gratitude: 14 Features Average Accuracy

	arousal	valence	happy	anger	sad	fear	surprise	disgust	control	fairness	self-related	other-related	expectedness	non-novelty
Subjective	0.321	0.880	0.961	0.005	0.040	0.053	0.271	0.008	0.372	0.704	0.654	0.926	0.516	0.604
500	0.964	0.964	0.964	0.965	0.963	0.964	0.964	0.965	0.965	0.964	0.966	0.964	0.962	0.961
1000	0.965	0.964	0.961	0.955	0.963	0.963	0.965	0.955	0.965	0.962	0.962	0.672	0.964	0.963
1500	0.962	0.961	0.946	0.960	0.965	0.963	0.964	0.964	0.964	0.951	0.963	0.958	0.965	0.963
2000	0.961	0.963	0.963	0.961	0.957	0.962	0.963	0.964	0.963	0.963	0.963	0.960	0.960	0.962
2500	0.955	0.964	0.962	0.926	0.956	0.964	0.961	0.960	0.962	0.963	0.963	0.958	0.960	0.959
3000	0.837	0.958	0.956	0.958	0.824	0.960	0.958	0.962	0.963	0.940	0.780	0.960	0.937	0.961
4000	0.770	0.952	0.959	0.960	0.951	0.910	0.944	0.943	0.958	0.955	0.958	0.961	0.960	0.954
5000	0.924	0.959	0.631	0.954	0.716	0.958	0.933	0.952	0.837	0.842	0.957	0.847	0.934	0.961
6000	0.856	0.641	0.759	0.942	0.938	0.934	0.954	0.892	0.844	0.939	0.950	0.879	0.935	0.661

Mask top k neurons