Personalized Federated Learning for Predicting Disability Progression in Multiple Sclerosis Using Real-World Routine Clinical Data

1 Supplementary Note

1.1 Recommendation Model in Detail

In this section, we presented the detailed results of the different paradigms, including the federated strategies for the federated approaches, across countries for both ROC–AUC and AUC–PR metrics. Supplementary Table 1 and 2 highlight the averages from 10 runs, while Supplementary Table 3 and 4 represent the standard deviations of these runs.

On top of the weighted average within Supplementary Table 1 and 2, we also report the macro average, where each client contributes equally, regardless of its dataset size. This complementary metric provides additional insight into model performance across the full client population. While the weighted average accounts for differences in client dataset sizes, it can be disproportionately influenced by clients with larger datasets and stronger performance. Moreover, it is important to recognize that federated optimization itself employs weighted aggregation during training, where client updates are combined in proportion to their local dataset sizes. As a result, larger clients have a stronger influence not only on evaluation but also on the global model during training. The smaller discrepancy between weighted and macro averages observed in personalized models suggests that personalization helps reduce this bias and enables better adaptation to clients with fewer samples. These findings highlight the need for caution when interpreting any single evaluation metric, as relying solely on weighted averages may overlook model behavior on underrepresented clients. Reporting both weighted and macro averages provides a more comprehensive understanding of model generalization under heterogeneous federated conditions.

Supplementary Table 1. Mean ROC–AUC scores, averaged over 10 runs, are presented for each learning paradigm: Centralized, Local, Federated, and Personalized Federated strategies (FedAVG, FedProx, FedAdam, FedYogi, FedAdagrad). The weighted average and macro average across all countries is also reported.

Country	Dataset Size	Centralized	Local	Federated					Fine-tuned					Adaptive				
				FedAVG	FedProx	FedAdam	FedYogi	FedAdagrad	FedAVG	FedProx	FedAdam	FedYogi	FedAdagrad	FedAVG	FedProx	FedAdam	FedYogi	FedAdagrad
CZ	55435	0.8781	0.8669	0.8918	0.8909	0.8948	0.9010	0.8883	0.9194	0.9193	0.9158	**0.9222**	0.9173	0.9199	0.9211	0.9198	0.9143	0.9184
IT	54354	0.7769	0.7603	0.796	0.7946	0.7949	0.7951	0.7890	0.8099	0.8108	0.8108	0.8151	0.8045	0.8140	**0.8168**	0.8121	0.7966	0.8118
TR	37853	0.8467	0.8498	0.8355	0.8365	0.8478	0.8482	0.8339	0.8810	0.8854	0.886	0.8881	0.8809	**0.8891**	0.8887	0.8872	0.8696	0.8843
ES	33396	0.7713	0.7609	0.7691	0.7680	0.7784	0.7793	0.7642	0.7972	0.7969	0.8013	0.802	0.7938	**0.8028**	0.8027	0.8006	0.7789	0.7990
