Model Architecture

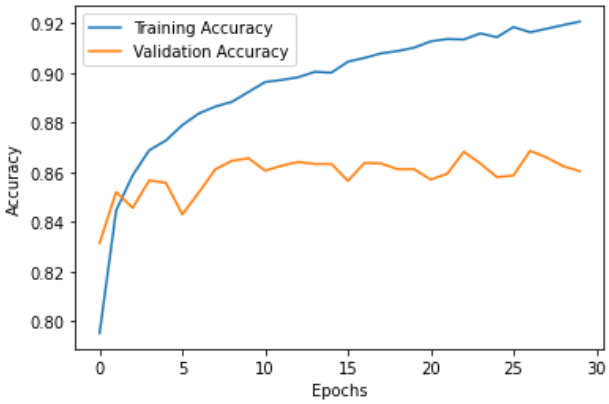
Model No.	Activation Function	Loss Function	No. of hidden layer	No. of nodes	Training Accuracy	Validation Accuracy	Total Runtime (s)	Total No. of parameters
1	ReLU,ReLU	Categorical Cross Entropy	2	64;32	0.9207	0.8604	112.59	52650
2	ReLU,Sigmoid	Categorical Cross Entropy	2	64;32	0.9226	0.8629	138.64	52650
3	ReLU,Sigmoid	KL divergence	2	64;32	0.92	0.8655	120.15	52650
4	ReLU,ReLU	KL divergence	2	128;64	0.935	0.8659	142.4	109386
5	ReLU,ReLU, ReLU	Categorical Cross Entropy	3	128;64;32	0.9288	0.863	111.5	111146
6	Sigmoid	KL divergence	1	256	0.9519	0.8907	187.305	2,03,530
7	Sigmoid,Sigmoid	KL divergence	2	128;64	0.9597	0.8831	209.59	109386
8	Sigmoid,Sigmoid	Categorical Cross Entropy	2	128;128	0.9631	0.8818	142.55	118282
9	ReLU,Sigmoid,Sigmoid	Categorical Cross Entropy	3	128;64;32	0.9399	0.8676	139.24	1,11,146
10	ReLU	Categorical Cross Entropy	1	256	0.9227	0.8677	164.13	1,01,770
11	Tanh	Categorical Cross Entropy	1	256	0.8734	0.8532	189.102	2,03,530
12	Tanh, Tanh	Categorical Cross Entropy	2	256;128	0.8941	0.848	234.038	2,35,146
13	Tanh,Tanh,Tanh	Categorical Cross Entropy	3	256;128;64	0.8919	0.8485	193.124	2,42,762
14	Tanh	KL divergence	1	128	0.8765	0.8553	165.301	1,01,770
15	Tanh, Tanh	KL divergence	2	128;64	0.8991	0.8542	164.101	1,09,386
16	Sigmoid, Sigmoid, Sigmoid, Sigmoid, Sigmoid	Categorical Cross Entropy	5	1028;512;256;128;64	0.9579	0.873	862.942	15,06,958

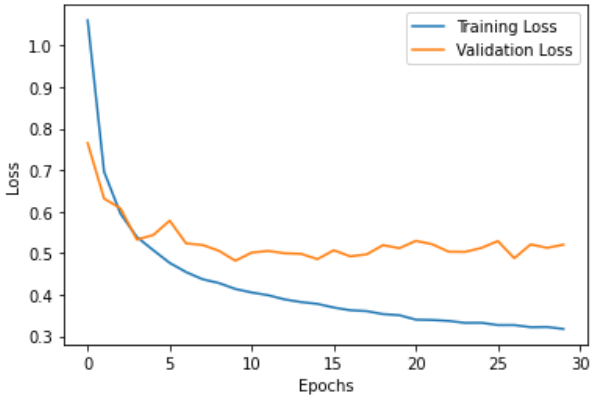
For every model :-

- L2 regularization of 0.01 for every hidden layer
- Dropout of 0.20 has been used after the last hidden layer
- Learning rate = 0.0005
- Batch size = 32
- Epochs = 30

