Appendix A   CSL Results Tables

Table A1: Model performance comparison for 10K, 20K, 50K and 100K sample sizes with IR weighting

Metric	LR	RF	XGB	LGBM	NB	TabNet	MLP	RF+TabNet	Stacking
10k Sample Size									
Test Accuracy	0.7675	0.8390	0.8740	0.8375	0.6660	0.8287	0.8453	0.9027	0.8685
AUC-ROC	0.8593	0.8532	0.8118	0.8410	0.7385	0.8114	0.8281	0.777227	0.4641
F1-Score (L0)	0.8571	0.9076	0.9302	0.9062	0.7838	0.9013	0.9115	0.948592	0.9274
F1-Score (L1)[1]	0.3775	0.3735	0.3571	0.3925	0.2659	0.3494	0.3862	0.0875	0.3024
Recall (L0)	0.763	0.8659	0.9184	0.8593	0.6628	0.8658	0.8813	0.993363	0.9195
Recall (L1)[1]	0.8150	0.5549	0.4046	0.6069	0.6994	0.4792	0.5069	0.048611	0.3295
20k Sample Size									
Test Accuracy	0.7570	0.8058	0.8433	0.7920	0.7095	0.8107	0.8877	0.911667	0.9048
AUC-ROC	0.8602	0.8512	0.8056	0.8437	0.7788	0.8247	0.7946	0.813989	0.5747
F1-Score (L0)	0.8497	0.8847	0.9110	0.8752	0.8164	0.8902	0.9390	0.95376	0.9498
F1-Score (L1)[1]	0.3655	0.3848	0.3435	0.3754	0.3042	0.3140	0.2935	0.01487	0.0685
Recall (L0)	0.7531	0.8167	0.8794	0.7997	0.7081	0.8417	0.9481	0.999269	0.9879
Recall (L1)[1]	0.7977	0.6923	0.4672	0.7123	0.7236	0.4906	0.2642	0.007547	0.0399
50k Sample Size									
Test Accuracy	0.7592	0.7844	0.8143	0.7709	0.7303	0.7845	0.9049	0.9104	0.9020
AUC-ROC	0.8469	0.8368	0.8065	0.8432	0.7780	0.8297	0.8017	0.833998	0.6443
F1-Score (L0)	0.8518	0.8700	0.8918	0.8600	0.8325	0.8716	0.9494	0.952769	0.9483
F1-Score (L1)[1]	0.3582	0.3692	0.3468	0.3690	0.3083	0.3306	0.2258	0.129534	0.0684
Recall (L0)	0.7584	0.7905	0.8384	0.7714	0.7345	0.8025	0.9778	0.991661	0.9845
Recall (L1)[1]	0.7680	0.7211	0.5634	0.7657	0.6869	0.6000	0.1564	0.075188	0.0411
100k Sample Size									
Test Accuracy	0.7489	0.7589	0.7823	0.7446	0.7681	0.7381	0.9145	0.917667	0.9128
AUC-ROC	0.8415	0.8359	0.8221	0.8440	0.7767	0.8356	0.8125	0.839335	0.7000
F1-Score (L0)	0.8442	0.8516	0.8691	0.8410	0.8602	0.8368	0.9550	0.956799	0.9542
F1-Score (L1)[1]	0.3540	0.3577	0.3530	0.3513	0.3185	0.3376	0.1395	0.125973	0.0802
Recall (L0)	0.7458	0.7585	0.7927	0.7407	0.7828	0.2145	0.9910	0.995632	0.9968
Recall (L1)[1]	0.7814	0.7626	0.6746	0.7853	0.6156	0.7919	0.0823	0.070411	0.0432

[1]Minority class (Level 1: heart condition) metric.

Table A2: Model performance comparison for 10K, 20K, 50K and 100K sample sizes with Custom weighting

