Findings Report

	f1: digit 0	f1: digit 1	f1: digit 2	f1: digit 3	f1: digit 4	f1: digit 5	f1: digit 6	f1: digit 7	f1: digit 8	f1: digit 9	overall accuracy
Model 1	0.987303	0.993843	0.97971	0.977228	0.982653	0.980942	0.988506	0.979055	0.976935	0.970574	0.9818
Model 2	0.986274	0.992508	0.978764	0.98419	0.983155	0.988182	0.985945	0.981132	0.980759	0.981692	0.9843
Model 3	**0.996438**	0.99561	**0.995658**	**0.995059**	**0.992857**	**0.992693**	**0.99319**	**0.993649**	**0.994872**	**0.992556**	**0.9943**
Model 4	0.994408	**0.996037**	0.991752	0.988653	0.990291	0.986471	0.989551	0.983559	0.988175	0.987103	0.9897

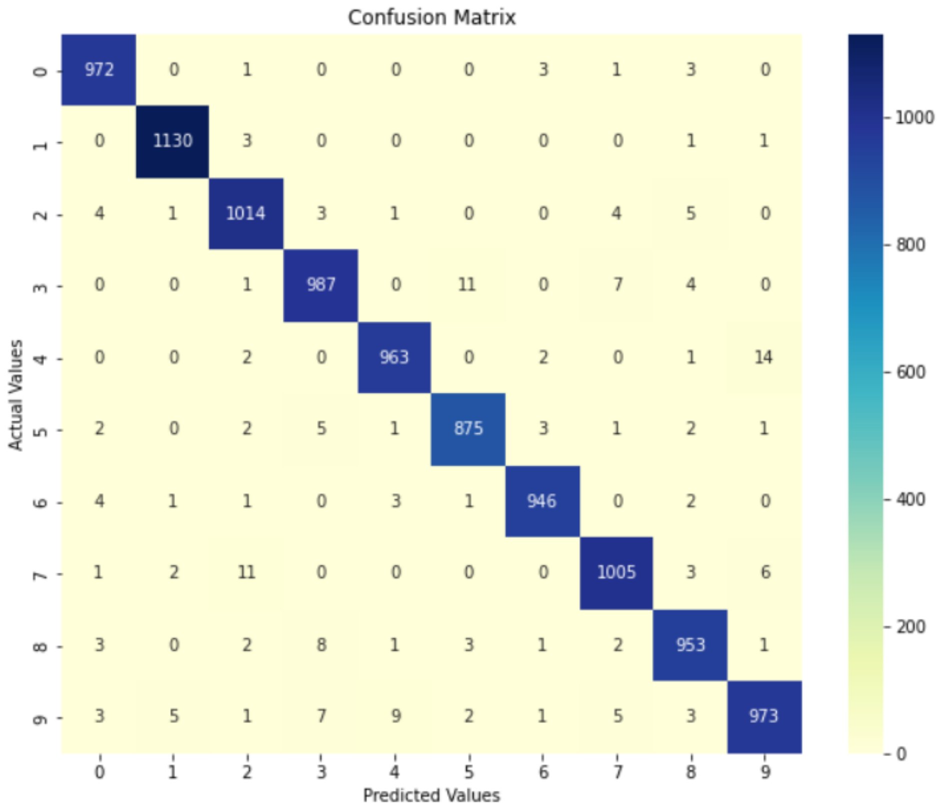
Overall, the stacked ensemble CNN model with global average pooling (model 3) performed the best during the test run, with a test set accuracy of 99.43%. This model was the only one to achieve an accuracy of more than 99% out of all 4 models.

