apr 29 14:57 DNN
drwxrwx---+ 3 apalazzo gruppo1 4096 Apr 28 20:17 pDNN
[apalazzo@wn-10 NN_Models]$ ll DNN/
total 16
drwxrwx---+ 3 apalazzo gruppo1 4096 Apr 28 20:15 merged
drwxrwx---+ 4 apalazzo gruppo1 4096 Apr 29 15:00 resolved
[apalazzo@wn-10 NN_Models]$ ll DNN/merged/
total 8
drwxrwx---+ 6 apalazzo gruppo1 4096 Apr 28 20:16 ggF
[apalazzo@wn-10 NN_Models]$ ll DNN/merged/ggF/
total 32
drwxrwx---+ 2 apalazzo gruppo1 4096 Apr 29 11:44 1000
drwxrwx---+ 2 apalazzo gruppo1 4096 Apr 29 13:08 3000
drwxrwx---+ 2 apalazzo gruppo1 4096 Apr 29 13:07 5000
drwxrwx---+ 2 apalazzo gruppo1 4096 Apr 29 13:07 700
[apalazzo@wn-10 NN_Models]$ ll DNN/merged/ggF/1000/
total 272
-rw-rw----+ 1 apalazzo gruppo1 42539 May  4 09:20 Accuracy.png
-rw-rw----+ 1 apalazzo gruppo1 38745 May  4 09:20 ConfusionMatrix.png
-rw-rw----+ 1 apalazzo gruppo1 34424 May  4 09:20 Higgs.h5
-rw-rw----+ 1 apalazzo gruppo1 34400 Apr 29 11:37 Higgs_t0.7_n32_l2_e150_v0.2.h5
-rw-rw----+ 1 apalazzo gruppo1  3115 Apr 29 11:37 Higgs_t0.7_n32_l2_e150_v0.2.yaml
-rw-rw----+ 1 apalazzo gruppo1  3115 May  4 09:20 Higgs.yaml
-rw-rw----+ 1 apalazzo gruppo1   942 May  4 09:20 logFile.txt
-rw-rw----+ 1 apalazzo gruppo1 30835 May  4 09:20 Loss.png
-rw-rw----+ 1 apalazzo gruppo1 29735 May  4 09:20 ROC.png
