Testing our Model on Different Split_train and validation_train dataset:-

Adam Operator :-

Dataset	Accuracy	Hyperparameter	Batch size	Epochs
Dataset 1	94.55	Beta_1 = 0.87, Beta_2=0.991, Learning_rate=0.02, Epsilon = 1e-07	68	100
Dataset 2	94.23	Beta_1 = 0.88, Beta_2=0.992, Learning_rate=0.02, Epsilon = 1e-07	68	100
Dataset 3	93.93	Beta_1 = 0.89, Beta_2=0.993, Learning_rate=0.03, Epsilon = 1e-07	68	100
Dataset 4	93.5	Beta_1 = 0.9, Beta_2=0.999, Learning_rate=0.03, Epsilon = 1e-07	68	100
Dataset 5	92.87	Beta_1 = 0.91, Beta_2=0.999, Learning_rate=0.04, Epsilon = 1e-07	68	100
Dataset 6	92.27	Beta_1 = 0.92, Beta_2=0.999, Learning_rate=0.04, Epsilon = 1e-07	68	100
Dataset 7	91.73	Beta_1 = 0.93, Beta_2=0.991, Learning_rate=0.05, Epsilon = 1e-07	68	100
Dataset 8	91.66	Beta_1 = 0.9, Beta_2=0.999, Learning_rate=0.05, Epsilon = 1e-07	68	100
Dataset 9	91.99	Beta_1 = 0.92, Beta_2=0.999, Learning_rate=0.05, Epsilon = 1e-07	68	100
Dataset 10	89.1	Beta_1 = 0.87, Beta_2=0.991, Learning_rate=0.06, Epsilon = 1e-07	68	100

SGD Optimizer:-

Dataset	Accuracy	Hyperparameter	Batch size	Epochs
Dataset 1	73.4	Momentum=0,nesterov=False, Learning_rate=0.2	68	100
Dataset 2	79.03	Momentum=0,nesterov=False, Learning_rate=0.2	68	100
Dataset 3	77.34	Momentum=0,nesterov=False, Learning_rate=0.3	68	100
Dataset 4	79.39	Momentum=0,nesterov=False, Learning_rate=0.3	68	100
Dataset 5	80.33	Momentum=0,nesterov=False, Learning_rate=0.3	68	100
Dataset 6	77.47	Momentum=0,nesterov=False, Learning_rate=0.2	68	100
Dataset 7	79.7	Momentum=0,nesterov=False, Learning_rate=0.2	68	100
Dataset 8	75.72	Momentum=0,nesterov=False, Learning_rate=0.4	68	100
Dataset 9	85.006	Momentum=0,nesterov=False, Learning_rate=0.4	68	100
Dataset 10	85.29	Momentum=0,nesterov=False, Learning_rate=0.4	68	100

ADAGRAD Optimizer :-

Dataset	Accuracy	Hyperparameter	Batch size	Epochs
Dataset 1	69.095	Learning rate = 0.02,initial_accumulator =0.5	68	100
Dataset 2	62.844	Learning rate = 0.02,initial_accumulator =0.5	68	100
Dataset 3	63.13	Learning rate = 0.03,initial_accumulator =0.6	68	100
Dataset 4	60.36	Learning rate = 0.03,initial_accumulator =0.6	68	100
Dataset 5	68.48	Learning rate = 0.03,initial_accumulator =0.5	68	100
Dataset 6	73.75	Learning rate = 0.06,initial_accumulator =0.7	68	100
Dataset 7	69.98	Learning rate = 0.06,initial_accumulator =0.7	68	100

Dataset 8	67.44	Learning rate = 0.05,initial_accumulator =0.8	68	100
Dataset 9	70.01	Learning rate = 0.05,initial_accumulator =0.7	68	100
Dataset 10	80.29	Learning rate = 0.06,initial_accumulator =0.9	68	100

RMS prop optimizer :-

Dataset	Accuracy	Hyperparameter	Batch size	Epochs
Dataset 1	82.13	Learning_Rate=0.02, rho=0.2,momentum=0.9	68	100
Dataset 2	85.44	Learning_Rate=0.04, rho=0.3,momentum=0.9	68	100
Dataset 3	87.72	Learning_Rate=0.03, rho=0.2,momentum=0.9	68	100
Dataset 4	83.92	Learning_Rate=0.04, rho=0.2,momentum=0.9	68	100
Dataset 5	91.3	Learning_Rate=0.04, rho=0.2,momentum=0.9	68	100
Dataset 6	82.23	Learning_Rate=0.04, rho=0.2,momentum=0.9	68	100
Dataset 7	81.44	Learning_Rate=0.08, rho=0.2,momentum=0.9	68	100
Dataset 8	80.23	Learning_Rate=0.08, rho=0.2,momentum=0.9	68	100
Dataset 9	87.54	Learning_Rate=0.07, rho=0.2,momentum=0.9	68	100
Dataset 10	89.24	Learning_Rate=0.07, rho=0.2,momentum=0.9	68	100

As we can see our best model :-

For Adam Optimizer

Dataset 1	94.55	Beta_1 = 0.87, Beta_2=0.991, Learning_rate=0.02, Epsilon = 1e-07	68	100

`Checking the training accuracy - 95.48`
Checking the Test accuracy – 94.51

Choosing the best architecture with the above specification:-

Architecture	Test accuracy	Train accuracy	Hyper Parameter	Number Of layers
1	97.01	98.41	Beta_1 = 0.87, Beta_2=0.991, Learning_rate=0.02, Epsilon = 1e-07	Hidden Layer – 3 (128,64,32)
2	97.20	98.35	Beta_1 = 0.87, Beta_2=0.991, Learning_rate=0.02, Epsilon = 1e-07	Hidden Layer – 4 (128,64,64,32)
3	97.16	98.28	Beta_1 = 0.87, Beta_2=0.991, Learning_rate=0.02, Epsilon = 1e-07	Hidden Layer – 4 (128,128,64,32)