Single rater VS Multi rater					
Dataset	Sample Number	Ensemble	Bayesian	Probabilistic	Hierarchical Probabilistic
Heart T1	23	0.016	0.344	0.030	3.2e-05
Heart T2	23	0.104	0.393	0.300	4.7e-07
SIJ	69	0.001	0.0002	0.0001	0.008
Knee	10	0.625	0.083	0.845	0.013
Lung	268	7.4e-09	3.1e-08	1.6e-06	1.5e-23
Brain Growth	8	0.94	0.25	0.007	0.640
Kidney	5	0.062	0.812	0.062	0.062
Pancreas	12	0.733	1	0.0004	0.007
Prostate T1	11	0.519	0.764	0.320	0.0009
Prostate T2	11	0.898	0.041	0.018	0.0009
Brain Tumor T1	7	0.296	0.296	0.468	0.156

Table 1: The GED result of the Wilcoxon test results comparing SIN and MUL settings