-rw-rw----+ 1 apalazzo gruppo1 37710 May  4 09:20 Scores.png
```

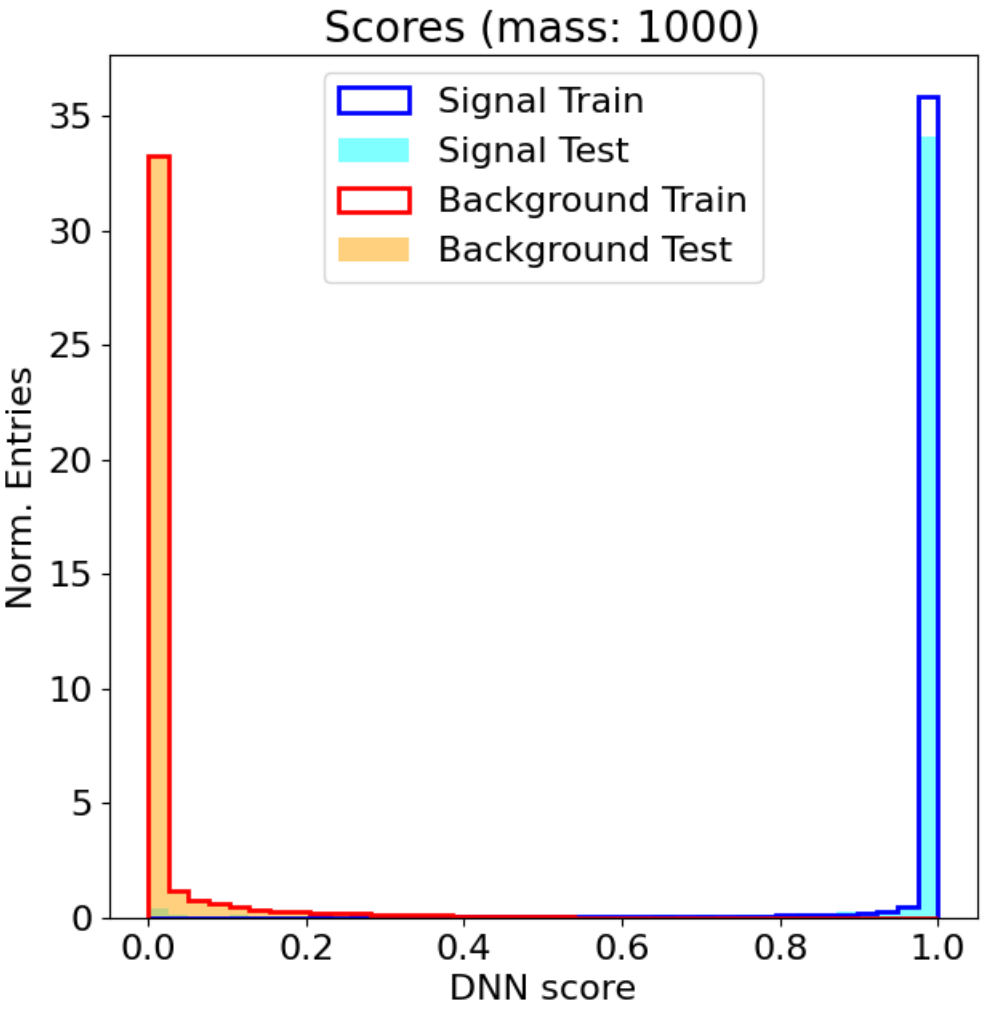
DNN merged ggF



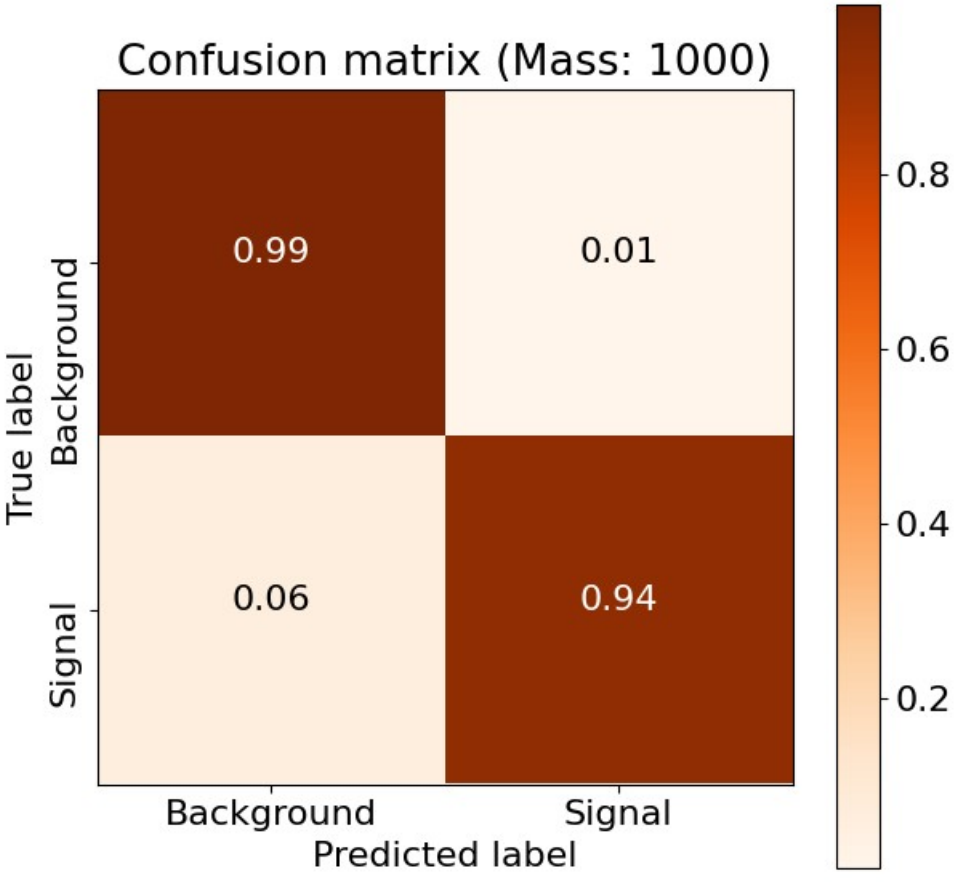


DNN merged ggF





DNN merged ggF

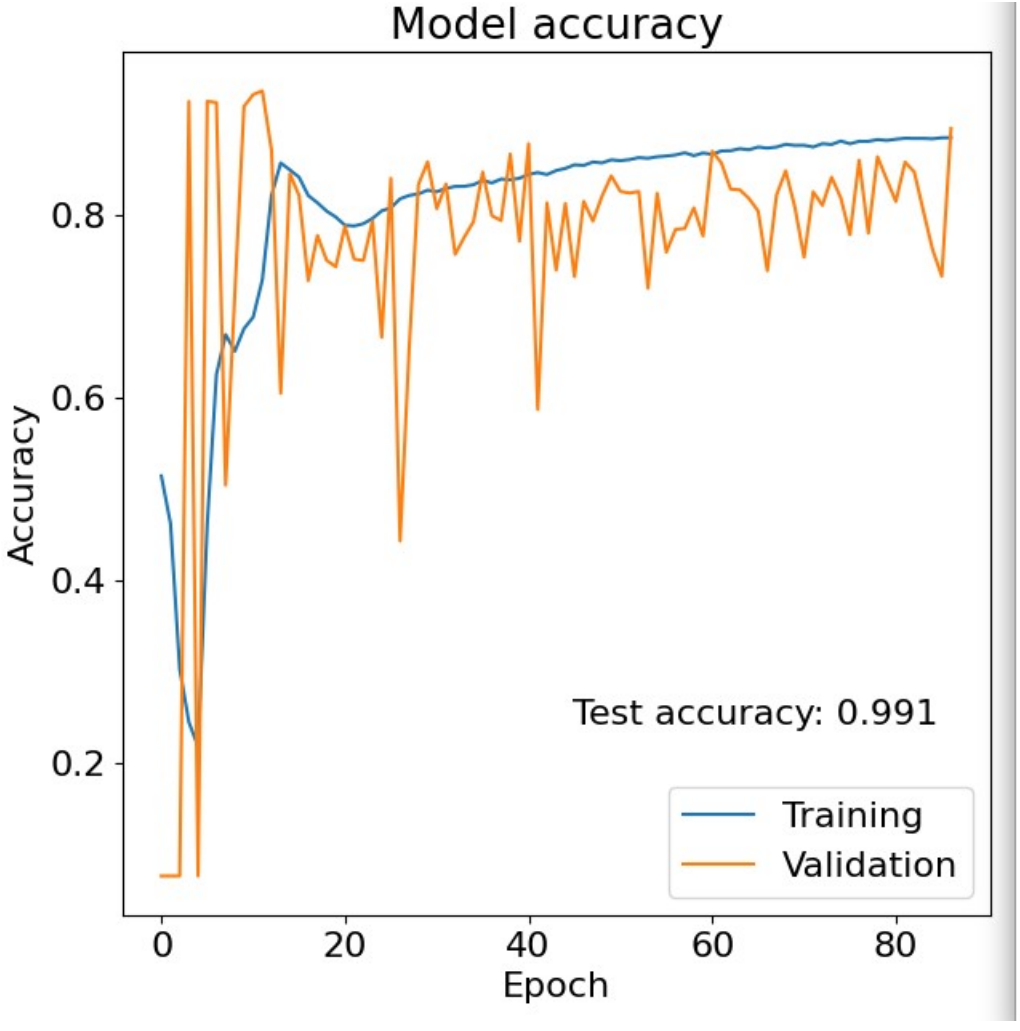


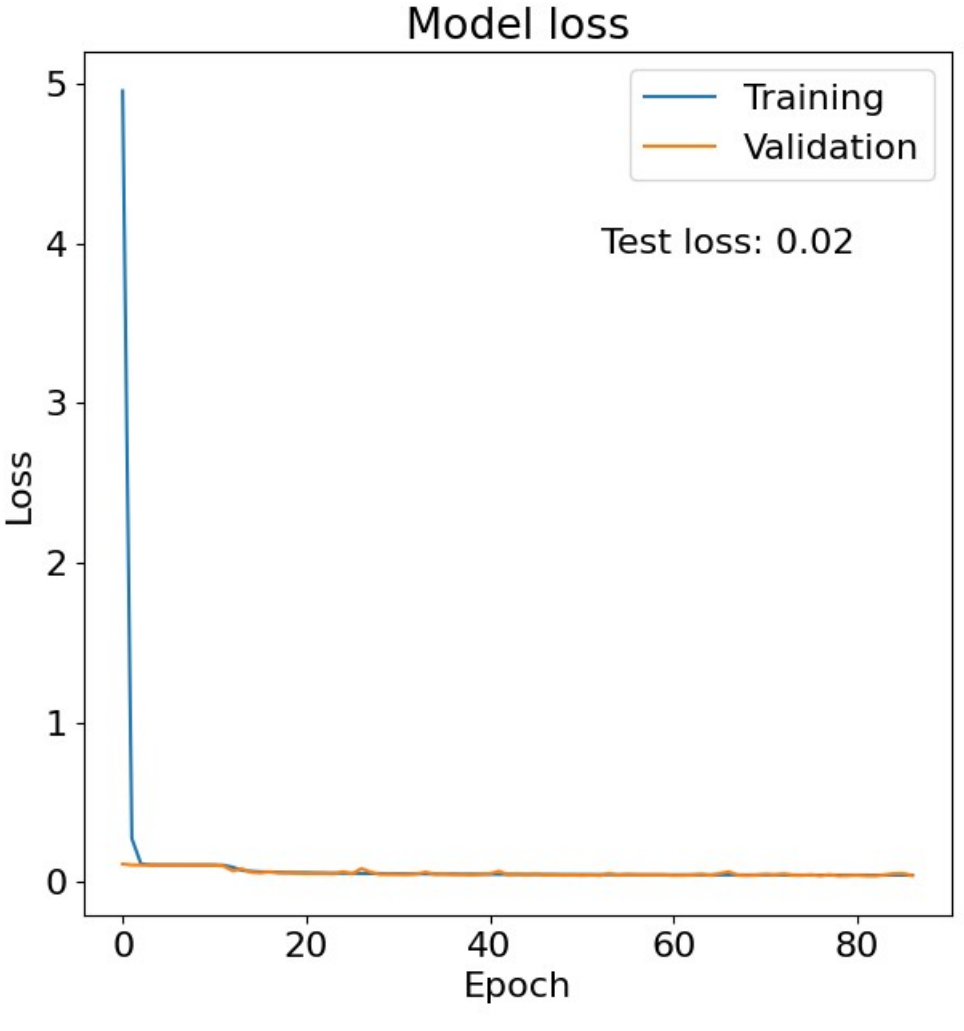
pDNN

- Loading dei dataframe di input
- Distinzione segnale/background
- Salvataggio delle masse del segnale
- Associazione al background di una massa casuale tra quelle del segnale
- Calcolo di un peso per il segnale ed uno per il background per compensare la diversa statistica
- Creazione di un unico dataframe contenente segnale e background
- Shuffling dei dati
- Scaling dei dati
