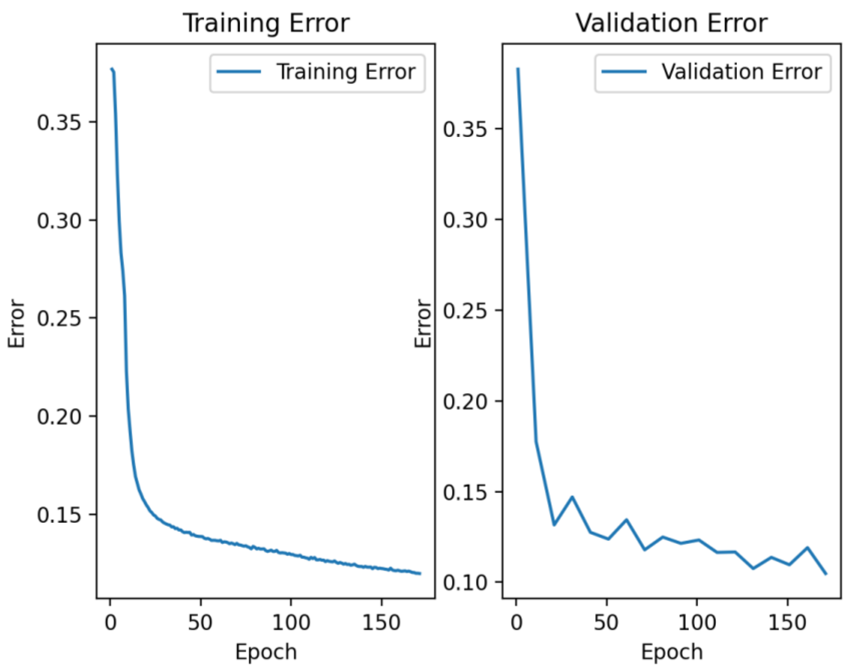
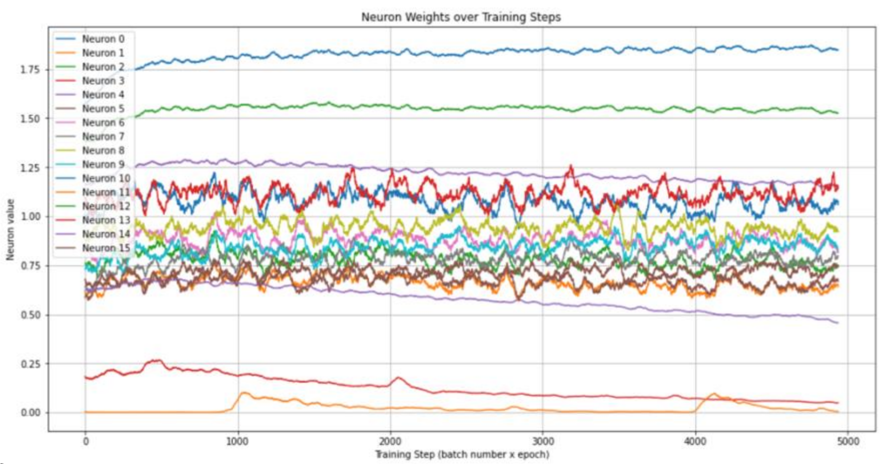
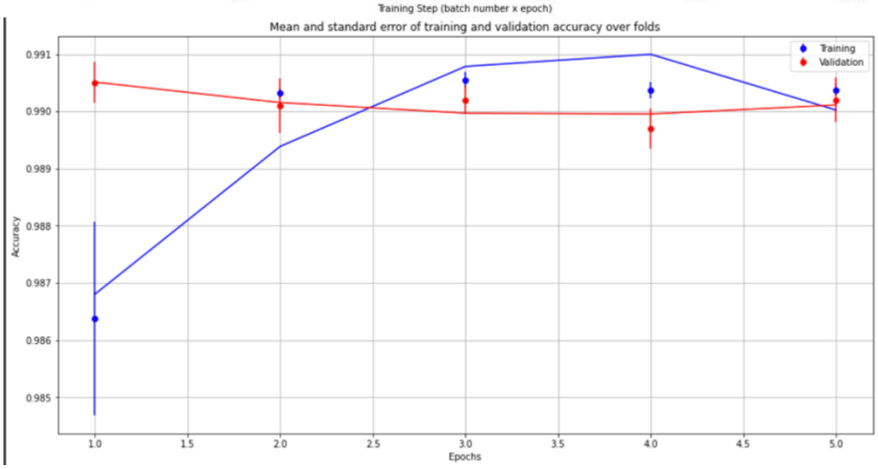

Training Error
Training Error
Error
Epoch
Validation Error
Validation Error
Error
Epoch

1.

2.


Neuron Weights over Training Steps
Neuron 0
Neuron 1
Neuron 2
Neuron 3
Neuron 4
Neuron 5
Neuron 6
Neuron 7
Neuron 8
Neuron 9
Neuron 10
Neuron 11
Neuron 12
Neuron 13
Neuron 14
Neuron 15
Neuron value
Training Step (batch number x epoch)


Mean and standard error of training and validation accuracy over folds
Training
Validation
Accuracy
Epochs

3.

$$J(z) = -\log\left(\frac{1}{1+exp\big(z(1-2y)\big)}\right) = \log\Big(1+exp\big(z(1-2y)\big)\Big)$$

$$\frac{\partial J}{\partial z} = \frac{1}{1+exp\big(z(1-2y)\big)} \cdot exp\big(z(1-2y)\big) \cdot (1-2y)$$

$$= (1-2y)\frac{e^{z(1-2y)}}{e^{z(1-2y)}+1}$$

i.

$$(1-2y)\frac{e^{z(1-2y)}}{e^{z(1-2y)}+1} \to -\frac{e^{-z}}{e^{-z}+1} \approx 0$$

ii.

$$(1-2y)\frac{e^{z(1-2y)}}{e^{z(1-2y)}+1} \to \frac{e^{z}}{e^{z}+1} \approx 1$$

iii.

$$(1-2y)\frac{e^{z(1-2y)}}{e^{z(1-2y)}+1} \to -\frac{e^{-z}}{e^{-z}+1} \approx 1$$

iv.

$$(1-2y)\frac{e^{z(1-2y)}}{e^{z(1-2y)}+1} \to \frac{e^{z}}{e^{z}+1} \approx 0$$

The derivative of the loss function vanishes when *z* is large and positive while *y=1* and when *z* is large and negative while *y=0*. This means that the vanishing gradient may occur for large values of *z* which could prevent the effectiveness of gradient decent techniques by stopping changes in model parameters. Since we know that $z = w^T h + b$, a possible countermeasure to this problem could be to enforce some regularization penalty to discourage large weights, keeping *z* small in magnitude.