CA	27131	0.7488	0.7332	0.7366	0.7370	0.7440	0.741	0.7327	0.7592	0.7615	0.7570	0.7573	0.7617	0.7597	**0.7619**	0.7543	0.7435	0.758
AU	23906	0.7490	0.7287	0.7315	0.7300	0.7402	0.7419	0.7187	0.7678	0.7657	0.7623	0.767	0.7582	**0.7692**	0.7679	0.7594	0.7497	0.7681
PT	6884	0.8252	0.7972	0.7389	0.7449	0.7557	0.7514	0.7377	0.8539	0.8525	0.8285	0.8266	0.8517	0.8496	0.8475	0.8212	0.8214	**0.8542**
BE	6534	0.7963	0.7987	0.6560	0.6495	0.695	0.6678	0.6216	0.8183	0.8156	0.8134	0.8026	0.8186	**0.8225**	0.8115	0.8035	0.8084	0.8185
KW	5725	0.8761	0.9164	0.7533	0.7445	0.7149	0.7178	0.7071	**0.9200**	0.9128	0.897	0.9002	0.9113	0.9038	0.9137	0.9000	0.8915	0.9044
HU	4892	0.9495	0.9549	0.7065	0.7128	0.7883	0.7393	0.6892	0.9625	**0.9632**	0.9531	0.9507	0.9597	0.9559	0.9608	0.9391	0.8947	0.9557
NL	4869	0.6873	**0.7107**	0.5592	0.5614	0.5796	0.5743	0.5515	0.6761	0.6782	0.6531	0.6628	0.6747	0.6642	0.6595	0.6537	0.6259	0.6567
TN	4780	0.9319	0.9312	0.7816	0.7857	0.8194	0.7974	0.7577	**0.9538**	0.9535	0.946	0.9471	0.9492	0.9502	0.9535	0.946	0.9136	0.9451
CH	3836	**0.7925**	0.7274	0.6197	0.6212	0.608	0.6145	0.5837	0.7748	0.7700	0.7368	0.7428	0.7690	0.7768	0.7650	0.7498	0.7480	0.7571
IR	2980	0.8158	0.7471	0.6138	0.6396	0.6305	0.6205	0.6168	0.8190	**0.8330**	0.7838	0.7864	0.8148	0.8079	0.8269	0.7231	0.7032	0.8127
AR	2440	0.8274	0.8784	0.6705	0.6714	0.7058	0.6842	0.6452	0.8690	0.8719	0.8711	0.8716	0.8682	0.8508	**0.8856**	0.7469	0.7105	0.8419
LB	1937	0.7314	0.6553	0.6062	0.5955	0.6023	0.6149	0.6164	0.7462	0.7398	0.7544	0.7417	0.7619	0.7336	0.7589	**0.7741**	0.7534	0.7574
US	1344	0.7437	0.7044	0.5713	0.5627	0.5637	0.5888	0.5674	0.7418	0.7368	0.7160	**0.7439**	0.7308	0.7053	0.7303	0.6902	0.6014	0.7135
IL	1140	0.8537	0.8503	0.6918	0.6937	0.7225	0.6835	0.6942	**0.8795**	0.8782	0.8693	0.8745	0.8773	0.8698	0.875	0.8693	0.8075	0.8762
OM	969	0.8093	0.7981	0.5276	0.5339	0.5086	0.533	0.4890	0.8618	**0.8731**	0.8349	0.8361	0.8277	0.8650	0.8472	0.8531	0.7483	0.8590
CU	782	0.8062	**0.8266**	0.5768	0.5625	0.6726	0.6565	0.5890	0.8043	0.8050	0.8093	0.8112	0.8076	0.7924	0.7971	0.7775	0.7509	0.7848
BR	578	0.7307	0.7070	0.6355	0.5768	0.5129	0.5281	0.5540	0.7736	0.7680	0.7877	**0.8303**	0.7825	0.7960	0.8063	0.8000	0.7394	0.7733
SA	256	**0.9374**	0.8827	0.6170	0.6749	0.7494	0.6704	0.6915	0.8953	0.8677	0.8643	0.8629	0.8877	0.8974	0.8915	0.8740	0.8740	0.9166
GB	221	**0.8510**	0.5333	0.7580	0.6520	0.7210	0.6778	0.7160	0.7625	0.6880	0.7100	0.6261	0.7675	0.6410	0.6060	0.6475	0.7850	0.7055
NZ	110	0.419	0.5873	0.4667	0.3286	0.5524	0.4497	0.3571	0.6238	0.4857	0.6238	0.6243	0.6571	0.6286	0.6095	0.581	0.2714	**0.6905**
GR	99	**0.9292**	0.7998	0.7146	0.724	0.6932	0.7141	0.7167	0.8958	0.8703	0.8495	0.8976	0.8604	0.8609	0.8714	0.9062	0.8828	0.837
Weighted Average		0.8092	0.7983	0.784	0.7835	0.7919	0.7910	0.7762	0.837	0.8375	0.8339	0.8369	0.8340	0.8384	**0.8398**	0.8324	0.8178	0.8361
Macro Average		0.8034	0.7803	0.6809	0.6717	0.6958	0.6836	0.6651	**0.8226**	0.8121	0.8094	0.8116	0.8197	0.8130	0.8150	0.7996	0.7673	0.8159

