Supplementary Table 1 Comparison between each single and ensemble model on GZ_Capcam (artifact-affected) datasets

Model	SegNet		U-Net		Attention-UNet		ResNet-UNet		HarDMSEG	
Metric	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
Snapshot 0	0.285	0.404	0.317	0.445	0.335	0.467	0.207	0.314	0.532	0.635
Snapshot 1	0.317	0.441	0.335	0.459	0.336	0.453	0.338	0.457	**0.538**	**0.648**
Snapshot 2	0.322	0.446	0.367	0.496	0.305	0.417	0.365	0.475	0.520	0.631
Snapshot 3	0.321	0.441	0.378	0.504	**0.361**	**0.502**	0.414	0.521	0.505	0.606
Snapshot 4	0.294	0.412	0.340	0.464	0.335	0.453	**0.414**	**0.522**	0.516	0.622
Snapshot 5	0.274	0.373	0.379	0.495	0.334	0.459	0.371	0.480	0.509	0.618
Snapshot 6	0.333	0.450	0.364	0.484	0.359	0.501	0.391	0.499	0.524	0.629
Snapshot 7	0.338	**0.454**	**0.379**	**0.507**	0.307	0.427	0.358	0.460	0.523	0.626
Snapshot 8	**0.341**	0.450	0.372	0.498	0.308	0.430	0.341	0.459	0.525	0.649
Snapshot 9	0.314	0.421	0.357	0.481	0.314	0.447	0.353	0.459	0.529	0.645
Snapshot 10	0.310	0.417	0.322	0.448	0.300	0.429	0.377	0.466	0.471	0.581
Snapshot 11	0.302	0.407	0.330	0.450	0.257	0.367	0.373	0.491	0.509	0.626
Snapshot 12	0.305	0.401	0.335	0.452	0.325	0.457	0.377	0.473	0.492	0.609
3-Ensemble	0.328	0.450	0.397	0.519	**0.375**	**0.509**	0.431	0.535	**0.550**	**0.653**
5-Ensemble	0.326	0.446	**0.406**	**0.532**	0.365	0.491	**0.435**	**0.537**	0.523	0.634
7-Ensemble	**0.341**	**0.462**	0.404	0.530	0.354	0.477	0.435	0.536	0.521	0.632

Rows labeled as snapshot indicate the performance of every single model of the corresponding base model on the test set. Rows labeled as n-Ensemble indicate the performance of each ensemble learning model of the corresponding base model, where 3-Ensemble, 5-Ensemble, and 7-Ensemble indicate the integration results of three, five, and seven single models, respectively. Bolded numbers denote the best performance of the single or ensemble model of the corresponding base learner.

Supplementary Table 2 Comparison between each single and ensemble model on CVC_Colon (clear) datasets

Model	SegNet		UNet		Attention-UNet		ResNet-UNet		HarDMSEG	
Metric	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
Snapshot 0	0.541	0.631	0.543	0.634	0.601	0.695	0.628	0.738	0.824	0.890
Snapshot 1	0.634	0.727	0.594	0.675	0.621	0.705	0.718	0.811	0.829	0.892
Snapshot 2	0.665	0.759	0.640	0.722	0.659	0.744	0.697	0.772	0.830	0.893
Snapshot 3	0.661	0.753	0.666	0.746	0.681	0.764	0.738	0.815	0.836	0.897
Snapshot 4	0.649	0.736	0.686	0.763	0.697	0.776	0.722	0.799	0.835	0.896
Snapshot 5	0.676	0.759	0.685	0.764	0.711	0.795	0.743	0.813	0.835	0.898
Snapshot 6	0.672	0.757	0.700	0.777	0.726	0.806	0.745	**0.819**	0.827	0.892
Snapshot 7	0.692	0.774	0.711	0.784	0.713	0.791	0.732	0.805	0.838	0.899
Snapshot 8	0.681	0.763	0.702	0.776	0.726	0.801	0.703	0.795	0.839	0.901
Snapshot 9	0.698	0.777	0.705	0.785	0.740	0.814	0.697	0.767	**0.840**	**0.901**
Snapshot 10	0.696	0.778	**0.713**	**0.788**	**0.754**	0.823	0.738	0.802	0.838	0.898
Snapshot 11	0.685	0.767	0.708	0.783	0.750	**0.830**	0.734	0.799	0.840	0.900
Snapshot 12	**0.700**	**0.780**	0.703	0.784	0.746	0.824	**0.747**	0.811	0.837	0.899
3-Ensemble	0.700	0.778	**0.711**	**0.783**	**0.752**	**0.829**	0.746	0.815	**0.840**	**0.901**
5-Ensemble	**0.702**	**0.779**	0.710	0.781	0.750	0.822	0.746	0.816	0.840	0.901
7-Ensemble	0.701	0.778	0.709	0.783	0.744	0.818	**0.750**	**0.816**	0.840	0.901

