Cluster Improvement on Features from Neural Network

By:
Sai Kiran Putta

Dataset

➢ Experimentation has been done on Human Gender data. Features include :

Frequency – Mean, SD, Median, Q25, Q75, Centroid, Peak

Fundamental Freq – Mean, SD, Median, Max, Min

Dominant Freq – Mean, Min, Max, range

Modulation Index

Skewness

Kurtosis

Label – Male or Female

Scope of the project

- Run Kmeans on original data, features from Neural Networks and compare them.
- Run more experiments. They are as follows :
  - Create 10% of data as outlier for one feature. Rerun above step
  - Create 10% of data as outlier for all features. Rerun above step
  - Mislabel 10% of targets (Can Neural Network handle it?)
  - Mislabel 50% of targets

Primary Metrics

- **Cluster Purity:**

  Sum of Maximum class in each cluster / Total Number of obs

  Value ranges from 0 – 1

- **Improvement :**

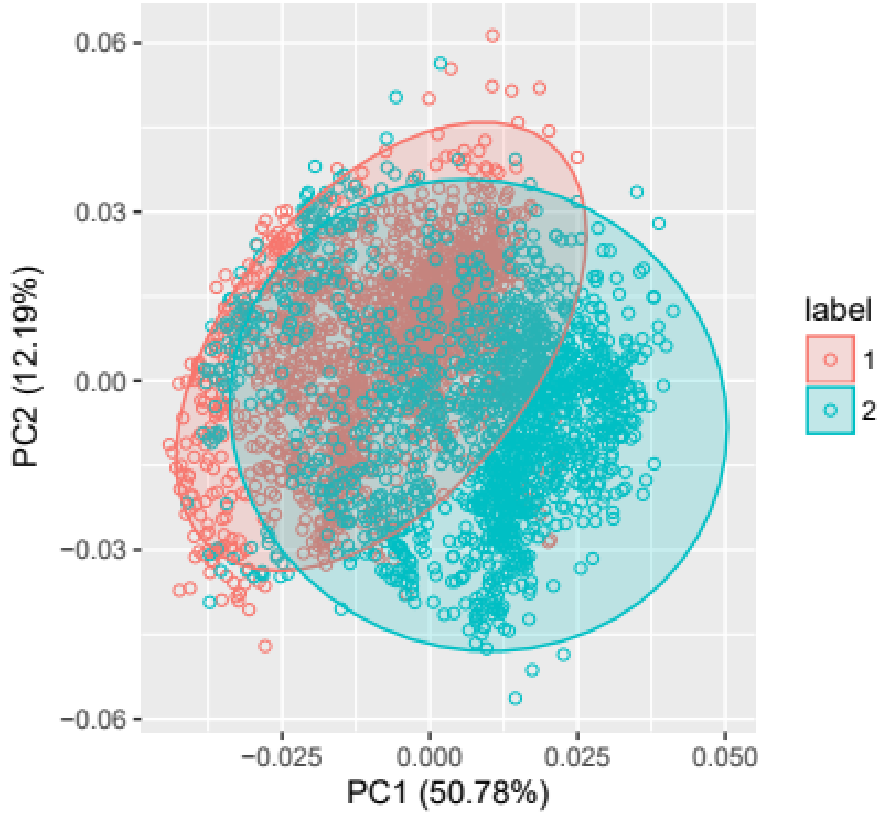
  How well off are we as compared to original Purity.

  ((Current_purity/Original_purity)-1) * 100

  Value ranges from 0 – 100%

Outlook of data – This is what we are dealing with

PCA on the original features

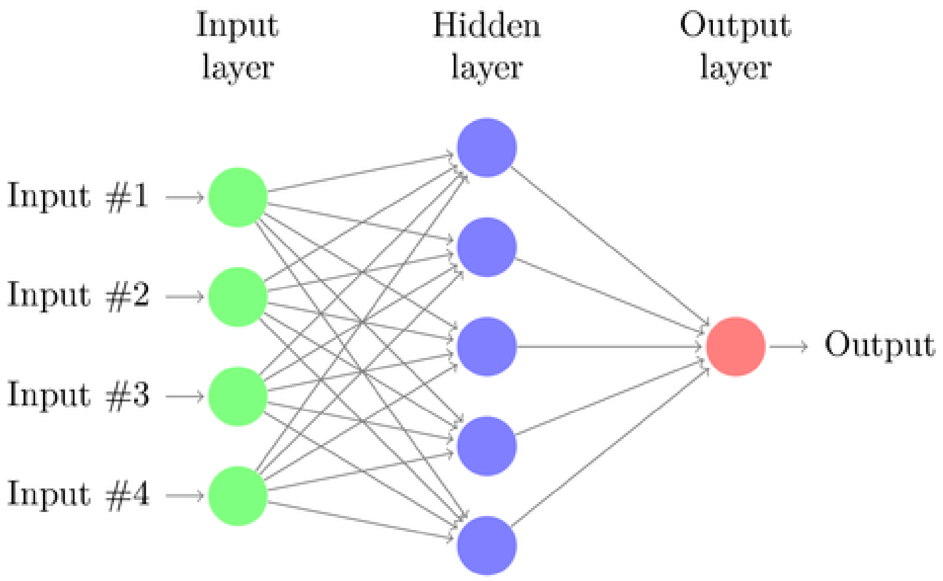


Classes are overlapping. Not in a great position to cluster both classes properly.

Cluster Purity : 0.65

Is there a way to seperate the data space?

- Here come Neural Networks!



Let’s change the underlying structure of data using Neural Networks!

Features :
We extract the value out of the Activation Function for all the neurons in the network.

Neural Network and Kmeans settings

- Since we have 2 classes we are assuming k should be 2

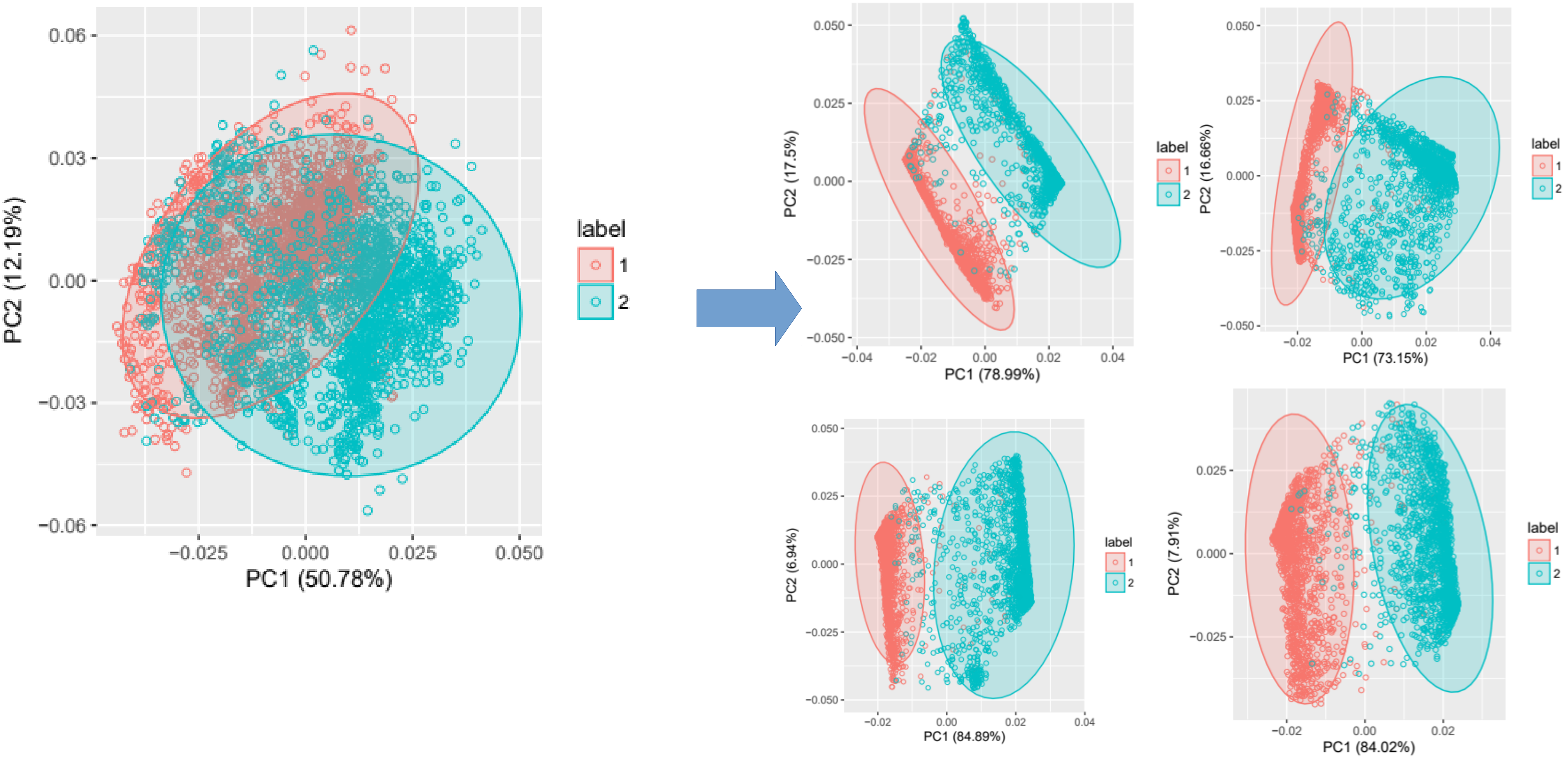
- With a little trail and error, here are our hyper-parameters.
- Learning Rate – 0.01

