Dataset	Single rater				Multi rater			
	Ensemble	Bayesian	Probabilistic	HP	Ensemble	Bayesian	Probabilistic	HP
Heart T1	0.018	0.02	0.021	0.065	0.019	0.019	0.016	0.049
Heart T2	0.026	0.023	0.022	0.197	0.025	0.024	0.025	0.118
SIJ	0.089	0.158	0.458	0.24	0.178	0.323	0.242	0.297
Knee	0.211	0.263	0.198	0.296	0.203	0.235	0.194	0.193
Lung	0.145	0.237	0.14	0.38	0.157	0.232	0.161	0.238
Brain Growth	0.105	0.112	0.502	0.103	0.091	0.107	0.1	0.095
Kidney	0.107	0.268	0.258	0.478	0.085	0.191	0.053	0.141
Pancreas	0.562	1.725	0.647	1.305	0.551	1.725	0.895	1.557
Prostate T1	0.039	0.037	0.042	0.046	0.041	0.051	0.035	0.48
Prostate T2	0.042	0.084	0.064	0.173	0.055	0.056	0.043	1.052
Brain Tumor T1	0.079	0.236	0.102	0.14	0.069	0.101	0.052	0.068
HP: Hierarchical Probabilistic								

Table 1: The GED scores for various models under both the MUL and SIN annotation settings.