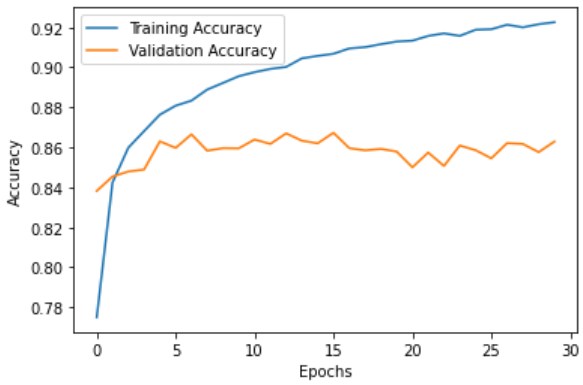
Accuracy vs Epochs and Loss vs Epochs

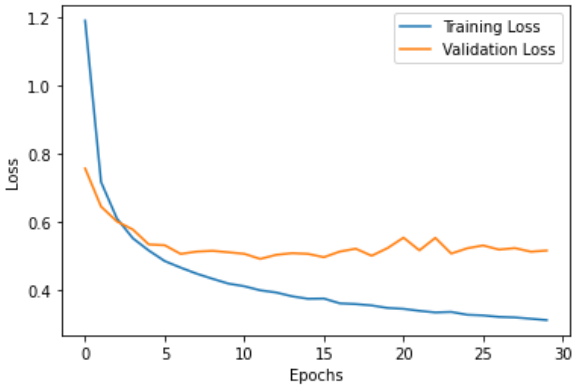
Model 1



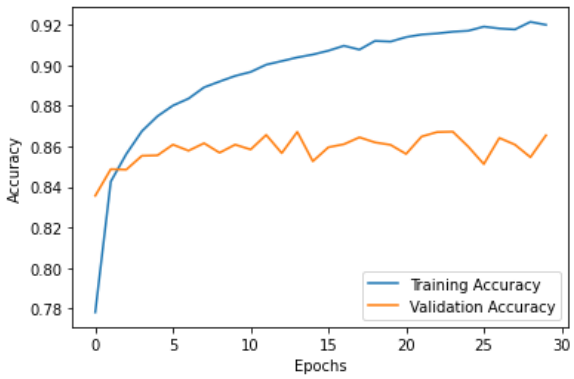


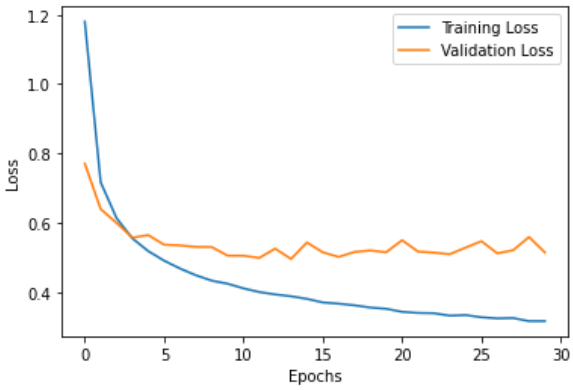
Model 2



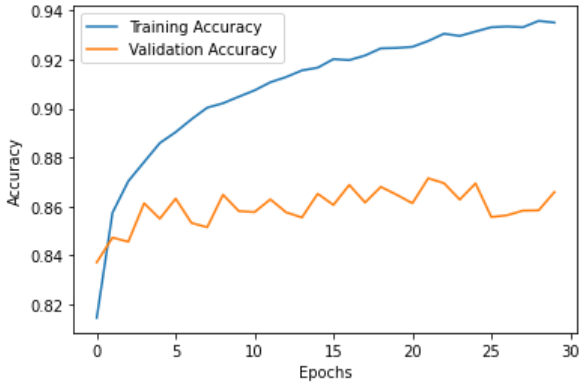


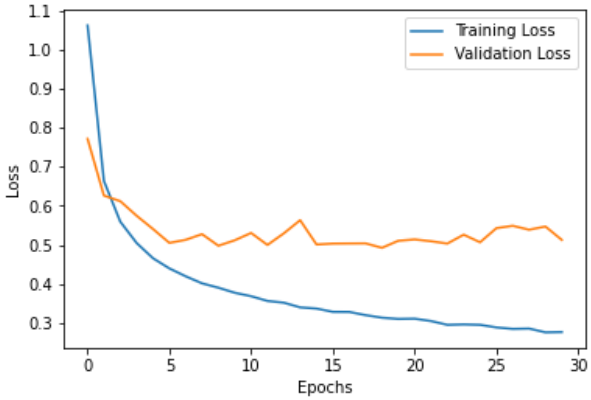
Model 3



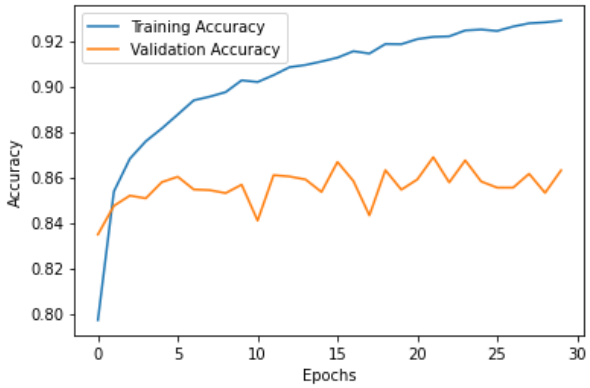


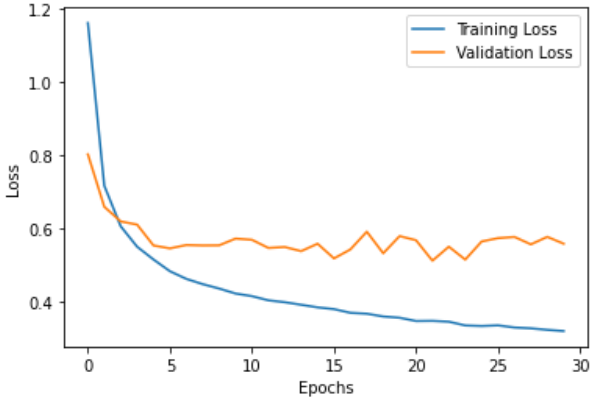
Model 4



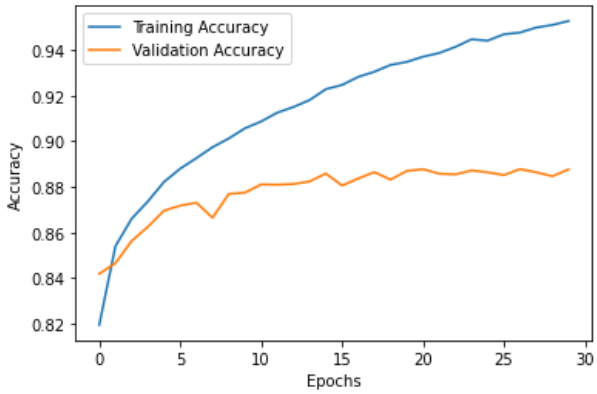


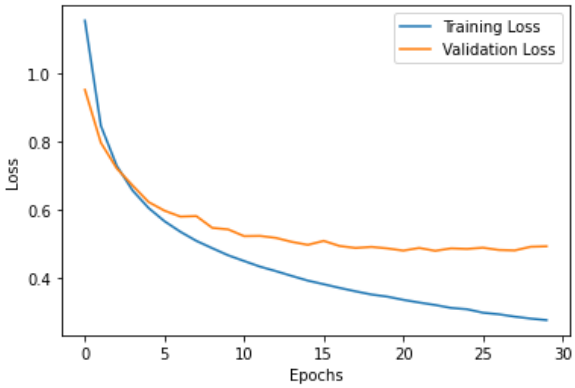
Model 5



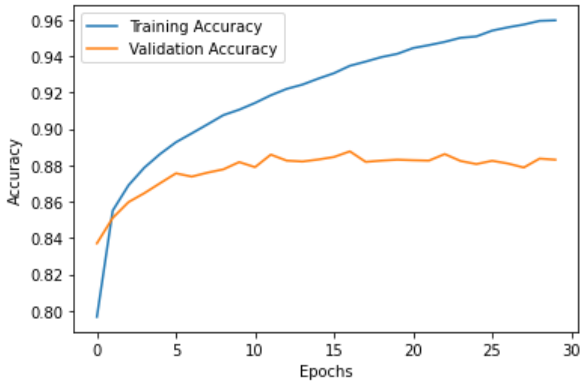


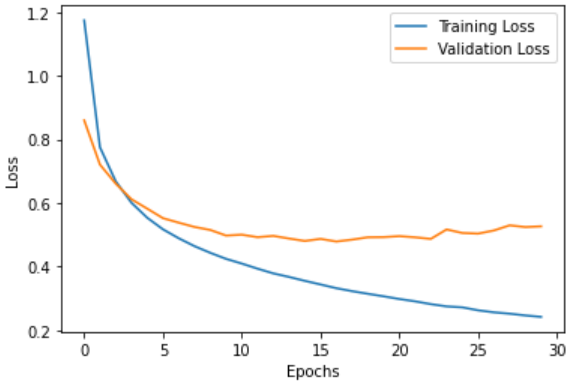
Model 6



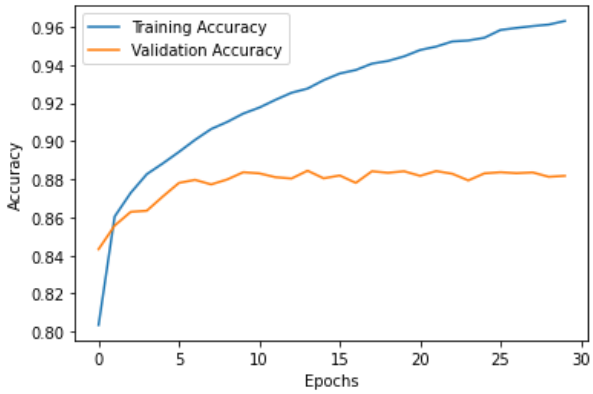