Metric	LR	RF	XGB	LGBM	NB	TabNet	MLP	RF+TabNet	Stacking
10k Sample Size									
Test Accuracy	0.4190	0.4625	0.7995	0.7210	0.6220	0.8287	0.8040	0.902667	0.8685
AUC-ROC	0.8559	0.8127	0.7871	0.8435	0.7383	0.8114	0.8286	0.777227	0.7851
F1-Score (L0)	0.5345	0.5851	0.8825	0.8231	0.7470	0.9013	0.8837	0.948592	0.9274
F1-Score (L1)[1]	0.2274	0.2370	0.3169	0.3404	0.2530	0.3494	0.3771	0.0875	0.3024
Recall (L0)	0.3651	0.4149	0.8243	0.7105	0.6108	0.8658	0.8813	0.993363	0.9195
Recall (L1)[1]	0.9884	0.9653	0.5376	0.8324	0.7399	0.4792	0.5069	0.048611	0.3295
20k Sample Size									
Test Accuracy	0.7568	0.8075	0.8433	0.9115	0.7095	0.8107	0.8977	0.911667	0.9050
AUC-ROC	0.8602	0.8521	0.8056	0.8527	0.7788	0.8247	0.7929	0.813989	0.5727
F1-Score (L0)	0.8495	0.8860	0.9110	0.9537	0.8164	0.8902	0.9449	0.95376	0.9499
F1-Score (L1)[1]	0.3653	0.3830	0.3435	0.0000	0.3042	0.3140	0.2776	0.01487	0.0821
Recall (L0)	0.7528	0.8197	0.8794	0.9992	0.7081	0.8417	0.9631	0.999269	0.9874
Recall (L1)[1]	0.7977	0.6809	0.4672	0.0000	0.7236	0.4906	0.2226	0.007547	0.0484
50k Sample Size									
Test Accuracy	0.7592	0.7850	0.8143	0.9131	0.7303	0.7845	0.9028	0.9104	0.9016
AUC-ROC	0.8469	0.8365	0.8065	0.8426	0.7780	0.8297	0.8161	0.833998	0.6438
F1-Score (L0)	0.8518	0.8705	0.8918	0.9545	0.8325	0.8716	0.9481	0.952769	0.9481
F1-Score (L1)[1]	0.3582	0.3680	0.3468	0.0159	0.3083	0.3306	0.2430	0.129534	0.0611
Recall (L0)	0.7584	0.7917	0.8384	0.9999	0.7345	0.8025	0.9735	0.991661	0.9845
Recall (L1)[1]	0.7680	0.7154	0.5634	0.0080	0.6869	0.6000	0.1759	0.46729	0.0366
100k Sample Size									
Test Accuracy	0.7488	0.7603	0.7823	0.9124	0.7681	0.7381	0.9140	0.917667	0.9119
AUC-ROC	0.8415	0.8361	0.8221	0.8410	0.7767	0.8356	0.8178	0.839335	0.6980
F1-Score (L0)	0.8441	0.8526	0.8691	0.9542	0.8602	0.8368	0.9546	0.956799	0.9538
F1-Score (L1)[1]	0.3539	0.3587	0.3530	0.0102	0.3185	0.3376	0.1783	0.125973	0.0638
Recall (L0)	0.7456	0.7601	0.7927	1.0000	0.7828	0.7332	0.9879	0.995632	0.9967
Recall (L1)[1]	0.7814	0.7615	0.6746	0.0051	0.6156	0.7919	0.1108	0.070411	0.0341

---

[1] Minority class (Level 1: heart condition) metric.

Table A3: Model performance comparison for 10K, 20K, 50K and 100K sample sizes with ICF weighting

Metric	LR	RF	XGB	LGBM	NB	TabNet	MLP	RF+TabNet	Stacking
10k Sample Size									
Test Accuracy	0.4190	0.4625	0.7995	0.7210	0.6220	0.7173	0.0960	0.902667	0.7935
AUC-ROC	0.8559	0.8127	0.7871	0.8435	0.7383	0.8294	0.8238	0.777227	0.7840
F1-Score (L0)	0.5345	0.5851	0.8825	0.8231	0.7470	0.8191	0.0000	0.948592	0.8786
F1-Score (L1)[1]	0.2274	0.2370	0.3169	0.3404	0.2530	0.3537	0.1752	0.0875	0.3105
Recall (L0)	0.3651	0.4149	0.8243	0.7105	0.6108	0.7080	0.0000	0.993363	0.8177
Recall (L1)[1]	0.9884	0.9653	0.5376	0.8324	0.7399	0.8056	1.0000	0.048611	0.5376
20k Sample Size									
Test Accuracy	0.3978	0.4218	0.7303	0.6313	0.6675	0.6533	0.8897	0.911667	0.7368
AUC-ROC	0.8545	0.8162	0.7715	0.8360	0.7786	0.8208	0.7862	0.813989	0.7570
F1-Score (L0)	0.5083	0.5375	0.8326	0.7498	0.7828	0.7696	0.9406	0.95376	0.8380
F1-Score (L1)[1]	0.2232	0.2287	0.3052	0.2993	0.2910	0.3001	0.2175	0.01487	0.2985
Recall (L0)	0.3412	0.3683	0.7355	0.6056	0.6569	0.6351	0.9590	0.999269	0.7368
Recall (L1)[1]	0.9858	0.9772	0.6752	0.8974	0.7778	0.8415	0.1736	0.007547	0.6382
50k Sample Size									
Test Accuracy	0.7585	0.7818	0.6282	0.5374	0.7303	0.7881	0.9061	0.9104	0.9024
AUC-ROC	0.8469	0.8353	0.7757	0.8308	0.7780	0.8278	0.8078	0.833998	0.6364
F1-Score (L0)	0.8513	0.8683	0.7504	0.6633	0.8325	0.8739	0.8078	0.952769	0.9485
F1-Score (L1)[1]	0.3575	0.3638	0.2713	0.2613	0.3083	0.3371	0.2143	0.129534	0.0740
Recall (L0)	0.7576	0.7884	0.6126	0.4993	0.7345	0.8057	0.9802	0.991661	0.9847
Recall (L1)[1]	0.7680	0.7131	0.7909	0.9349	0.6869	0.6075	0.1444	0.075188	0.0446
100k Sample Size									
Test Accuracy	0.3213	0.3080	0.5472	0.4630	0.6939	0.3429	0.9168	0.917667	0.5514
AUC-ROC	0.8407	0.8118	0.7979	0.8327	0.7764	0.8286	0.8186	0.839335	0.7260
F1-Score (L0)	0.4084	0.3898	0.6747	0.5849	0.8044	0.4418	0.9562	0.956799	0.6781
F1-Score (L1)[1]	0.2041	0.2008	0.2555	0.2395	0.2966	0.2013	0.1669	0.125973	0.2602
Recall (L0)	0.2569	0.2424	0.5148	0.4149	0.6901	0.2839	0.9921	0.995632	0.5181
Recall (L1)[1]	0.9881	0.9875	0.8825	0.9603	0.7331	0.9826	0.0989	0.070411	0.8961