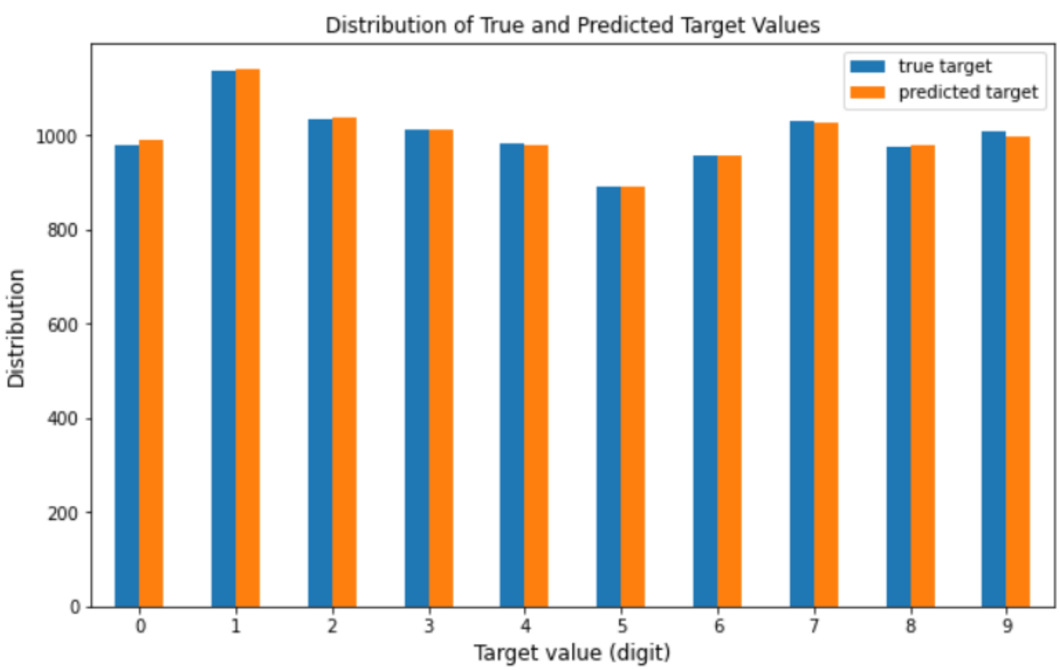
In terms of the performance in predicting the individual digits, model 3 had the best performance (highest F1 score) for all classes except the digit 1. Model 4 (stacked ensemble depthwise-separable CNN) slightly outperformed model 3 for digit 1. It should be noted that since the MNIST dataset is a fairly balanced dataset in terms of class distribution, the accuracy and F1-scores of models should be fairly similar as well. Thus, it is no surprise that the model with the highest individual F1-scores across the board also has the highest overall accuracy.

Before the test run, I predicted that either model 3 or model 4 will have the best performance, given their complexity compared to the first two models. I believe that one of the main reasons the depthwise-separable CNN model (model 4) failed to outperform its standard CNN counterpart is mainly due to the fact that MNIST image classification only uses one colour channel since all the images are in greyscale. In general, depthwise-separable CNN models are better suited for image classification problems that use all 3 colour channels. Another reason that model 3 outperformed model 4 was the incorporation of global average pooling for model 3. Experiments done in the past on the MNIST dataset have shown that global average pooling layers tends to outperform regular flattening layers.

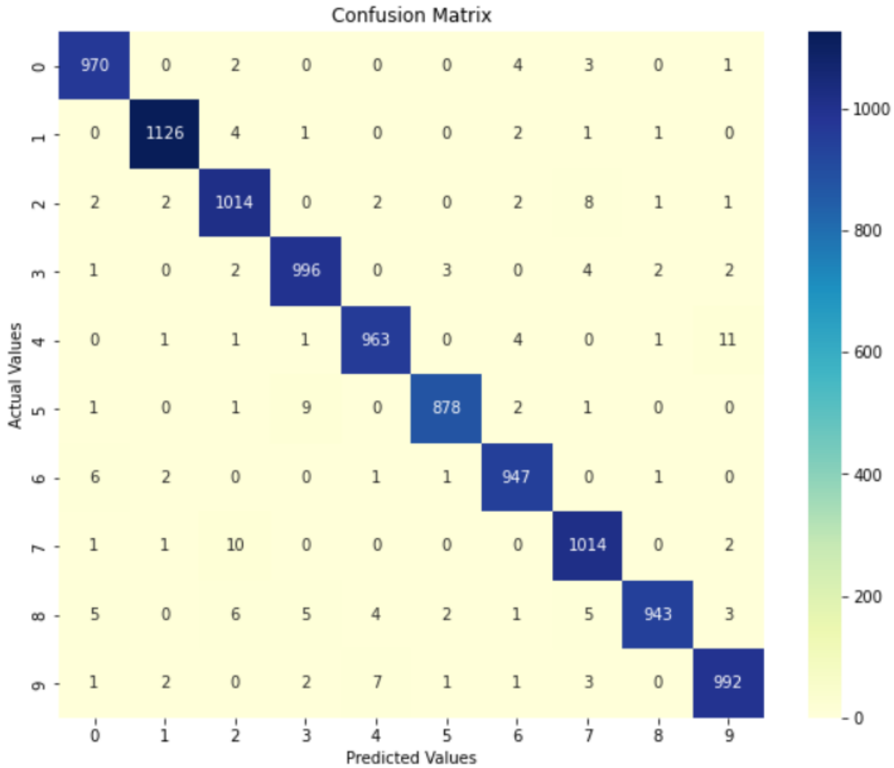
In the following pages you can find the confusion matrix as well as the true/predicted class distribution graphs for each of the 4 models.