- Addestramento della rete sul campione di train
- Plot di accuracy, loss
- Per ogni valore della massa del segnale:
  - Associazione della massa selezionata agli eventi di background del campione di test
  - Plot di ROC, scores e matrice di confusione

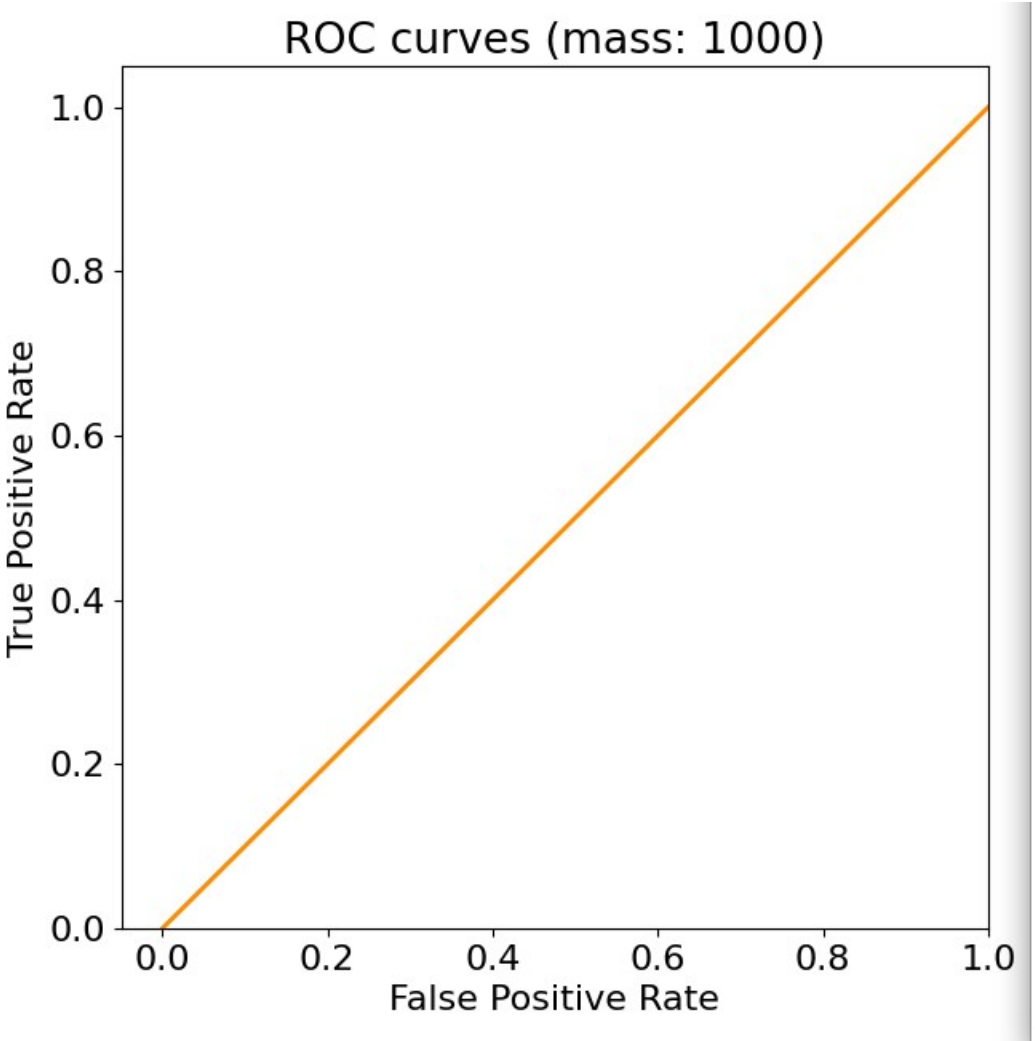
```
[apalazzo@wn-10 NN_Models]$ ll
total 16
drwxrwx---+ 4 apalazzo gruppo1 4096 Apr 29 14:57 DNN
drwxrwx---+ 3 apalazzo gruppo1 4096 Apr 28 20:17 pDNN
[apalazzo@wn-10 NN_Models]$ ll pDNN/
total 8
drwxrwx---+ 3 apalazzo gruppo1 4096 Apr 28 20:17 merged
[apalazzo@wn-10 NN_Models]$ ll pDNN/merged/
total 8
drwxrwx---+ 6 apalazzo gruppo1 4096 May  1 15:05 ggF