  Activation Function – Tanh

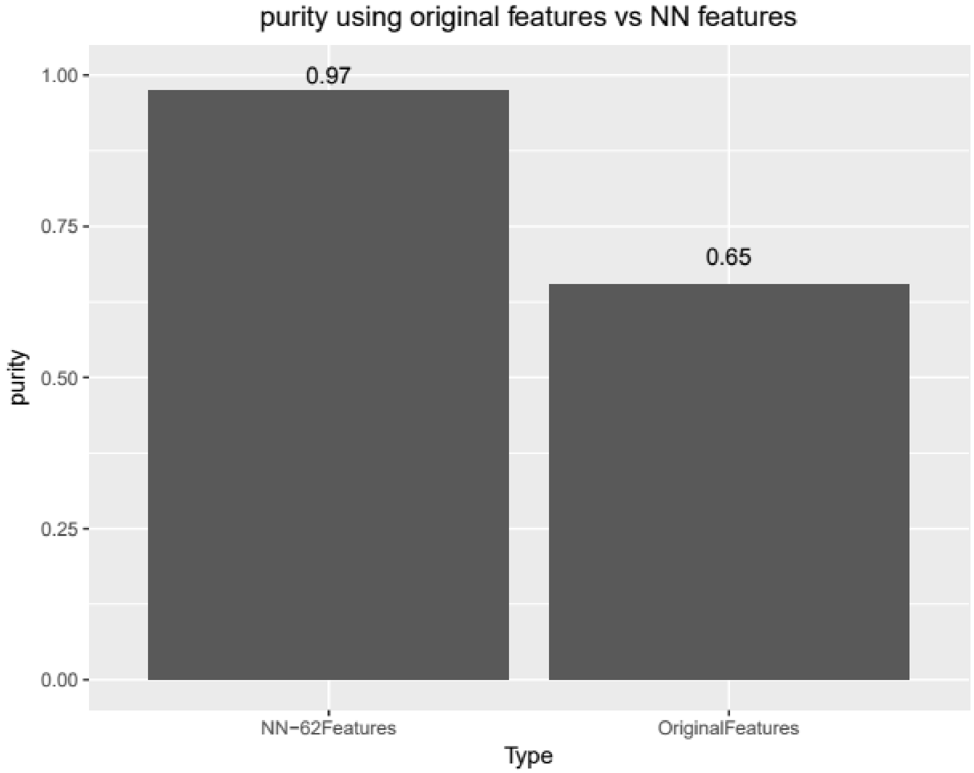
  Epochs – 15

  Hidden Layer and Neurons – Variable

Changing Feature Space



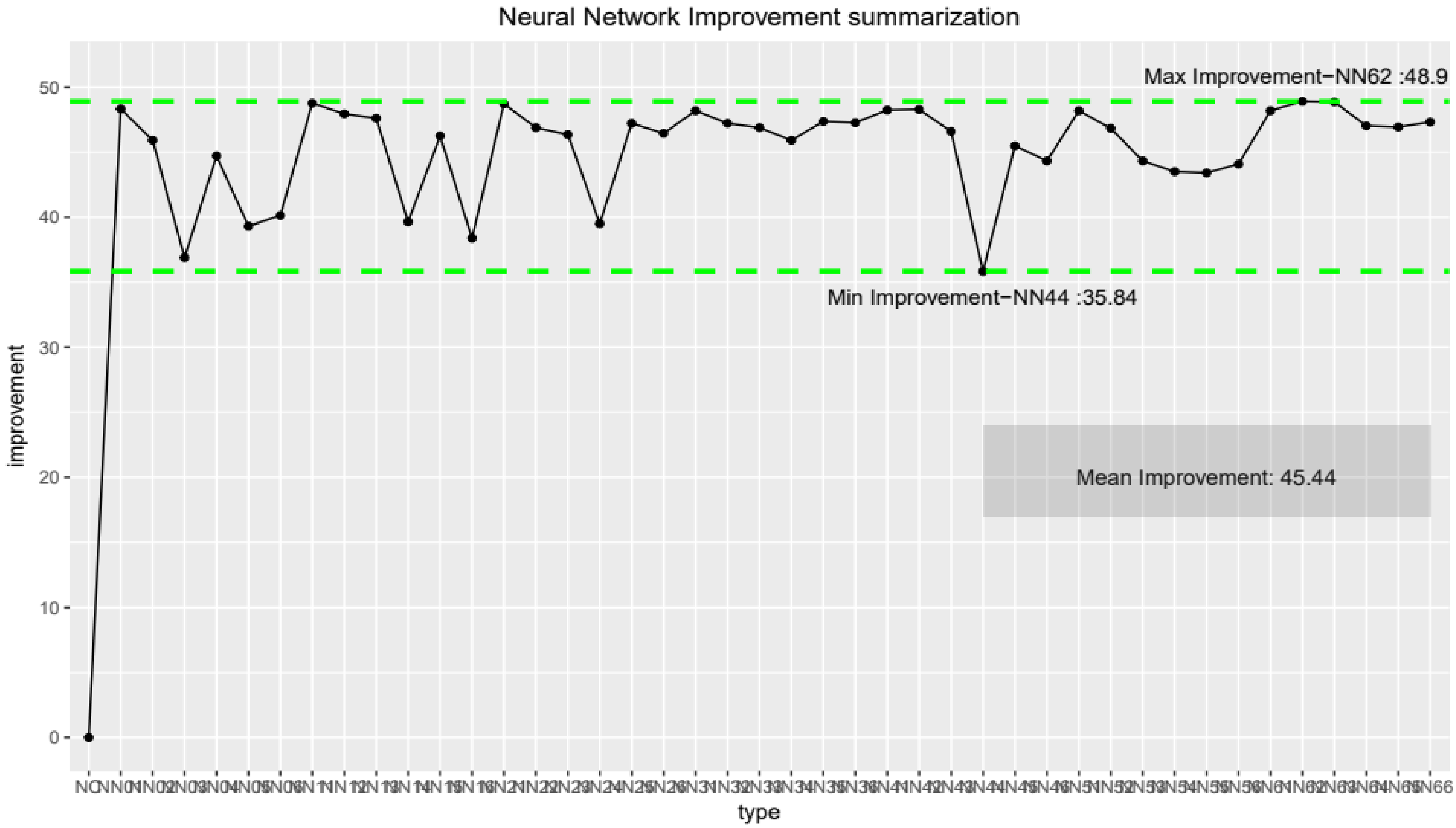
Purity Comparison



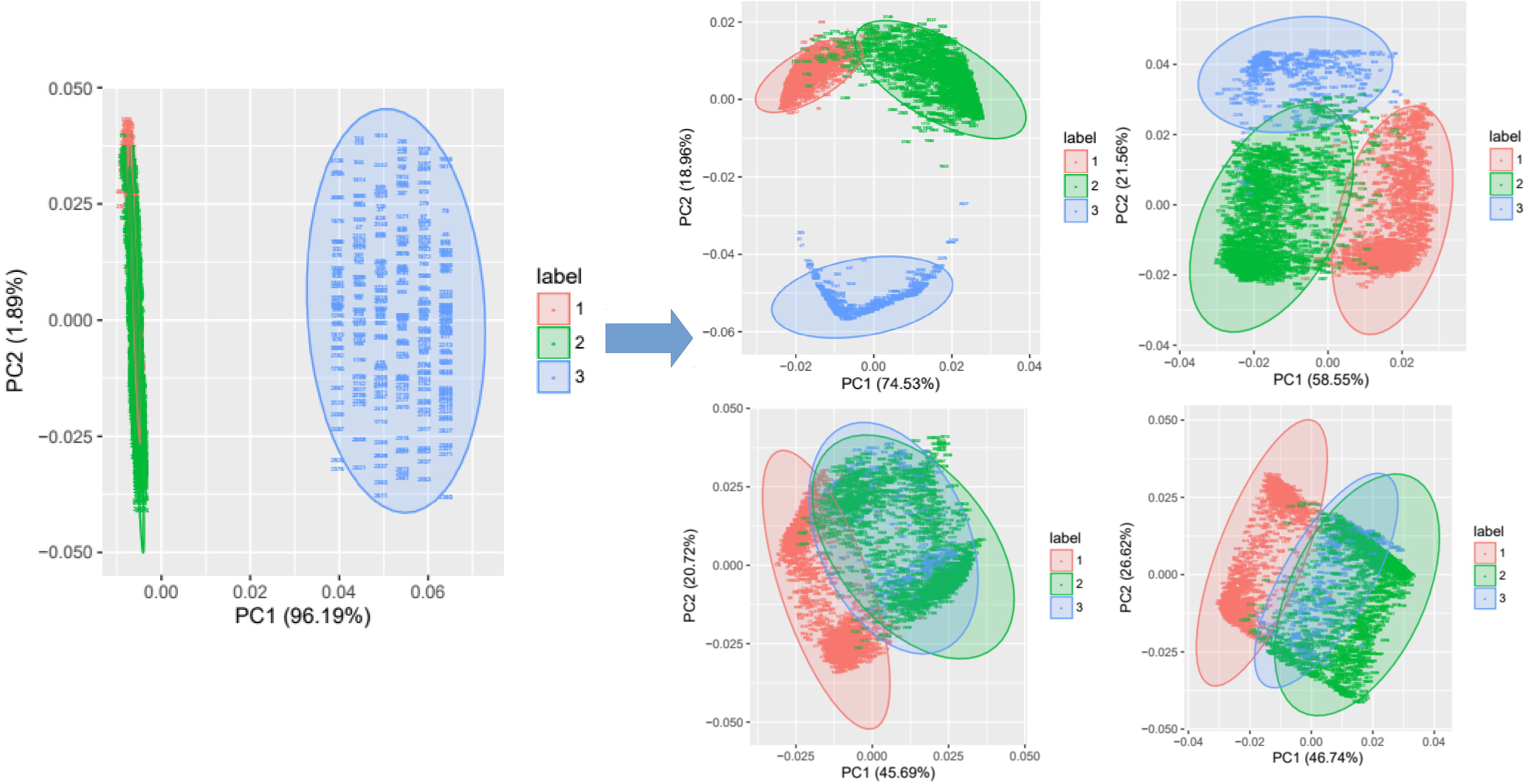
The best purity is obtained by a Neural Network with 6, 2 as Hidden Layer setting with purity – 0.97

A significant increase from 0.65!