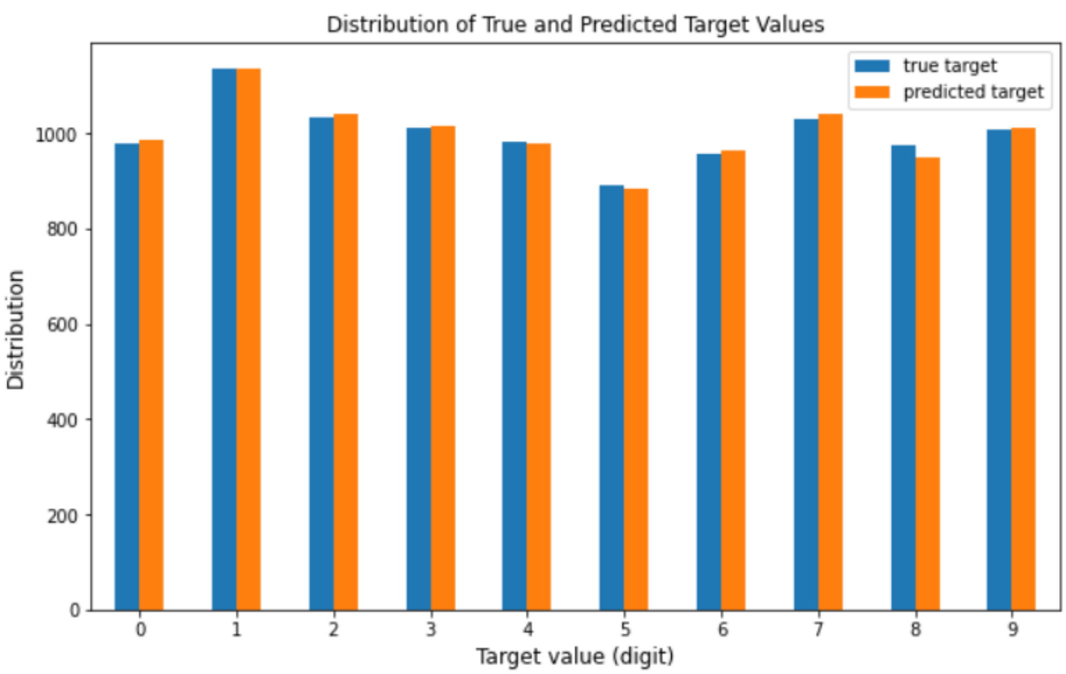
Model 1: Non-linear SVC with PCA



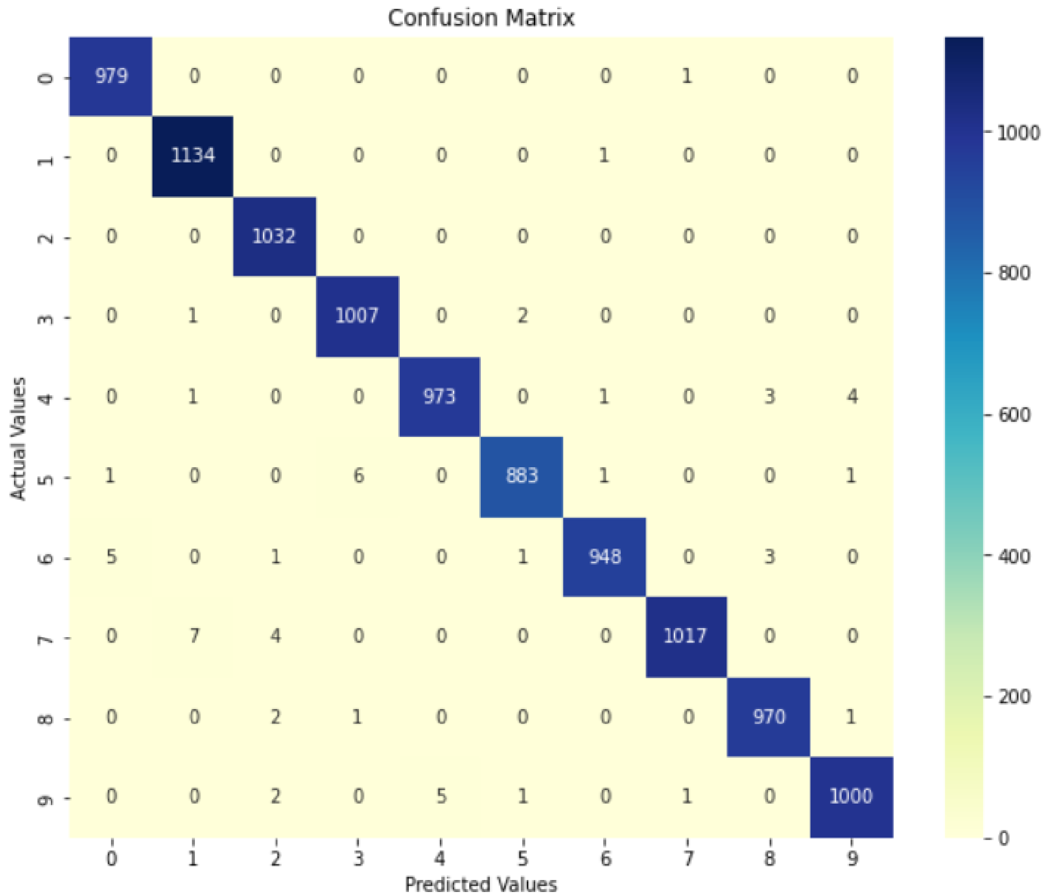


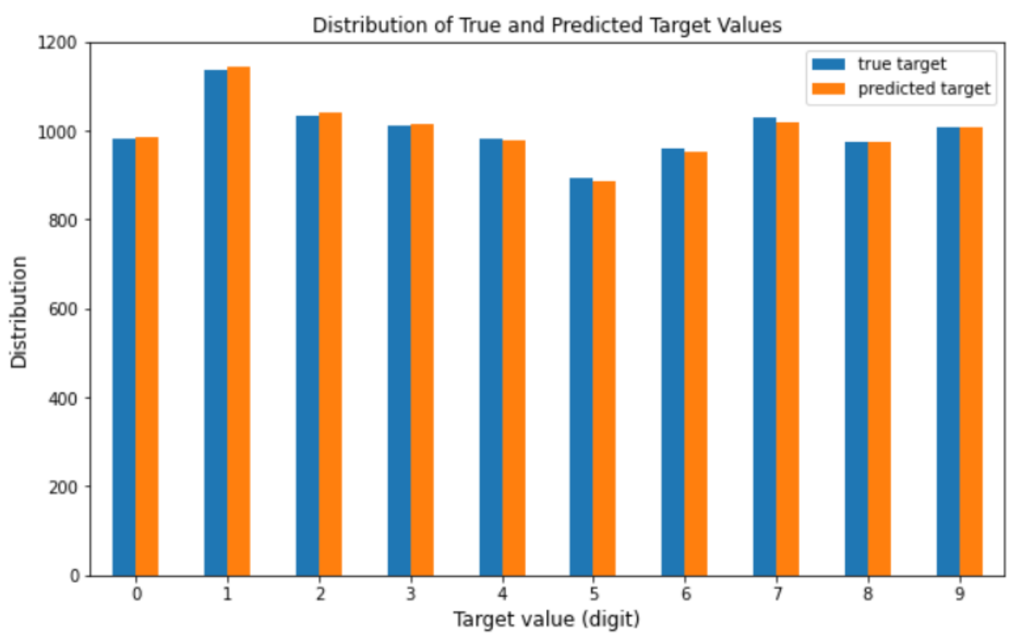
Model 2: Single Conv2D layer CNN





Model 3: Stacked ensemble CNN with global average pooling





Model 4: Stacked ensemble depthwise-separable CNN