1.2 The Best Analysis Paradigms for each Country

This section provides an overview to identify the best-performing paradigm for each country by comparing the results. It highlights the top-performing model for each country, along with its respective score, offering a clear overview of which paradigm performs best in each case.

1.3 Benchmarking Personalized Federated Models

This section provides implementation details for the benchmark PFL algorithms, Ditto[1] and FedPer[2], which are used for comparison against our proposed AdaptiveDualBranchNet. To ensure experimental consistency, both methods were implemented using

Supplementary Table 2. Average AUC–PR Performance Across Paradigms and Countries. Mean AUC–PR scores, averaged over 10 runs, are reported for each learning paradigm: Centralized, Local, Federated, and Personalized Federated strategies (FedAVG, FedProx, FedAdam, FedYogi, FedAdagrad). The weighted average and macro average across all countries is also provided.

Country	Dataset Size	Centralized	Local	Federated					Fine-tuned					Adaptive				
				FedAVG	FedProx	FedAdam	FedYogi	FedAdagrad	FedAVG	FedProx	FedAdam	FedYogi	FedAdagrad	FedAVG	FedProx	FedAdam	FedYogi	FedAdagrad
CZ	55435	0.5568	0.4600	0.5803	0.5875	0.6536	0.6482	0.5667	0.6885	0.6990	0.7197	**0.7266**	0.6719	0.7192	0.7214	0.7128	0.6633	0.6985
IT	54354	0.4312	0.3491	0.4509	0.459	0.4798	0.4841	0.4377	0.4831	0.4924	0.5175	**0.5247**	0.4687	0.5034	0.5099	0.5116	0.4710	0.4909
TR	37853	0.483	0.4161	0.4402	0.4427	0.5248	0.5146	0.4305	0.5583	0.5716	**0.6139**	0.6041	0.5509	0.5852	0.5953	0.5936	0.5278	0.5631
ES	33396	0.4055	0.3381	0.3762	0.3819	0.4085	0.4055	0.373	0.4302	0.4393	**0.4644**	0.4582	0.4313	0.4503	0.4564	0.4504	0.4018	0.4294
CA	27131	0.3893	0.3407	0.3408	0.3383	0.3606	0.3465	0.3369	0.3958	0.3949	**0.4108**	0.4044	0.3957	0.4032	0.3965	0.3982	0.3764	0.391
AU	23906	0.3882	0.3129	0.3037	0.3098	0.3372	0.3372	0.2884	0.3782	0.3805	0.3955	0.3925	0.3558	0.3920	**0.3962**	0.3769	0.3456	0.3769
PT	6884	0.449	0.3562	0.2445	0.2627	0.2744	0.2599	0.2440	0.4485	**0.4774**	0.4464	0.4434	0.4610	0.4545	0.4524	0.4208	0.4040	0.4658
BE	6534	0.3553	0.2987	0.153	0.1559	0.2077	0.1876	0.1356	0.4072	0.405	**0.4130**	0.3978	0.3868	0.4022	0.398	0.3737	0.3586	0.3797
KW	5725	0.4558	0.4850	0.1719	0.1661	0.1526	0.1518	0.1408	**0.5404**	0.5111	0.4789	0.4919	0.5143	0.5004	0.5104	0.4729	0.4280	0.4659
HU	4892	0.7311	0.5924	0.2944	0.3099	0.4085	0.3254	0.2388	**0.7887**	0.779	0.7667	0.7221	0.7623	0.7501	0.781	0.6076	0.3464	0.7533
NL	4869	**0.2826**	0.2650	0.1750	0.1797	0.1872	0.1876	0.176	0.2733	0.2756	0.2529	0.2619	0.2734	0.2611	0.2539	0.2519	0.2114	0.2556
TN	4780	0.8178	0.804	0.4926	0.503	0.5926	0.5535	0.4682	0.8627	**0.8664**	0.8432	0.8369	0.8452	0.8441	0.8639	0.8416	0.7411	0.8313
CH	3836	0.3084	0.1952	0.1187	0.1232	0.1183	0.1204	0.1121	**0.3400**	0.3196	0.2887	0.3005	0.3121	0.3336	0.3042	0.2877	0.2548	0.2983
IR	2980	0.5345	0.3682	0.2382	0.2514	0.2628	0.2277	0.2276	0.5369	**0.5702**	0.5119	0.5294	0.5261	0.5265	0.5570	0.3684	0.3216	0.5212
AR	2440	0.5801	0.5946	0.2617	0.2625	0.3011	0.2578	0.2211	**0.6332**	0.6331	0.6255	0.5897	0.6111	0.5212	0.6287	0.2828	0.2372	0.5229
LB	1937	0.2756	0.2255	0.1287	0.1171	0.1193	0.1259	0.1232	0.3334	0.3187	0.3139	0.3467	0.3219	0.2788	0.3343	**0.3853**	0.3245	0.3376
US	1344	0.3128	0.2433	0.1441	0.1383	0.1266	0.1382	0.1493	0.3142	0.2797	0.2337	0.2860	**0.3219**	0.2552	0.2493	0.1899	0.1313	0.239
IL	1140	0.5198	0.5181	0.2702	0.2604	0.2581	0.2743	0.2397	0.6307	**0.6310**	0.5660	0.6065	0.5879	0.5899	0.5986	0.5540	0.4258	0.5727
OM	969	0.4763	0.2995	0.1024	0.0962	0.0858	0.093	0.0994	0.5753	**0.5897**	0.5157	0.5192	0.5477	0.5480	0.5421	0.5505	0.2538	0.5178
CU	782	0.4775	**0.5260**	0.1882	0.1864	0.2688	0.2458	0.1967	0.458	0.4744	0.4554	0.4369	0.4491	0.432	0.419	0.3859	0.3309	0.4225
BR	578	0.4677	0.4605	0.174	0.1434	0.1214	0.1118	0.1181	**0.4911**	0.4200	0.4832	0.4511	0.4209	0.3897	0.4308	0.3653	0.2558	0.3937
SA	256	0.6851	**0.7674**	0.1858	0.2466	0.2986	0.2577	0.2888	0.6288	0.5619	0.4946	0.4998	0.6547	0.5997	0.659	0.529	0.4330	0.6318
GB	221	**0.4774**	0.2529	0.3233	0.2576	0.2889	0.3177	0.287	0.3653	0.3007	0.3117	0.3141	0.3541	0.3371	0.3250	0.3100	0.3643	0.3142
NZ	110	0.0810	0.1057	0.0866	0.0738	0.1533	0.1247	0.0717	0.1377	0.1094	**0.2039**	0.1890	0.1857	0.1332	0.1247	0.1185	0.0722	0.1577
GR	99	**0.9082**	0.8048	0.6472	0.6495	0.5825	0.6245	0.6739	0.8955	0.8777	0.8127	0.8837	0.8692	0.8577	0.8794	0.8982	0.8531	0.8428
Weighted Average		0.4605	0.3874	0.4030	0.4081	0.4488	0.442	0.3913	0.5156	0.5221	**0.5383**	0.5379	0.5043	0.5291	0.5346	0.5197	0.4703	0.5131
Macro Average		0.4740	0.4152	0.2757	0.2761	0.3029	0.2928	0.2658	**0.5038**	0.4951	0.4855	0.4886	0.4911	0.4827	0.4955	0.4495	0.3813	0.4749

Supplementary Table 3. Standard Deviation of the ROC–AUC Scores for Each Paradigm Across Countries