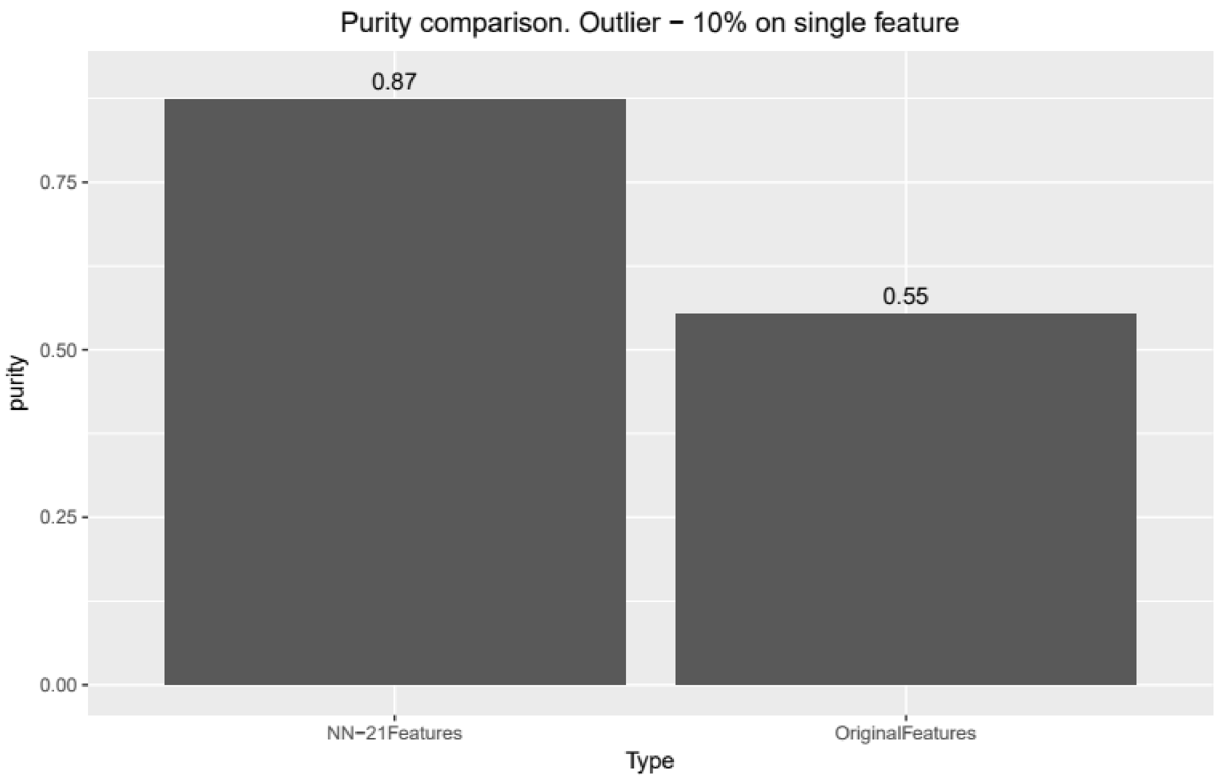
Result Summary



Experiment – 1
10% of data into outlier for one feature

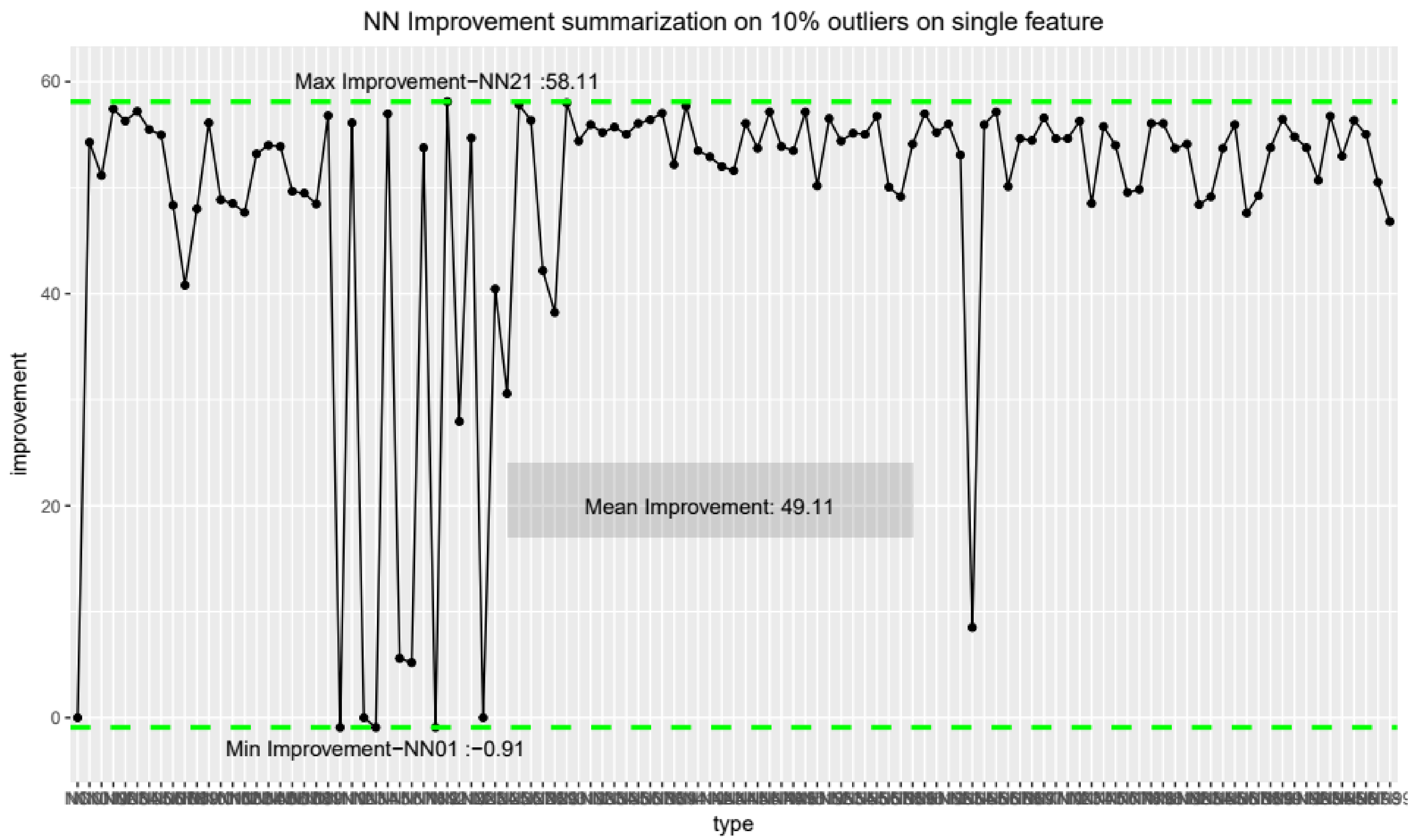