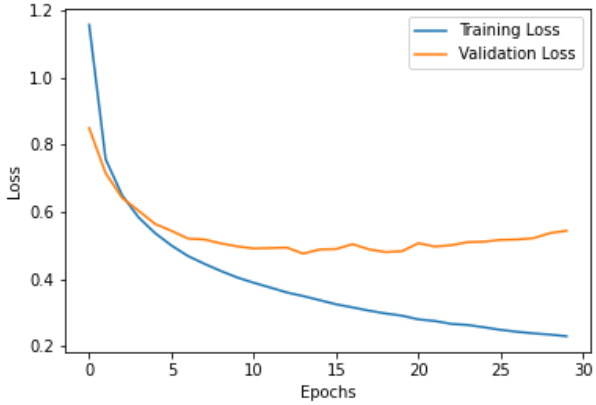
Model 7



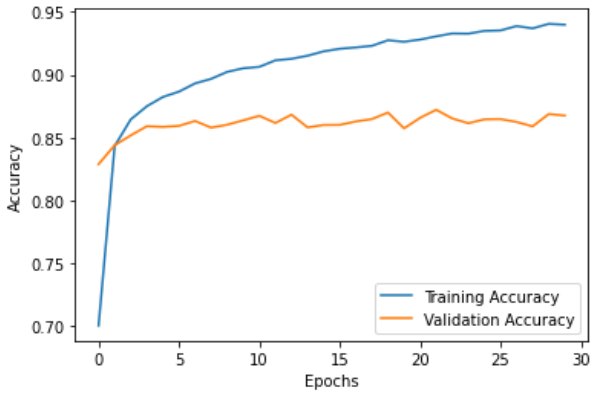


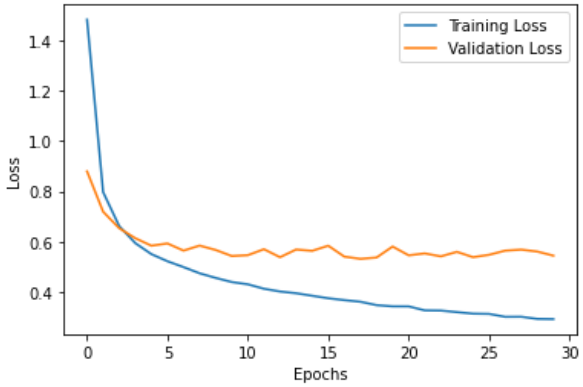
Model 8



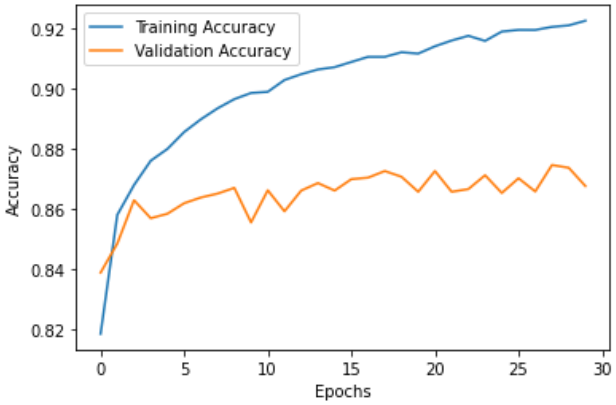


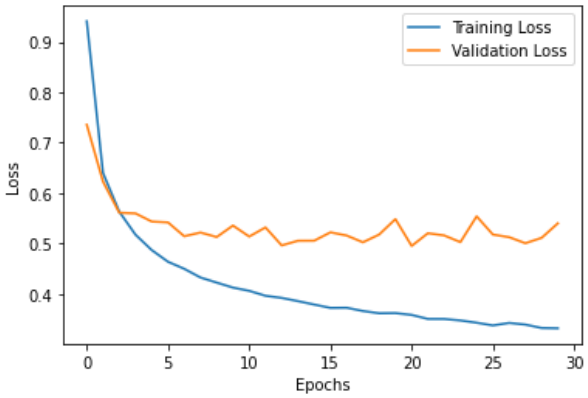
Model 9



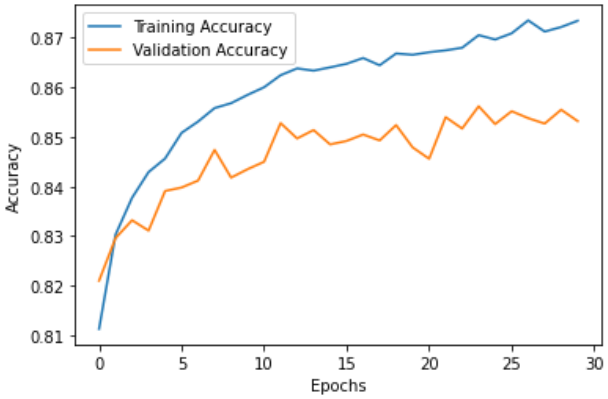


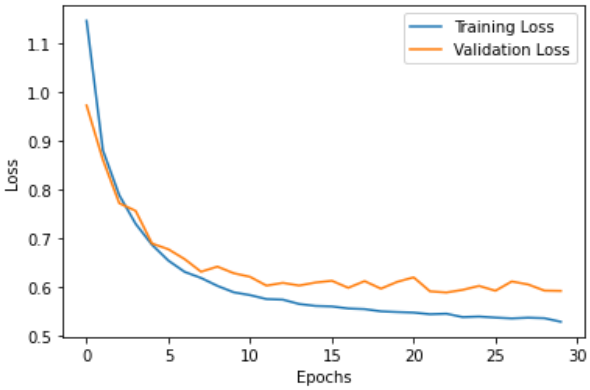
Model 10



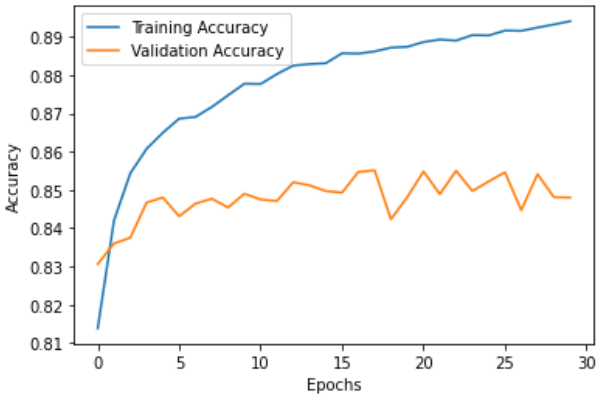


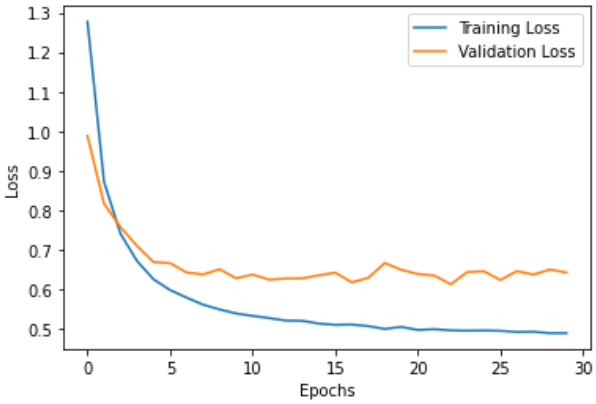
Model 11



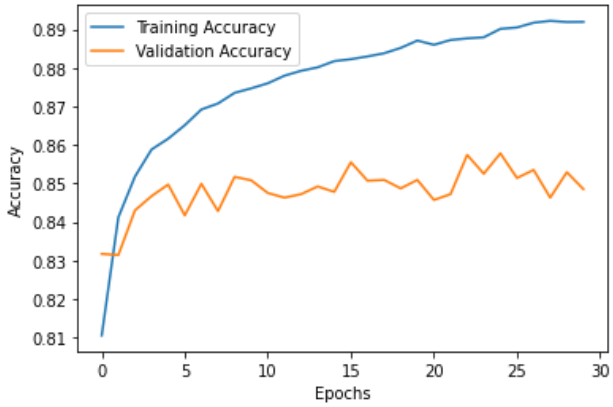


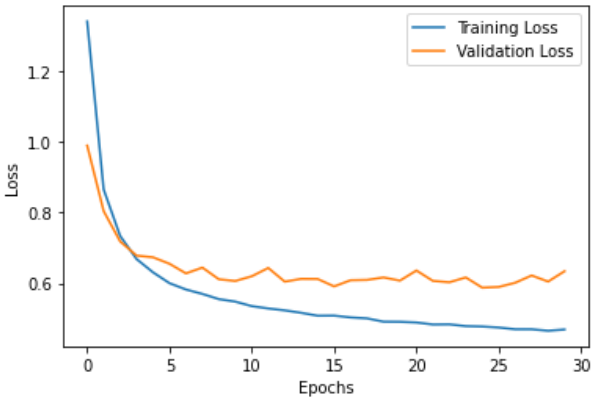
Model 12



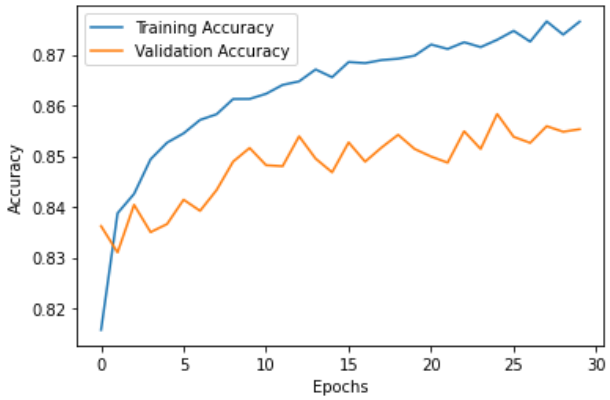


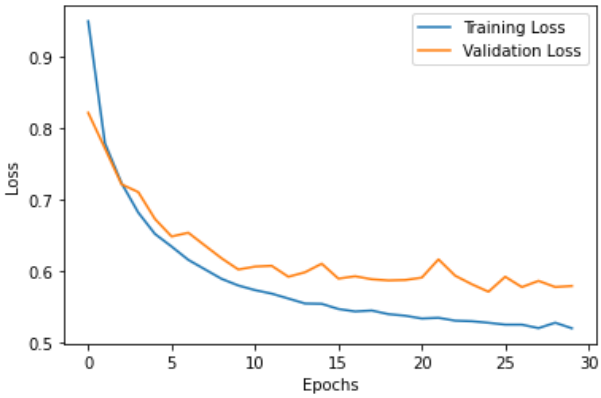
Model 13



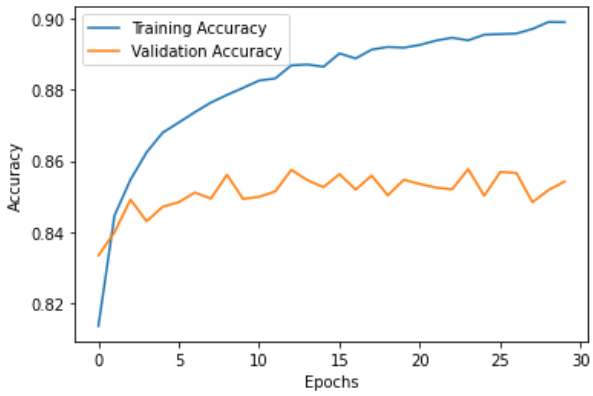