[1]Minority class (Level 1: heart condition) metric.

Table A4: Model performance comparison for 10K, 20K, 50K and 100K sample sizes with LICF weighting

Metric	LR	RF	XGB	LGBM	NB	TabNet	MLP	RF+TabNet	Stacking
10k Sample Size									
Test Accuracy	0.6200	0.7005	0.8375	0.7685	0.6490	0.8101	0.6107	0.902667	0.8460
AUC-ROC	0.8578	0.8381	0.7763	0.8259	0.7384	0.8016	0.8232	0.777227	0.8048
F1-Score (L0)	0.7404	0.8067	0.9080	0.8597	0.7697	0.8888	0.8311	0.948592	0.9128
F1-Score (L1)[1]	0.2910	0.3352	0.3041	0.3395	0.2626	0.3632	0.3113	0.0875	0.3419
Recall (L0)	0.5933	0.6842	0.8779	0.7761	0.6420	0.8370	0.5782	0.993363	0.8823
Recall (L1)[1]	0.9017	0.8728	0.4104	0.6879	0.7225	0.5625	0.9167	0.048611	0.4624
20k Sample Size									
Test Accuracy	0.6060	0.6490	0.7865	0.7220	0.6938	0.7760	0.8903	0.911667	0.9123
AUC-ROC	0.8584	0.8410	0.7998	0.8393	0.7787	0.8232	0.7979	0.813989	0.7872
F1-Score (L0)	0.7271	0.7644	0.8731	0.8238	0.8040	0.8653	0.9410	0.95376	0.9540
F1-Score (L1)[1]	0.2914	0.3118	0.3286	0.3412	0.3004	0.3347	0.2148	0.01487	0.0436
Recall (L0)	0.5755	0.6243	0.8049	0.7125	0.6884	0.7894	0.9601	0.999269	0.9978
Recall (L1)[1]	0.9231	0.9060	0.5954	0.8205	0.7493	0.6378	0.1698	0.007547	0.0228
50k Sample Size									
Test Accuracy	0.7923	0.8224	0.7348	0.6570	0.7303	0.8001	0.9021	0.9104	0.9132
AUC-ROC	0.8468	0.8378	0.7968	0.8402	0.7780	0.8315	0.7995	0.833998	0.6387
F1-Score (L0)	0.8753	0.8961	0.8353	0.7717	0.8325	0.8824	0.9478	0.952769	0.9545
F1-Score (L1)[1]	0.3794	0.3889	0.3200	0.3107	0.3083	0.3358	0.2175	0.129534	0.0667
Recall (L0)	0.7987	0.8393	0.7369	0.6353	0.7345	0.8225	0.9750	0.991661	0.9974
Recall (L1)[1]	0.7257	0.6457	0.7131	0.7131	0.6869	0.5699	0.1534	0.075188	0.0354
100k Sample Size									
Test Accuracy	0.5828	0.5661	0.6779	0.6153	0.7199	0.6193	0.9143	0.917667	0.9070
AUC-ROC	0.8415	0.8299	0.8145	0.8420	0.7765	0.8341	0.8157	0.839335	0.7129
F1-Score (L0)	0.7063	0.6908	0.7899	0.7356	0.8246	0.7405	0.9549	0.956799	0.9510
F1-Score (L1)[1]	0.2795	0.2724	0.3097	0.2938	0.3041	0.2857	0.1472	0.125973	0.0403
Recall (L0)	0.5502	0.5316	0.6641	0.5869	0.7222	0.5931	0.9903	0.995632	0.9906
Recall (L1)[1]	0.9194	0.9228	0.8206	0.9091	0.6951	0.9035	0.0878	0.070411	0.0403

---

[1]Minority class (Level 1: heart condition) metric.