Experiment1 – Purity Comparison

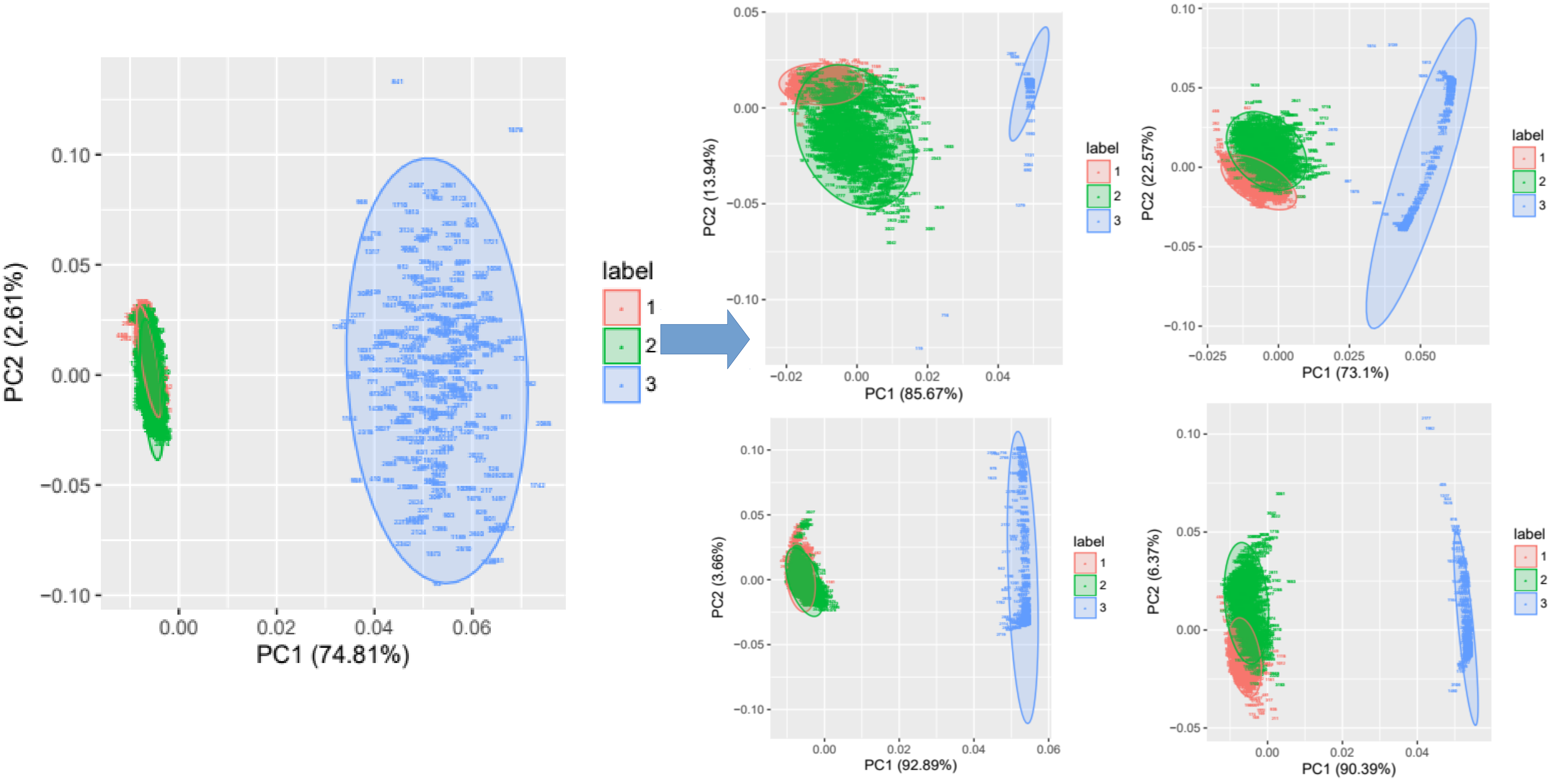


The best purity is obtained by a Neural Network with 2, 1 as Hidden Layer setting with purity – 0.87