Country	Dataset Size	Centralized	Local	Federated					fine-tuned					Adaptive				
				FedAVG	FedProx	FedAdam	FedYogi	FedAdagrad	FedAVG	FedProx	FedAdam	FedYogi	FedAdagrad	FedAVG	FedProx	FedAdam	FedYogi	FedAdagrad
CZ	55435	0.0017	0.0133	0.0055	0.0026	0.0063	0.0063	0.0045	0.0023	0.0030	0.0034	0.0021	0.0028	0.0019	0.0021	0.0027	0.0029	0.0021
IT	54354	0.0029	0.0068	0.0034	0.0044	0.0060	0.0054	0.0013	0.0046	0.0034	0.0051	0.0036	0.0033	0.0042	0.0032	0.0064	0.0036	0.0038
TR	37853	0.0045	0.0078	0.0048	0.0059	0.0065	0.0045	0.0046	0.0034	0.0043	0.0043	0.0029	0.0041	0.0031	0.0038	0.0046	0.0042	0.0040
ES	33396	0.0051	0.0115	0.0049	0.0038	0.0091	0.0070	0.0051	0.0046	0.0031	0.0068	0.0050	0.0055	0.0049	0.0046	0.0073	0.0051	0.0059
CA	27131	0.0058	0.0068	0.0055	0.0053	0.0058	0.0060	0.0080	0.0043	0.0040	0.0063	0.0068	0.0053	0.0041	0.0056	0.0040	0.0061	0.0069
AU	23906	0.0057	0.0041	0.0086	0.0090	0.0076	0.0088	0.0049	0.0044	0.0096	0.0058	0.0067	0.0049	0.0062	0.0100	0.0072	0.0061	0.0104
PT	6884	0.0119	0.0040	0.0145	0.0129	0.0151	0.0109	0.0151	0.0123	0.0058	0.0128	0.0170	0.0121	0.0123	0.0087	0.0181	0.0146	0.0085
BE	6534	0.0109	0.0088	0.0094	0.0104	0.0183	0.0234	0.0104	0.0091	0.0082	0.0114	0.0110	0.0093	0.0125	0.0117	0.0101	0.0118	0.0117
KW	5725	0.0114	0.0027	0.0180	0.0216	0.0234	0.0253	0.0285	0.0134	0.0118	0.0080	0.0095	0.0121	0.0077	0.0065	0.0133	0.0128	0.0097
HU	4892	0.0037	0.0033	0.0309	0.0464	0.0624	0.0525	0.0290	0.0070	0.0047	0.0085	0.0095	0.0044	0.0056	0.0077	0.0138	0.0197	0.0046
NL	4869	0.0114	0.0077	0.0147	0.0203	0.0141	0.0136	0.0089	0.0245	0.0227	0.0148	0.0221	0.0143	0.0237	0.0210	0.0349	0.0100	0.0183
TN	4780	0.0051	0.0081	0.0153	0.0087	0.0194	0.0140	0.0211	0.0038	0.0038	0.0076	0.0063	0.0044	0.0053	0.0045	0.0078	0.0058	0.0037
CH	3836	0.0151	0.0061	0.0258	0.0331	0.0216	0.0331	0.0319	0.0277	0.0130	0.0117	0.0209	0.0249	0.0213	0.0271	0.0106	0.0225	0.0147
IR	2980	0.0142	0.0067	0.0212	0.0256	0.0175	0.0368	0.0167	0.0211	0.0150	0.0183	0.0241	0.0195	0.0131	0.0162	0.0476	0.0202	0.0124
AR	2440	0.0109	0.0042	0.0251	0.0221	0.0153	0.0344	0.0248	0.0116	0.0153	0.0172	0.0180	0.0167	0.0102	0.0138	0.0539	0.0331	0.0108
LB	1937	0.0401	0.0161	0.0337	0.0297	0.0435	0.0396	0.0272	0.0230	0.0265	0.0280	0.0402	0.0175	0.0335	0.0353	0.0250	0.0305	0.0210
US	1344	0.0303	0.0154	0.0415	0.0294	0.0410	0.0447	0.0330	0.0319	0.0247	0.0295	0.0460	0.0328	0.0312	0.0339	0.0381	0.0402	0.0352
IL	1140	0.0116	0.0112	0.0369	0.0250	0.0249	0.0423	0.0264	0.0136	0.0197	0.0126	0.0225	0.0087	0.0163	0.0150	0.0152	0.0192	0.0129
OM	969	0.0445	0.0135	0.0504	0.0429	0.0292	0.0835	0.0745	0.0248	0.0326	0.0266	0.0499	0.0454	0.0415	0.0350	0.0406	0.0345	0.0360
CU	782	0.0188	0.0089	0.0248	0.0435	0.0491	0.0278	0.0341	0.0196	0.0151	0.0371	0.0214	0.0174	0.0229	0.0133	0.0331	0.0188	0.0227
BR	578	0.0473	0.0298	0.0670	0.0587	0.0965	0.0476	0.0496	0.0670	0.0590	0.0730	0.0603	0.0522	0.0617	0.0696	0.0523	0.0693	0.0649
SA	256	0.0362	0.0211	0.0581	0.0642	0.0685	0.0767	0.0807	0.0383	0.0562	0.0701	0.0846	0.0664	0.0685	0.0694	0.0681	0.0602	0.0225
GB	221	0.0512	0.0400	0.0764	0.0563	0.1026	0.1047	0.0669	0.0754	0.0653	0.1139	0.0950	0.0709	0.0496	0.0524	0.0823	0.0706	0.0512
NZ	110	0.1409	0.0673	0.1274	0.1800	0.2644	0.2511	0.1190	0.1600	0.2571	0.2772	0.3130	0.2201	0.1518	0.1605	0.1741	0.2054	0.1317
GR	99	0.0212	0.0095	0.0412	0.0716	0.0748	0.0877	0.0803	0.0557	0.0354	0.0607	0.0418	0.0637	0.0528	0.0472	0.0586	0.0636	0.0467

