Date: July 2, 2025

From G. Pang , **File is BRCA1-Simulation and Dataset-creation2**

-------------------------------------------------------------------------------------------------------------------

**Two jupyter files attached:**

**Dataset-creation2 to create two dataset files: train\_vector2.csv and test\_vector2.csv**

**Run the above to generate the two files.**

**BRAC1-Simulation : in this file, note that I have messages tat says “NO NOT RUN …” for it is related to another ANN model not useful any more.**

**Dataset description:**

Training samples: 4800 items: 16384 inputs, one binary output (0 or 1)

File is train\_vectors2.csv

new\_train.shape = (4800, 16385)

The input vector is concatenated from the Reference vector (8192 elements) and the Variant vector (8192 elements).

* 2400 items are labelled 0, which have Variant vector 0.1%, 0.2% or 0.3% of elements different from Reference vector.
* 2400 items are labelled 1, which have Variant vector 2%, 3% or 4% of elements different from Reference vector.

Testing samples: 600 items: 16384 inputs, one binary output (0 or 1)

File is test\_vectors2.csv

(7373250,)

new\_test.shape = (600, 16385)

* 300 items are labelled 0, which have Variant vector 0.1%, 0.2% or 0.3% of elements different from Reference vector.
* 300 items are labelled 1, which have Variant vector 2%, 3% or 4% of elements different from Reference vector.

**Result description:**

**ANN Model 512, 128,32 hidden units**

Model: "sequential\_4"

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Layer (type) Output Shape Param #

=================================================================

dense\_16 (Dense) (None, 512) 8389120

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

batch\_normalization\_12 (Batc (None, 512) 2048

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dropout\_8 (Dropout) (None, 512) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dense\_17 (Dense) (None, 128) 65664

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

batch\_normalization\_13 (Batc (None, 128) 512

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dropout\_9 (Dropout) (None, 128) 0

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dense\_18 (Dense) (None, 32) 4128

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

batch\_normalization\_14 (Batc (None, 32) 128

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

dense\_19 (Dense) (None, 1) 33

=================================================================

Total params: 8,461,633

Trainable params: 8,460,289

Non-trainable params: 1,344

**history = ANN\_model.fit(train\_x, train\_y, epochs=50, batch\_size=64,validation\_split=0.1,verbose = 2)**

Execution time: 87.8541247844696

19/19 - 0s - loss: 3.4850 - accuracy: 0.4783

Accuracy Score: 0.47833333333333333

error0to1 = 143 ; label is 0

error1to0 = 170 ; label is 1

total error = 313 percentError = 52.16666666666666

600 test cases: 300 label 0; 300 label 1

Even the training cases has 286 errors out of 4800:

Accuracy Score: 0.9404166666666667

error0to1 = 0 ; label is 0

error1to0 = 286 ; label is 1

total error = 286 percentError = 5.9583333333333335

train cases: 4800

**Notice: input values not scaled**

**np.max(train\_x) =** 0.01398 (which is 0.014 or 0.01 x 1.4)