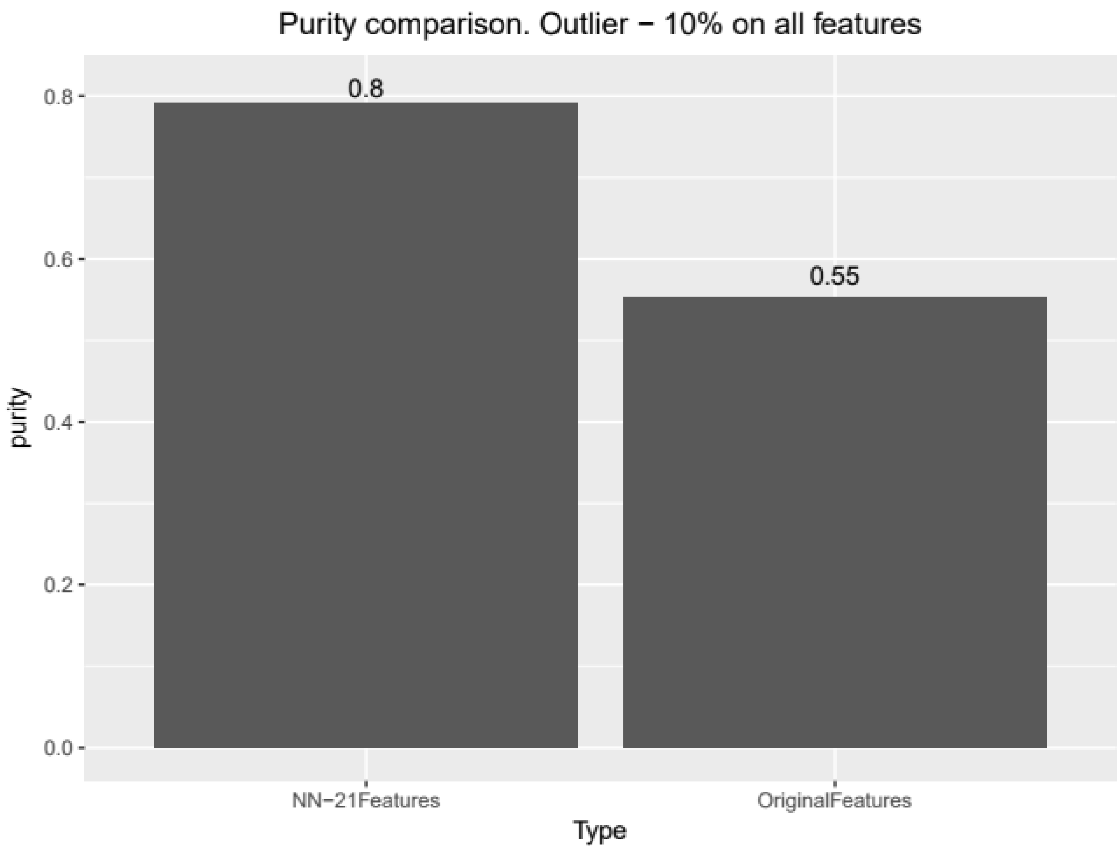
Experiment1 – Result Summary



Experiment – 2
10% of data into outlier for all features

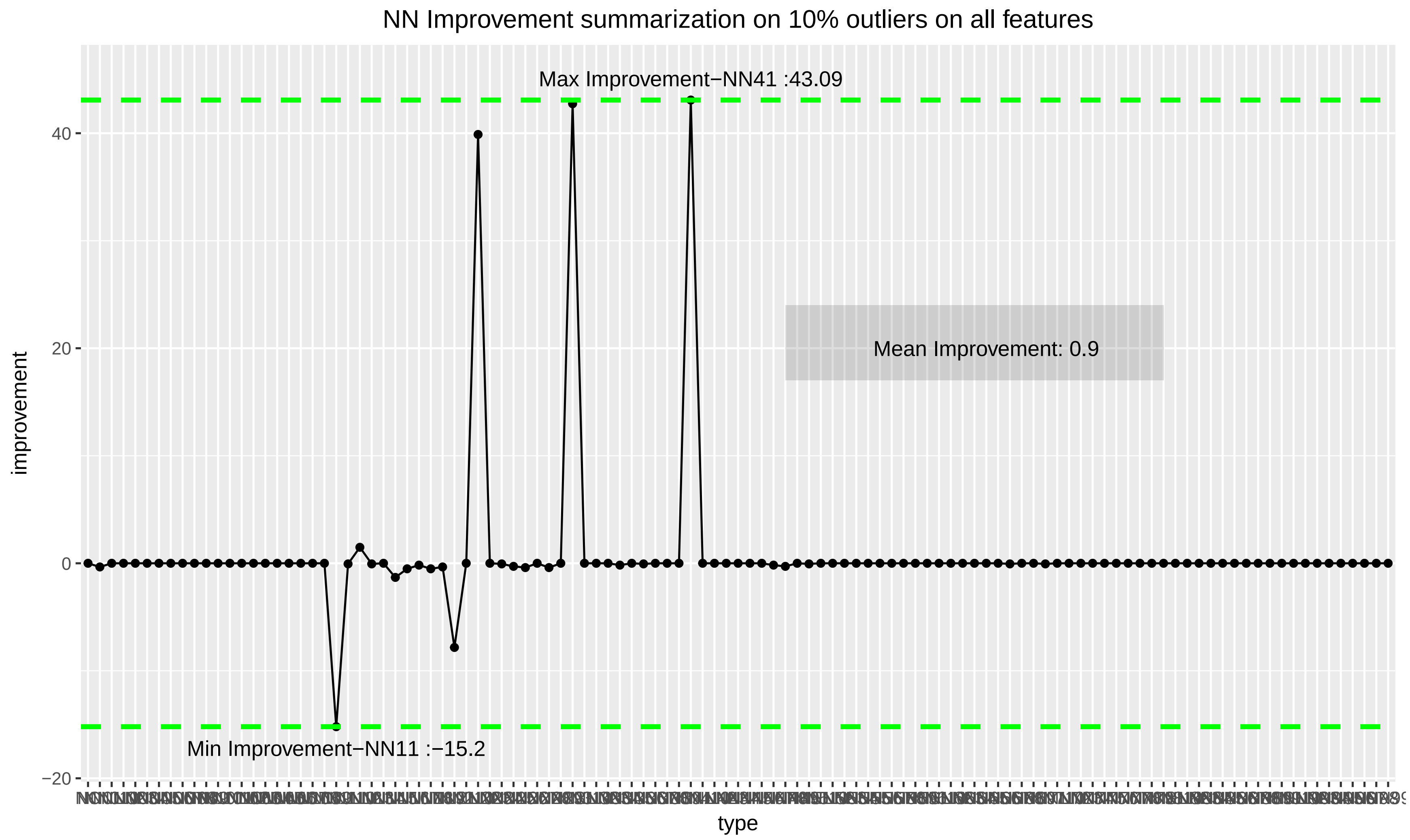