Supplementary Table 4. Standard Deviation of the AUC–PR Scores for Each Paradigm Across Countries

Country	Dataset Size	Centralized	Local	Federated					fine-tuned					Adaptive				
				FedAVG	FedProx	FedAdam	FedYogi	FedAdagrad	FedAVG	FedProx	FedAdam	FedYogi	FedAdagrad	FedAVG	FedProx	FedAdam	FedYogi	FedAdagrad
CZ	55435	0.0095	0.0463	0.0209	0.0155	0.0153	0.0131	0.0161	0.0128	0.0129	0.0080	0.0122	0.0141	0.0087	0.0059	0.0128	0.0146	0.0122
IT	54354	0.0032	0.0129	0.0121	0.0086	0.0139	0.0110	0.0086	0.0096	0.0080	0.0090	0.0082	0.0078	0.0106	0.0043	0.0140	0.0062	0.0106
TR	37853	0.0118	0.0210	0.0144	0.0121	0.0172	0.0163	0.0158	0.0134	0.0094	0.0102	0.0127	0.0143	0.0167	0.0090	0.0157	0.0241	0.0167
ES	33396	0.0096	0.0166	0.0109	0.0106	0.0148	0.0143	0.0115	0.0095	0.0104	0.0154	0.0113	0.0133	0.0118	0.0076	0.0119	0.0095	0.0133
CA	27131	0.0066	0.0148	0.0097	0.0110	0.0135	0.0072	0.0130	0.0124	0.0077	0.0111	0.0078	0.0098	0.0108	0.0087	0.0124	0.0109	0.0066
AU	23906	0.0118	0.0068	0.0104	0.0125	0.0153	0.0198	0.0097	0.0130	0.0125	0.0086	0.0137	0.0082	0.0100	0.0168	0.0131	0.0076	0.0191
PT	6884	0.0227	0.0078	0.0216	0.0211	0.0316	0.0155	0.0210	0.0264	0.0211	0.0271	0.0282	0.0287	0.0154	0.0257	0.0297	0.0205	0.0312
BE	6534	0.0173	0.0136	0.0137	0.0087	0.0177	0.0243	0.0102	0.0120	0.0229	0.0254	0.0204	0.0156	0.0339	0.0195	0.0314	0.0218	0.0294
KW	5725	0.0246	0.0294	0.0204	0.0208	0.0210	0.0240	0.0152	0.0593	0.0601	0.0366	0.0506	0.0350	0.0502	0.0441	0.0505	0.0576	0.0542
HU	4892	0.0211	0.0358	0.0327	0.0432	0.0440	0.0364	0.0298	0.0281	0.0481	0.0310	0.0431	0.0294	0.0343	0.0336	0.1354	0.0904	0.0244
NL	4869	0.0102	0.0137	0.0099	0.0096	0.0107	0.0119	0.0112	0.0290	0.0185	0.0147	0.0185	0.0128	0.0144	0.0162	0.0407	0.0087	0.0218
TN	4780	0.0132	0.0271	0.0208	0.0204	0.0483	0.0377	0.0341	0.0114	0.0188	0.0243	0.0244	0.0160	0.0173	0.0138	0.0191	0.0280	0.0114
CH	3836	0.0214	0.0096	0.0112	0.0146	0.0109	0.0201	0.0122	0.0375	0.0340	0.0468	0.0530	0.0414	0.0512	0.0485	0.0315	0.0360	0.0276
IR	2980	0.0298	0.0117	0.0174	0.0251	0.0314	0.0337	0.0200	0.0223	0.0393	0.0380	0.0360	0.0291	0.0181	0.0186	0.0832	0.0541	0.0317
AR	2440	0.0245	0.0155	0.0350	0.0333	0.0328	0.0197	0.0287	0.0291	0.0289	0.0212	0.0747	0.0314	0.0458	0.0380	0.1085	0.0597	0.0457
LB	1937	0.0468	0.0198	0.0262	0.0248	0.0281	0.0254	0.0220	0.0494	0.0628	0.0316	0.0672	0.0364	0.0518	0.0603	0.0606	0.0531	0.0308
US	1344	0.0275	0.0285	0.0302	0.0227	0.0211	0.0224	0.0455	0.0640	0.0554	0.0445	0.0491	0.0456	0.0300	0.0319	0.0357	0.0158	0.0432
IL	1140	0.0224	0.0264	0.0415	0.0340	0.0552	0.0574	0.0285	0.0350	0.0348	0.0671	0.0762	0.0439	0.0550	0.0416	0.0521	0.0374	0.0443
OM	969	0.0663	0.0439	0.0310	0.0139	0.0111	0.0178	0.0269	0.0583	0.0797	0.0318	0.0993	0.0899	0.0445	0.0669	0.1074	0.0442	0.0561
CU	782	0.0243	0.0275	0.0237	0.0361	0.0558	0.0380	0.0304	0.0460	0.0524	0.0735	0.0418	0.0376	0.0582	0.0366	0.0547	0.0339	0.0405
BR	578	0.0425	0.0339	0.0559	0.0493	0.0431	0.0266	0.0152	0.0807	0.0590	0.1318	0.1231	0.0900	0.1174	0.1543	0.1008	0.0793	0.0596
SA	256	0.1077	0.0441	0.0542	0.0930	0.1281	0.1371	0.1394	0.1843	0.2266	0.2258	0.2392	0.2324	0.2366	0.2157	0.2314	0.1147	0.1189
GB	221	0.0826	0.0624	0.0854	0.0691	0.1233	0.1206	0.1204	0.0560	0.0546	0.0999	0.0434	0.1064	0.0804	0.0733	0.0912	0.0978	0.0877
NZ	110	0.0242	0.0155	0.0217	0.0322	0.1304	0.1341	0.0160	0.0751	0.0602	0.1633	0.1165	0.1334	0.0564	0.0465	0.0506	0.0433	0.0671
GR	99	0.0274	0.0075	0.0637	0.1099	0.0816	0.1163	0.0971	0.0601	0.0358	0.0743	0.0450	0.0481	0.0471	0.0410	0.0506	0.0923	0.0582