[apalazzo@wn-10 NN_Models]$ ll pDNN/merged/ggF/
total 196
drwxrwx---+ 2 apalazzo gruppo1  4096 May  4 18:07 1000
drwxrwx---+ 2 apalazzo gruppo1  4096 Apr 28 20:18 3000
drwxrwx---+ 2 apalazzo gruppo1  4096 May  3 14:47 5000
drwxrwx---+ 2 apalazzo gruppo1  4096 Apr 28 20:18 700
-rw-rw----+ 1 apalazzo gruppo1 45878 May  4 18:06 Accuracy.png
-rw-rw----+ 1 apalazzo gruppo1 34424 May  4 18:06 Higgs.h5
-rw-rw----+ 1 apalazzo gruppo1 34424 Apr 28 20:18 Higgs_t0.7_n32_l2_e150_v0.2.h5
-rw-rw----+ 1 apalazzo gruppo1  3115 Apr 28 20:18 Higgs_t0.7_n32_l2_e150_v0.2.yaml
-rw-rw----+ 1 apalazzo gruppo1  3115 May  4 18:06 Higgs.yaml
-rw-rw----+ 1 apalazzo gruppo1   777 May  4 18:06 logFile.txt
-rw-rw----+ 1 apalazzo gruppo1 28684 May  4 18:06 Loss.png
[apalazzo@wn-10 NN_Models]$ ll pDNN/merged/ggF/1000/
total 164
-rw-rw----+ 1 apalazzo gruppo1 37845 May  4 18:06 ConfusionMatrix.png
-rw-rw----+ 1 apalazzo gruppo1 34400 Apr 29 12:35 Higgs.h5
-rw-rw----+ 1 apalazzo gruppo1  3115 Apr 29 12:35 Higgs.yaml
-rw-rw----+ 1 apalazzo gruppo1   925 May  4 18:06 logFile.txt
-rw-rw----+ 1 apalazzo gruppo1 39400 May  4 18:06 ROC.png
-rw-rw----+ 1 apalazzo gruppo1 38413 May  4 18:06 Scores.png
```

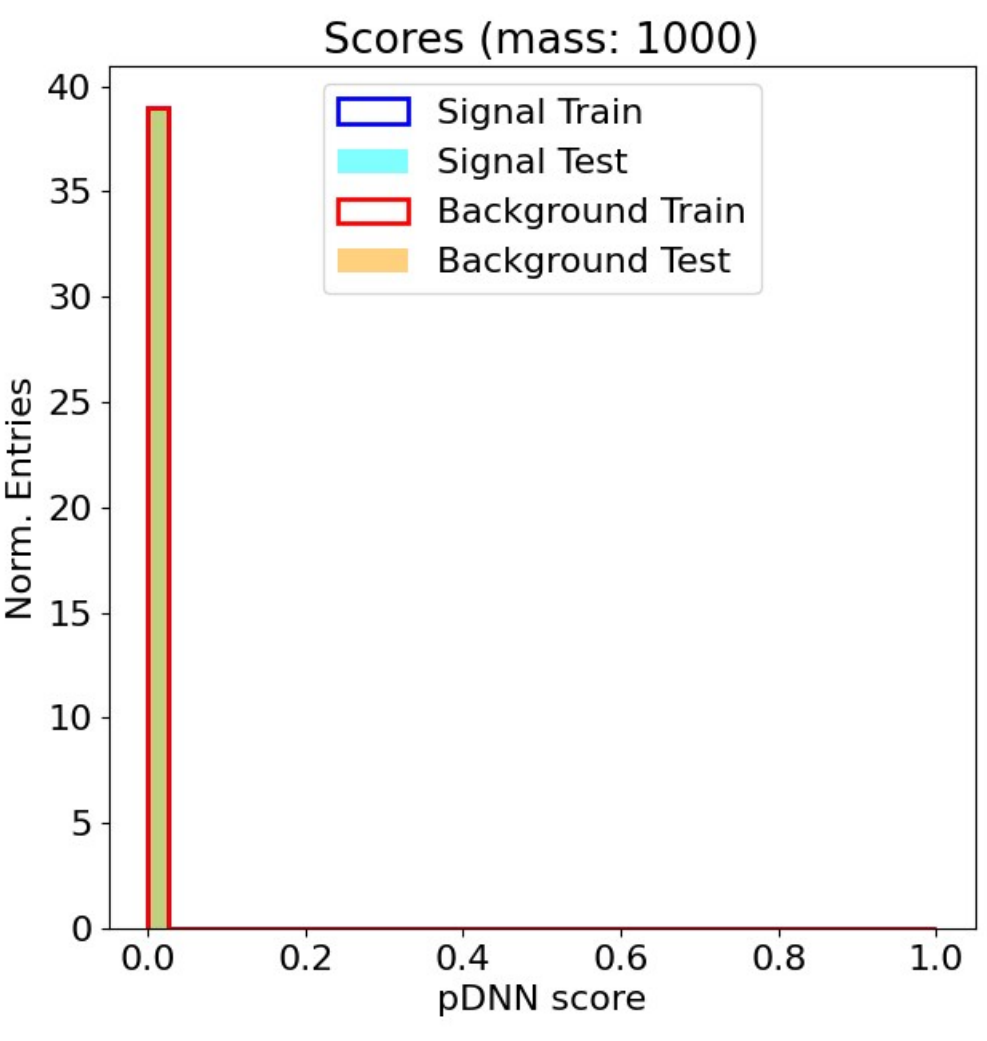
pDNN merged ggF





pDNN merged ggF





pDNN merged ggF