Experiment2 – Purity Comparison

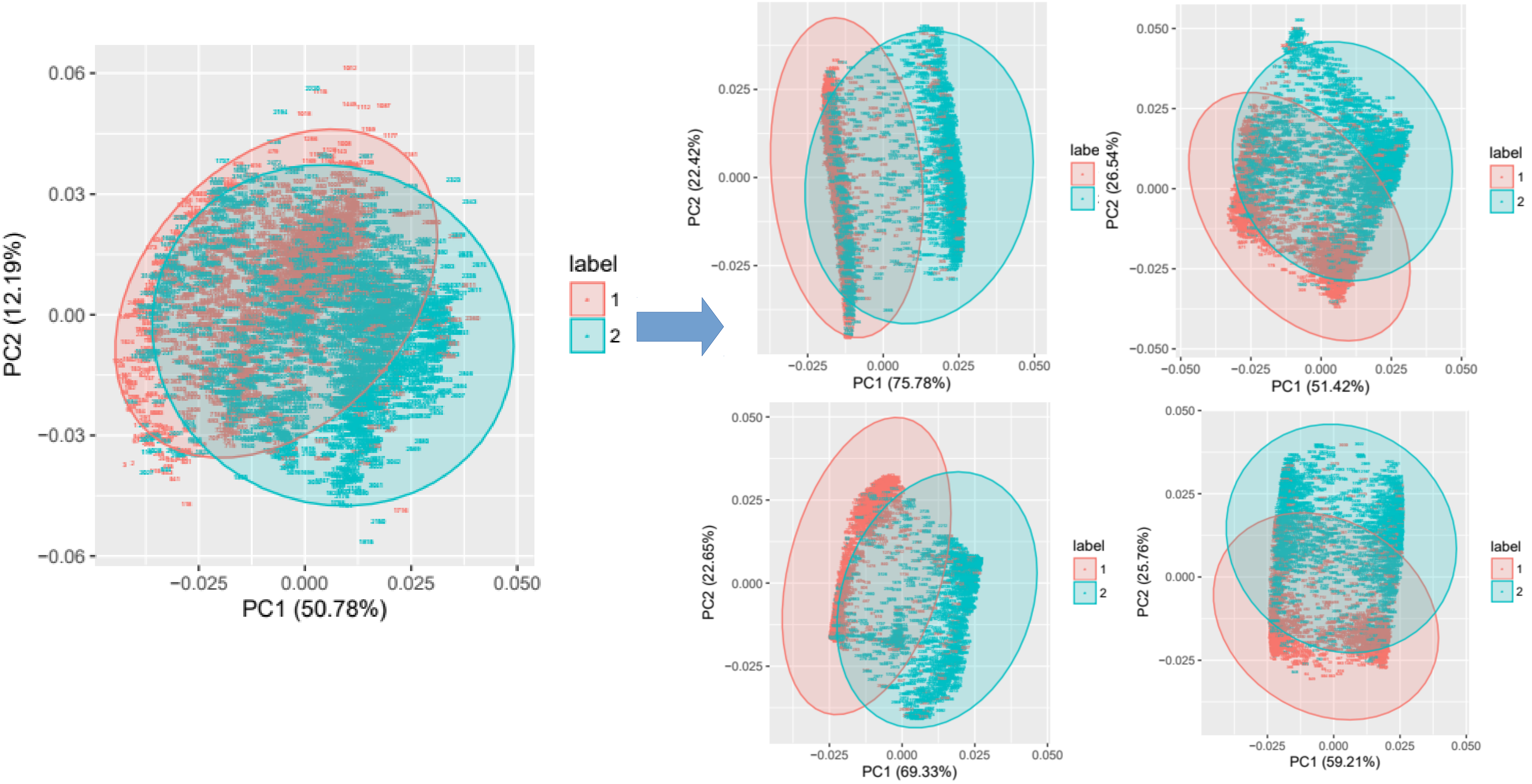


The best purity is obtained by a Neural Network with 2, 1 as Hidden Layer setting with purity – 0.8