Supplementary Table 5. Best Paradigm for ROC–AUC and AUC–PR Scores by Country

Country	ROC–AUC		AUC–PR	
	Paradigm	Score	Paradigm	Score
CZ	Fine-tuned - FedYogi	0.9222	Fine-tuned - FedYogi	0.7266
IT	Adaptive - FedProx	0.8168	Fine-tuned - FedYogi	0.5247
TR	Adaptive - FedAVG	0.8891	Fine-tuned - FedAdam	0.6139
ES	Adaptive - FedAVG	0.8028	Fine-tuned - FedAdam	0.4644
CA	Adaptive - FedProx	0.7619	Fine-tuned - FedAdam	0.4108
AU	Adaptive - FedAVG	0.7692	Adaptive - FedAVG	0.3962
PT	Adaptive - FedAdagrad	0.8542	Fine-tuned - FedProx	0.4774
BE	Adaptive - FedAVG	0.8225	Fine-tuned - FedProx	0.4130
KW	Fine-tuned - FedAVG	0.9200	Fine-tuned - FedAVG	0.5404
HU	Fine-tuned - FedProx	0.9632	Fine-tuned - FedAVG	0.7887
NL	Local	0.7107	Centralized	0.2826
TN	Fine-tuned - FedAVG	0.9538	Fine-tuned - FedProx	0.8664
CH	Centralized	0.7925	Centralized	0.3400
IR	Fine-tuned - FedProx	0.8330	Fine-tuned - FedProx	0.5702
AR	Adaptive - FedProx	0.8856	Fine-tuned - FedProx	0.6332
LB	Adaptive - FedAdagrad	0.7741	Adaptive - FedAdagrad	0.3853
US	fine-tuned - FedYogi	0.7439	Fine-tuned - FedYogi	0.3219
IL	Fine-tuned - FedAVG	0.8795	Fine-tuned - FedAVG	0.6310
OM	Fine-tuned - FedProx	0.8731	Fine-tuned - FedProx	0.5897
CU	Local	0.8266	Local	0.5260
BR	Fine-tuned - FedYogi	0.8303	Fine-tuned - FedProx	0.4911
SA	Centralized	0.9374	Local	0.7674
GB	Centralized	0.8510	Centralized	0.4774
NZ	Adaptive - FedAdagrad	0.6905	Adaptive - FedAdagrad	0.2039
GR	Centralized	0.9292	Centralized	0.9082
Weighted Average		0.8398	0.5383	

the Flower simulation framework[3]. A high-level comparison of their conceptual differences relative to AdaptiveDualBranchNet is provided in Table 6.

Supplementary Table 6. Comparison of Personalization Federated Learning Strategies

Aspect	**AdaptiveDualBranchNet**	**FedPer**	**Ditto**
Architecture Type	Dual-branch (core + extension)	Split-layer MLP	Single MLP
Personalization Scope	Extension branch	Final layers	Entire model
Adaptivity	Yes (via data size n_i)	No (fixed layers)	No (regularization only)
Personalization Method	Structural + adaptive depth	Structural split	Optimization constraint
Synchronized Parameters	θ_c (core only)	θ_s (shared layers)	θ (global model)
Client-Specific Components	$\theta_e^{(i)}$	$\theta_p^{(i)}$	w_i

The Ditto[1] algorithm represents an optimization-based PFL strategy. Each client i maintains both a personalized model copy (w_i) and a local version of the global model (θ). While the global model θ is updated via standard FedAVG, the personalized model w_i is trained locally on data $\mathscr{D}_i$ by minimizing the regularized objective $\mathscr{L}_{\text{Ditto}}(w_i) = \mathscr{L}_{\text{local}}(w_i; \mathscr{D}_i) + \frac{\lambda}{2}\|w_i - \theta\|^2$. Here, the regularization hyperparameter λ controls the proximity to the global model. We used the same base network architecture as the FedAVG-Baseline and tested $\lambda \in \{0.05, 0.1, 0.3\}$. Local optimization involved 10 steps per round.