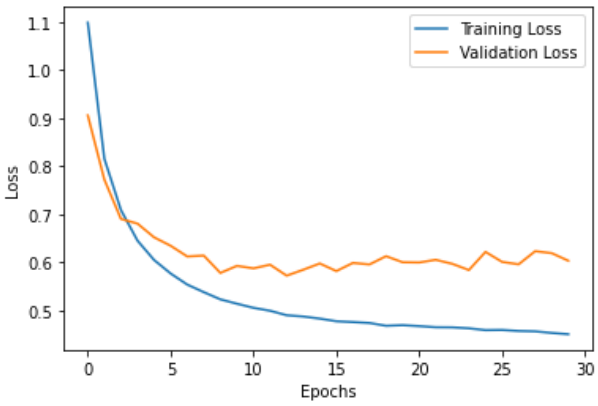
Model 14



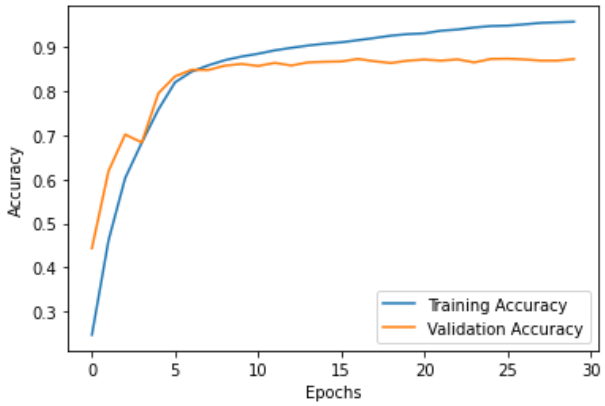


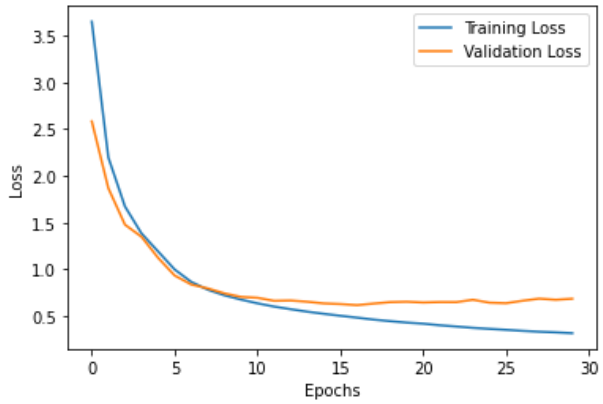
Model 15





Model 16





Comparative Study

Model No.	Activation Function	Loss Function	No. of hidden layer	No. of nodes	Training Accuracy	Validation Accuracy	Total Runtime (s)	Total No. of parameters
6	Sigmoid	KL divergence	1	256	0.9519	0.8907	187.305	2,03,530
7	Sigmoid,Sigmoid	KL divergence	2	128;64	0.9597	0.8831	209.59	109386
8	Sigmoid,Sigmoid	Categorical Cross Entropy	2	128;128	0.9631	0.8818	142.55	118282
16	Sigmoid, Sigmoid, Sigmoid, Sigmoid, Sigmoid	Categorical Cross Entropy	5	1028;512;256;128;64	0.9579	0.873	862.942	15,06,958