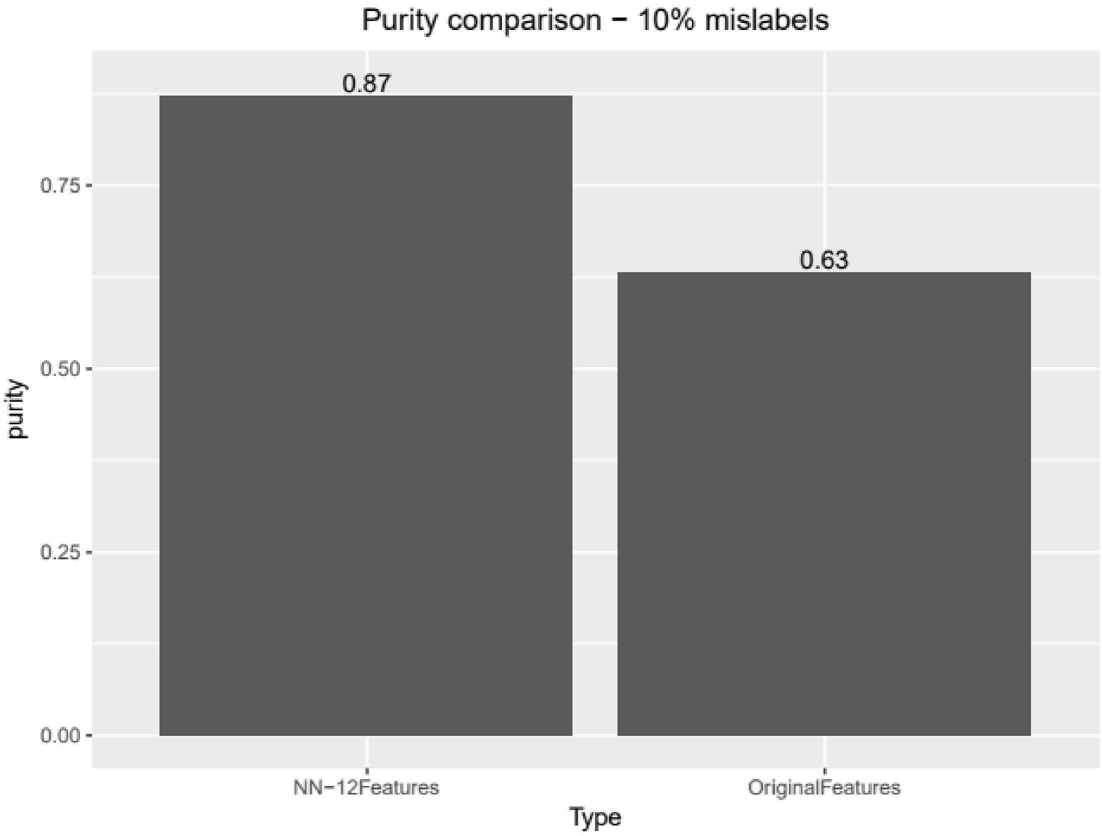
Experiment2 – Result Summary



Experiment – 3
10% of mislabels

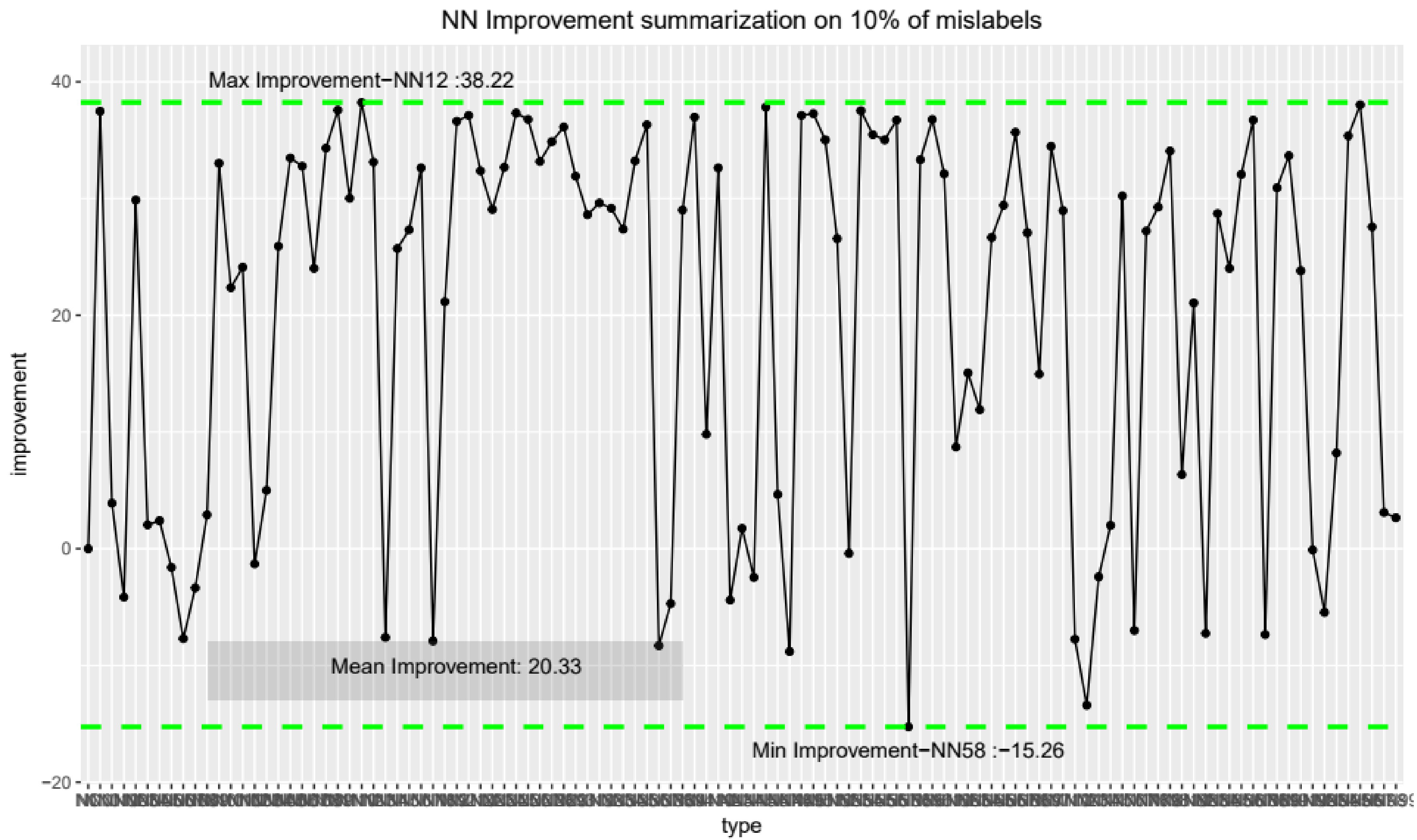


Experiment3 – Purity Comparison

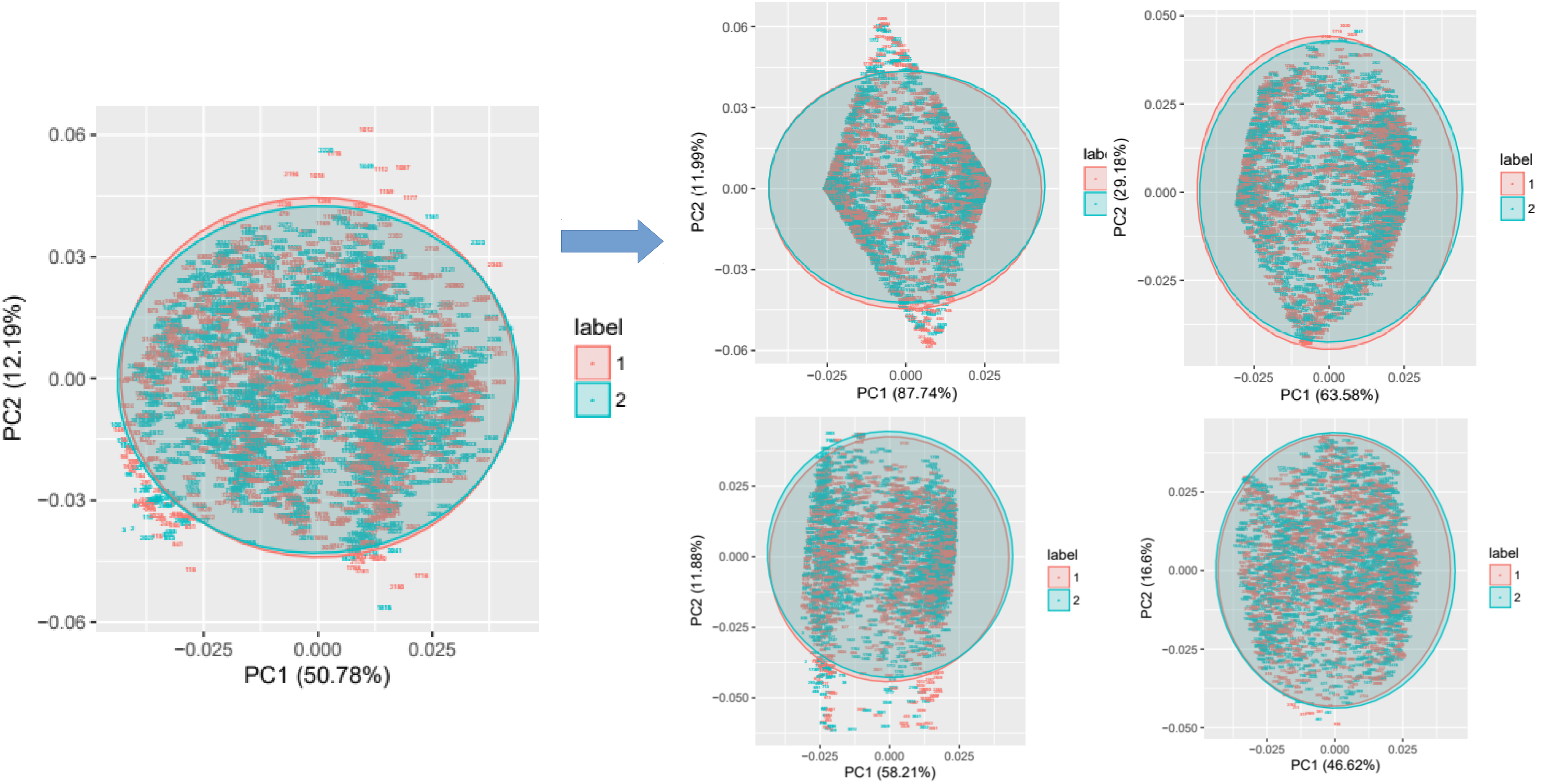


The best purity is obtained by a Neural Network with 1, 2 as Hidden Layer setting with purity – 0.87