On the other hand, the FedPer[2] algorithm employs architectural personalization by partitioning the model into shared base layers (θ_s) and client-specific layers ($\theta_p^{(i)}$). Only θ_s are globally aggregated via FedAVG; $\theta_p^{(i)}$ are trained locally and remain private. Using the FedAVG-Baseline architecture, we evaluated configurations where the final 1, 2, or 3 fully connected layers served as the personalized head (FedPer-1, FedPer-2, FedPer-3, respectively). Both shared and personalized parameters were updated during local training steps.

The comparative performance of these methods is summarized in Table 7 and visualized per-client in Figure 1. The key findings from these benchmark experiments indicate that, for this specific task and highly heterogeneous dataset, Ditto's performance remained comparable to the non-personalized FedAVG baseline across the tested λ values (e.g., ROC–AUC $\approx$ 0.782 vs. FedAVG's 0.7845). In contrast, FedPer demonstrated consistent improvement over FedAVG, with FedPer-3 achieving the best results among the fixed-layer variants (ROC–AUC 0.8135).

However, FedPer's strategy led to a weaker global model as expected, since part of original model stays at local (see Table 7, FedPer-GlobalModel rows), indicating a trade-off between personalization and the strength of the shared representation. In comparison, AdaptiveDualBranchNet achieved better average performance (ROC–AUC 0.8391, AUC–PR 0.5281), suggesting that its dynamic architectural adaptation enables more effective personalization while maintaining a strong global model.

The radar plot in Figure 1 visually corroborates these findings. The cyan line representing AdaptiveDualBranchNet consistently encloses a larger area, reflecting generally higher ROC–AUC scores across most of the 32 clients, compared to FedPer (red) and Ditto/FedAvg-Baseline (blue/pink). This suggests that the adaptive approach is better suited to handling the substantial client heterogeneity present in this dataset.

To further highlight the importance of adaptivity, rather than simply increasing parameter count by adding layers, we refer to Supplementary Table 8 , which summarizes the total number of layers in AdaptiveDualBranchNet. As expected, the core layers remain fixed at 3, while the adaptive extension layers vary across clients. The mean number of extension layers is 0.9375, effectively corresponding to one additional layer on average. This is particularly insightful when compared to FedPer with one personalization layer (FedPer-1), which employs a fixed architecture of 4 shared layers and 1 personalized layer, totaling 5 hidden layers. In contrast, AdaptiveDualBranchNet operates with fewer layers on average, yet consistently achieves better performance. This underscores that its strength lies not in parameter count, but in its ability to flexibly adjust model capacity to match client data heterogeneity, expanding for clients with abundant data and contracting for those with limited data.

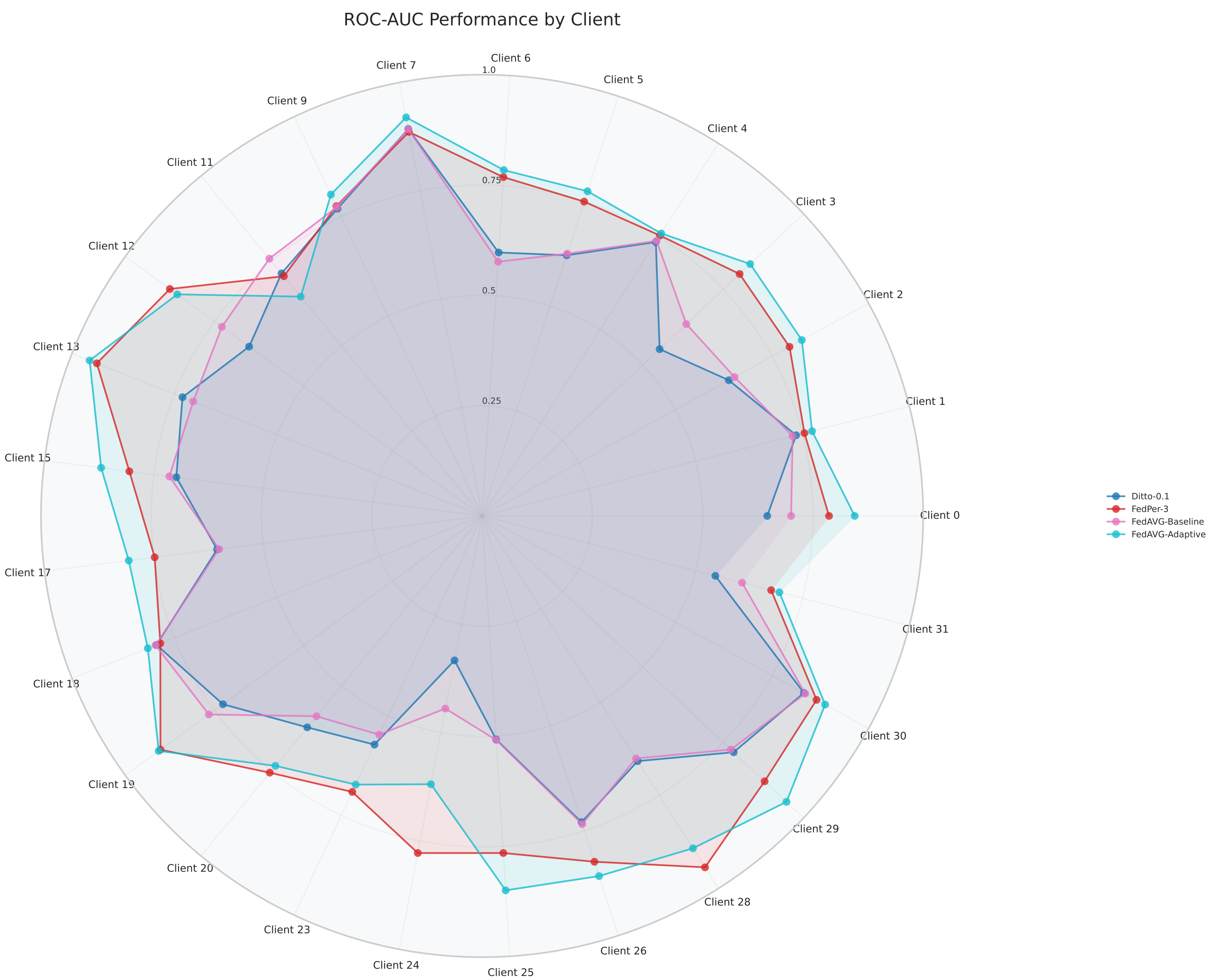
It is important to note that the reported results for Ditto and FedPer correspond to a specific set of parameter choices (e.g., λ or number of local epochs E, personalization layers); Further tuning of these hyperparameters may yield minor performance differences but is unlikely to alter the observed overall trends.