10	ReLU	Categorical Cross Entropy	1	256	0.9227	0.8677	164.13	1,01,770
9	ReLU,Sigmoid,Sigmoid	Categorical Cross Entropy	3	128;64;32	0.9399	0.8676	139.24	1,11,146
4	ReLU,ReLU	KL divergence	2	128;64	0.935	0.8659	142.4	109386
3	ReLU,Sigmoid	KL divergence	2	64;32	0.92	0.8655	120.15	52650
5	ReLU,ReLU, ReLU	Categorical Cross Entropy	3	128;64;32	0.9288	0.863	111.5	111146
2	ReLU,Sigmoid	Categorical Cross Entropy	2	64;32	0.9226	0.8629	138.64	52650
1	ReLU,ReLU	Categorical Cross Entropy	2	64;32	0.9207	0.8604	112.59	52650
14	Tanh	KL divergence	1	128	0.8765	0.8553	165.301	1,01,770
15	Tanh, Tanh	KL divergence	2	128;64	0.8991	0.8542	164.101	1,09,386
11	Tanh	Categorical Cross Entropy	1	256	0.8734	0.8532	189.102	2,03,530
13	Tanh,Tanh,Tanh	Categorical Cross Entropy	3	256;128;64	0.8919	0.8485	193.124	2,42,762
12	Tanh, Tanh	Categorical Cross Entropy	2	256;128	0.8941	0.848	234.038	2,35,146

Fig: Models sorted in decreasing order of validation accuracy

- The general trend that we can observe here is that sigmoid activation function works best followed by ReLU and then Tanh

- We can also say that increasing the number of hidden layers may not necessarily lead to better results (this is evident from how for a particular set of activation and loss function, increasing number of layer does not improve accuracy - Eg: Models 6&7, Models 14&15, Models 11&12 but increases accuracy in a few other instances - Eg: Models 12&13,)

    - Increasing the number of layers helps us learn more complex features, thus always increasing the training accuracy
    - Sometimes due to overfitting,i.e, due to the larger number of parameters, model fits too closely to the training data, thus losing its ability to generalize well and performing badly on the validation set (evident by lower validation accuracy in spite of higher training accuracy)

- Increasing the number of neurons in a hidden layer can lead to better results and can have a greater effect than changing activation function (evident from comparing Models 3&4 - Model 4 performed better even though in Model 3 sigmoid activation function has been used)
    - By increasing the number of neurons, we are able to learn more complex features. However activation functions only introduces non-linearity into the output and the choice of the activation function does not make a huge impact

- Another inference we can make is that KL divergence works better than Categorical Cross Entropy (Can be inferred by comparing Models 2&3)

- Increasing the number of hidden layers comprises the computational feasibility causing high runtimes. (We can see from Models 8&16 that this holds true.)