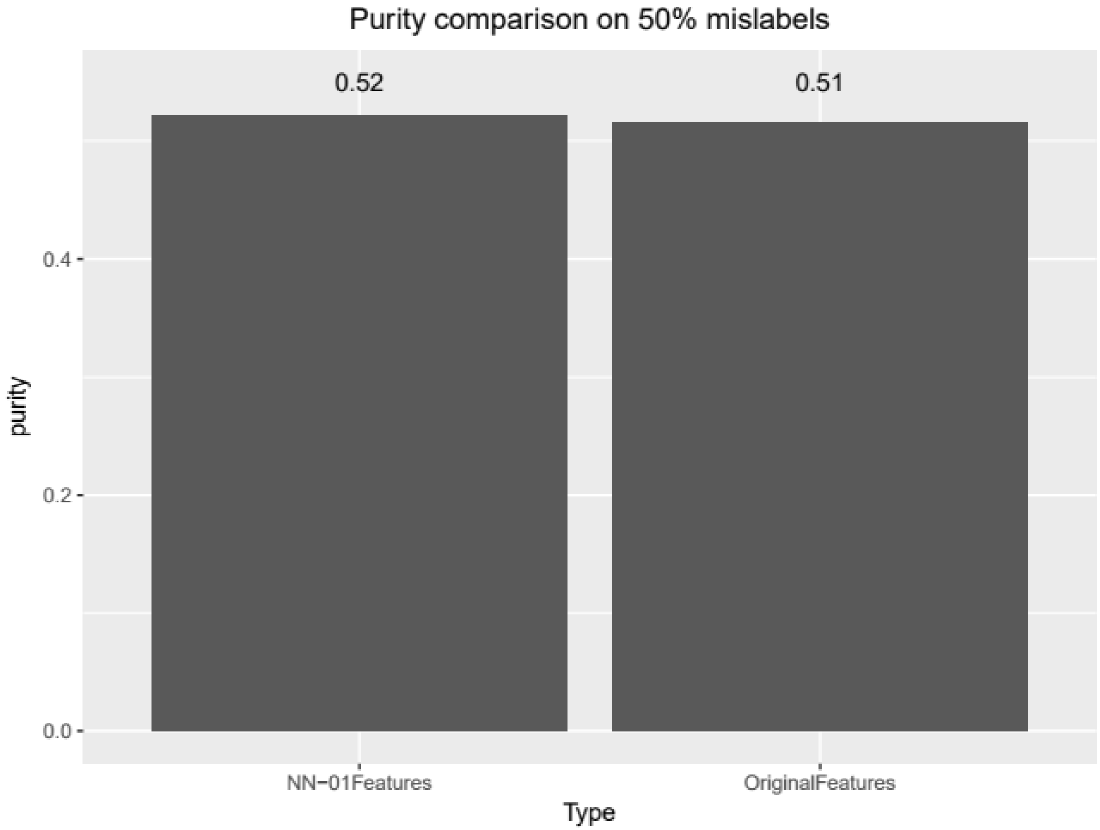
Experiment3 – Result Summary



Experiment – 4
50% of mislabels



Experiment4 – Purity Comparison

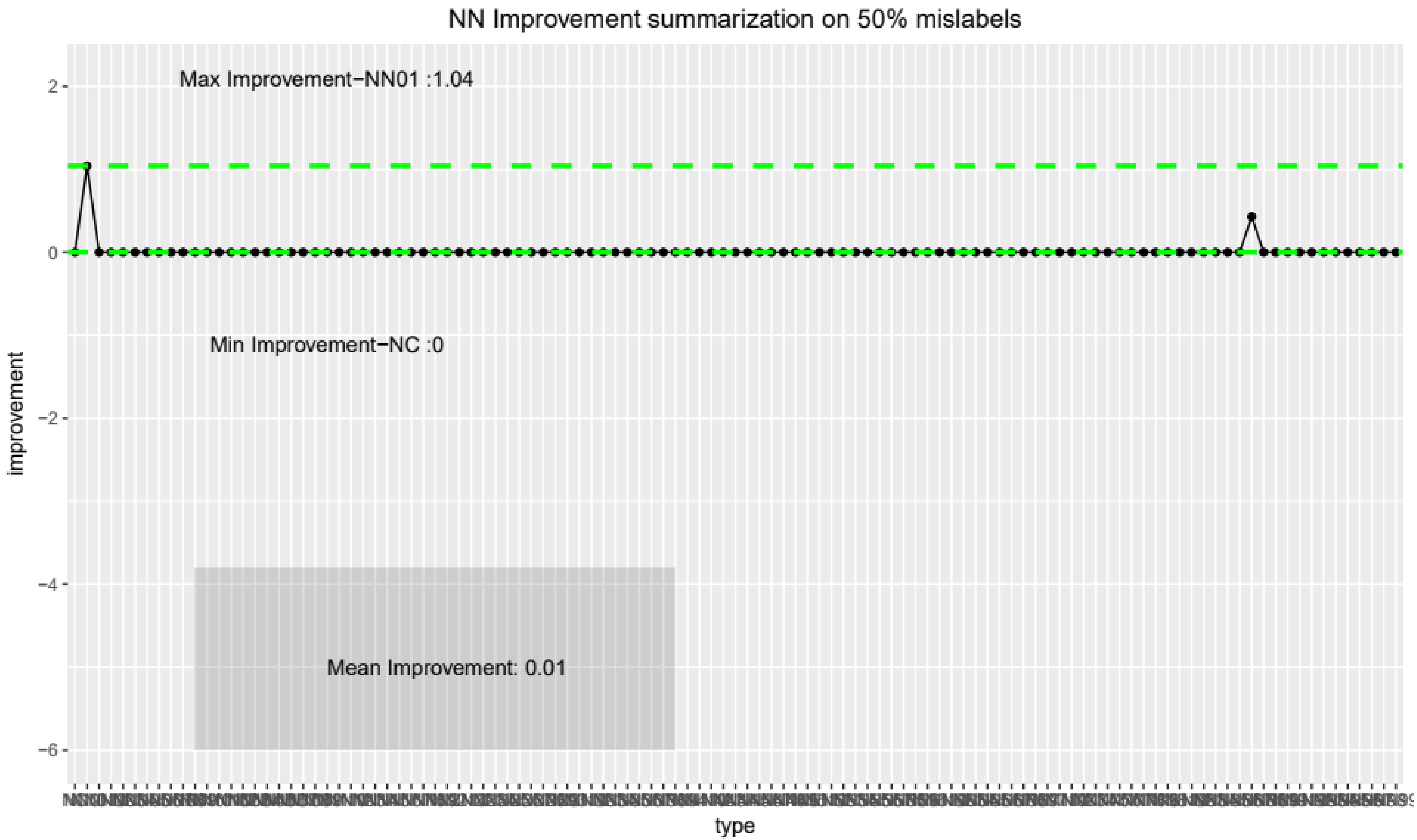


Neural Network doesn’t do well when big number of targets are mismatched.

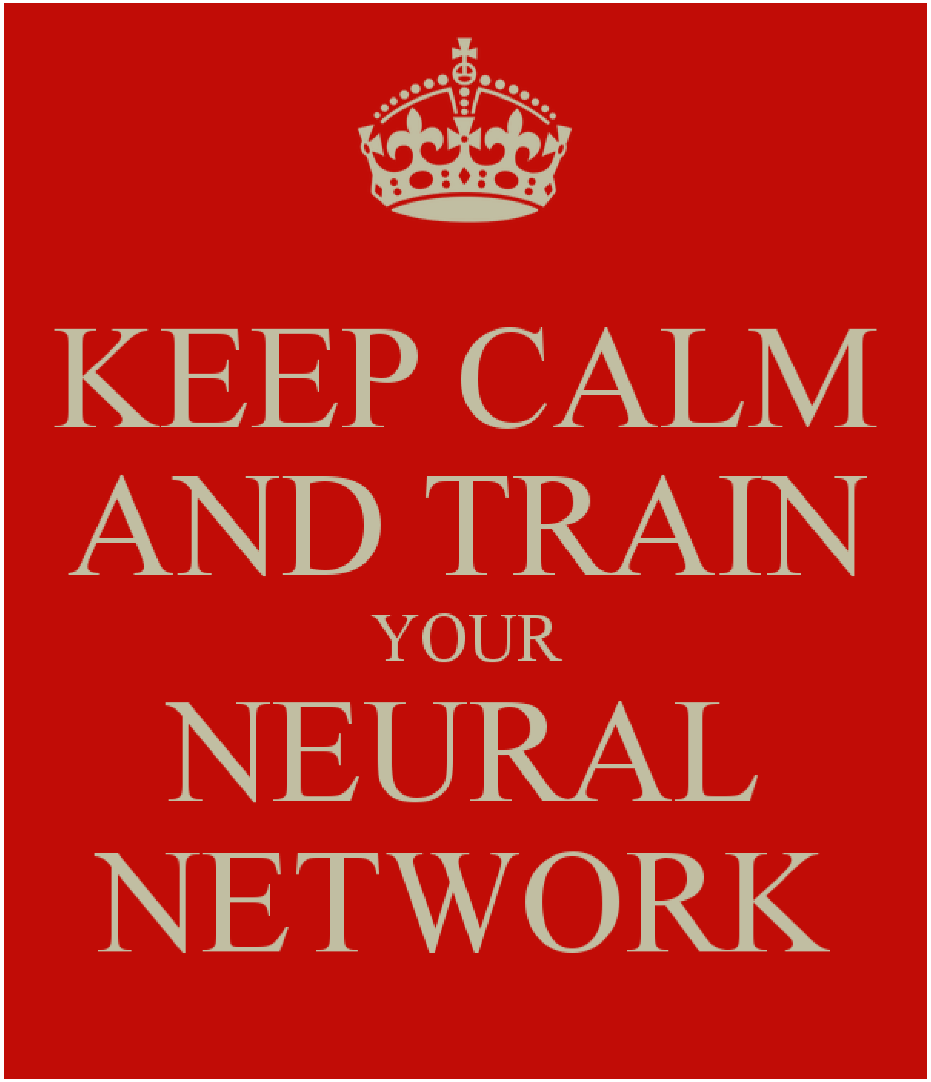
The best purity we obtained is 0.52 for a Neural Network setting 0,1

But it still leads by 0.01! 😛

Experiment4 – Result Summary



For better results apply Neural Networks somehow!



DEEP LEARNING
DEEP LEARNING EVERYWHERE
memes.com

Any Questions?

Thank you ☺