Rows labeled as snapshot indicate the performance of every single model of the corresponding base model on the test set. Rows labeled as n-Ensemble indicate the performance of each ensemble learning model of the corresponding base model, where 3-Ensemble, 5-Ensemble, and 7-Ensemble indicate the integration results of three, five, and seven single models, respectively. Bolded numbers denote the best performance of the single or ensemble model of the corresponding base learner.

Supplementary Table 3 Comparison between each single and ensemble model on CVC_Clinic (clear) datasets

Model	SegNet		UNet		Attention-UNet		ResNet-UNet		HarDMSEG	
Metric	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
Snapshot 0	0.701	0.794	0.742	0.825	0.762	0.848	0.789	0.862	0.838	0.890
Snapshot 1	0.766	0.852	0.778	0.857	0.802	0.877	0.808	0.877	0.844	0.893
Snapshot 2	0.790	0.871	0.794	0.864	0.821	0.889	0.789	0.858	0.842	0.891
Snapshot 3	0.799	0.876	0.806	0.876	0.831	0.894	0.815	0.880	0.842	0.891
Snapshot 4	0.795	0.873	0.810	0.880	0.825	0.889	0.815	0.878	0.841	0.890
Snapshot 5	0.807	0.883	0.811	0.881	0.834	0.897	0.817	0.882	0.843	0.891
Snapshot 6	0.801	0.880	**0.816**	**0.884**	**0.836**	0.896	0.823	0.883	0.839	0.888
Snapshot 7	0.808	0.884	0.807	0.879	0.830	0.893	0.820	0.881	**0.845**	**0.892**
Snapshot 8	0.809	0.885	0.815	0.884	0.835	**0.900**	0.818	0.880	0.840	0.887
Snapshot 9	0.812	0.888	0.808	0.879	0.831	0.897	**0.823**	**0.884**	0.844	0.891
Snapshot 10	0.809	0.886	0.807	0.877	0.833	0.897	0.818	0.880	0.843	0.890
Snapshot 11	0.813	0.890	0.807	0.877	0.832	0.895	0.818	0.878	0.841	0.890
Snapshot 12	**0.814**	**0.890**	0.805	0.876	0.830	0.894	0.818	0.879	0.842	0.891
3-Ensemble	0.813	0.889	0.818	0.885	0.835	0.897	**0.824**	**0.884**	0.844	**0.892**
5-Ensemble	0.813	0.888	0.826	0.891	0.835	0.897	0.824	0.884	0.844	0.892
7-Ensemble	**0.815**	**0.890**	**0.826**	**0.891**	**0.838**	**0.898**	0.823	0.883	**0.844**	0.892

Rows labeled as snapshot indicate the performance of every single model of the corresponding base model on the test set. Rows labeled as n-Ensemble indicate the performance of each ensemble learning model of the corresponding base model, where 3-Ensemble, 5-Ensemble, and 7-Ensemble indicate the integration results of three, five, and seven single models, respectively. Bolded numbers denote the best performance of the single or ensemble model of the corresponding base learner.