Test results I presented in last class on NECMEC dataset.

	Below 9	9 - 13	14 - 17	Above 18
40 class model	0.46	0.640	0.543	0.863
40 class with class weights	0.59	0.720	0.624	0.878
4 class with class weights	0.32	0.613	0.322	0.797

I wrote a test program like CK's so our projects can be benchmarked against each other on Dan' dataset. When I tested my models with that script on NECMEC dataset I got different results (refer following table).

models	Metric	Below 9	9-13	14-17	Above 18
40 class, no class weights	Accuracy	0.378			
	Precision	0.13	0.099	0.532	0.426
	Recall	0.221	0.344	0.401	0.366
40 class with class weights	Accuracy	0.412			
	Precision	0.130	0.114	0.535	0.457
	Recall	0.150	0.344	0.460	0.381
4 class with class weights	Accuracy	0.313			
	Precision	0.082	0.083	0.513	0.451
	Recall	0.195	0.486	0.433	0.144

Reason:

1. After analyzing I figured that my previous code for calculating recall was wrong, hence the inflated results. This has been corrected in new program.
2. There are many mis-labeled images in NECMEC dataset (Every age group has images of babies, there are images of cars also and there are images in which there are no persons or objects). Hence overall test performance is low.
   To overcome these issues, I created a custom dataset and tested models on that. Following table has the results:

models	Metric	Below 9	9-13	14-17	Above 18
40 class, no class weights	Accuracy	0.589			
	Precision	0.562	0.584	0.395	0.743
	Recall	0.161	0.428	0.563	0.640
40 class with class weights	Accuracy	0.576			
	Precision	0.600	0.571	0.388	0.755
	Recall	0.161	0.433	0.606	0.597
4 class with class weights	Accuracy	0.512			
	Precision	0.373	0.365	0.402	0.840
	Recall	0.339	0.642	0.734	0.382

I would like to go with above results.

Question:

1. Should I provide Dan all models for testing or just one ("4 class with weights" because it performs better than others for below 18 and I can probably try to improve it for above 18 in next week)?
   Currently the model I provided is "40 class with no class weights"