1.4 Client-Level Fine-Tuning Performance Analysis

While PFL aims to tailor models to local data, the strategy of fine-tuning pre-trained global models presents challenges, particularly under the extreme client data heterogeneity observed in this study. Adapting models on clients with very limited data risks overfitting to local noise or catastrophic forgetting of valuable knowledge acquired during federated training. Furthermore, reliably evaluating performance on correspondingly small, potentially imbalanced test sets is inherently difficult. This challenge was evident in our analysis of the seven smallest client cohorts ($<$1000 samples), where for the majority (six out of seven), the locally trained model did not outperform the PFL models or the centralized learning, underscoring the instability

Supplementary Table 7. Average ROC–AUC and AUC–PR across 3 runs for different personlized federated learning methods

Experiment	ROC–AUC	AUC–PR
FedAVG-Adaptive	0.8391	0.5281
FedPer-3	0.8135	0.4157
FedPer-2	0.8052	0.4042
FedPer-1	0.7967	0.3902
FedAVG-Baseline	0.7845	0.4034
Ditto-0.1	0.7826	0.4075
Ditto-0.05	0.7824	0.4023
Ditto-0.3	0.7814	0.4025
FedPer-3-GlobalModel	0.5206	0.1295
FedPer-2-GlobalModel	0.4684	0.1072
FedPer-1-GlobalModel	0.4527	0.1022



Supplementary Figure 1. Radar chart illustrating client-specific benchmarking across various PFL methods

of fine-tuning in data-scarce regimes. Performance metrics reported for such low-resource clients must therefore be interpreted with considerable caution.

Supplementary Table 8. Number of Core and Extension Layers per Country as Determined by the AdaptiveDualBranchNet Extension Layer Calculation Rule. The Total Hidden Layers are computed as the sum of core and extension layers.

Country	Core Layers	Extension Layers	Total Hidden Layers	Country	Core Layers	Extension Layers	Total Hidden Layers
AR	3	0	3	**IN**	3	0	3
AU	3	2	5	**IR**	3	0	3
BE	3	1	4	**IT**	3	5	8
BR	3	0	3	**KW**	3	1	4
CA	3	3	6	**LB**	3	0	3
CH	3	1	4	**MK**	3	0	3
CU	3	0	3	**MT**	3	0	3
CZ	3	5	8	**NL**	3	1	4
EG	3	0	3	**NZ**	3	0	3
ES	3	4	7	**OM**	3	0	3
FR	3	0	3	**PT**	3	1	4
GB	3	0	3	**RO**	3	0	3
GR	3	0	3	**SA**	3	0	3
HU	3	1	4	**TN**	3	1	4
IE	3	0	3	**TR**	3	4	7
IL	3	0	3	**US**	3	0	3

Our primary mitigation involved using the converged FL model as a strong starting point and employing constrained fine-tuning parameters (e.g., reduced learning rate, reduce batch size and add L2/early stopping when triggered). While alternative strategies like freezing initial layers were explored, they did not yield substantial average performance differences across the federation, likely due to the limited influence of these small clients on overall metrics. However, fine-tuning demonstrated clear value for specific client subgroups, as illustrated in Figure 2. This analysis shows the average ROC–AUC improvement post-fine-tuning, stratified by dataset size and initial FL performance quantiles. While clients with the smallest datasets (Q1/Q1) showed potentially unreliable average gains, clients with intermediate data sizes (Q2 and Q3), particularly those initially underserved by the global model (lower performance quantiles), experienced significant performance enhancements (e.g., average gains of 0.11-0.21 ROC–AUC points). This indicates that fine-tuning is most impactful for clients possessing sufficient local data to support stable adaptation.

Addressing the specific needs of low-resource clients requires further investigation. Future research could explore strategies such as client clustering or grouping based on data similarity (potentially identified using methods like OTDD[4]), enabling more robust collaborative adaptation within these clusters. Implementing such approaches in practice, however, must carefully navigate the associated privacy and governance complexities inherent in real-world clinical data sharing.

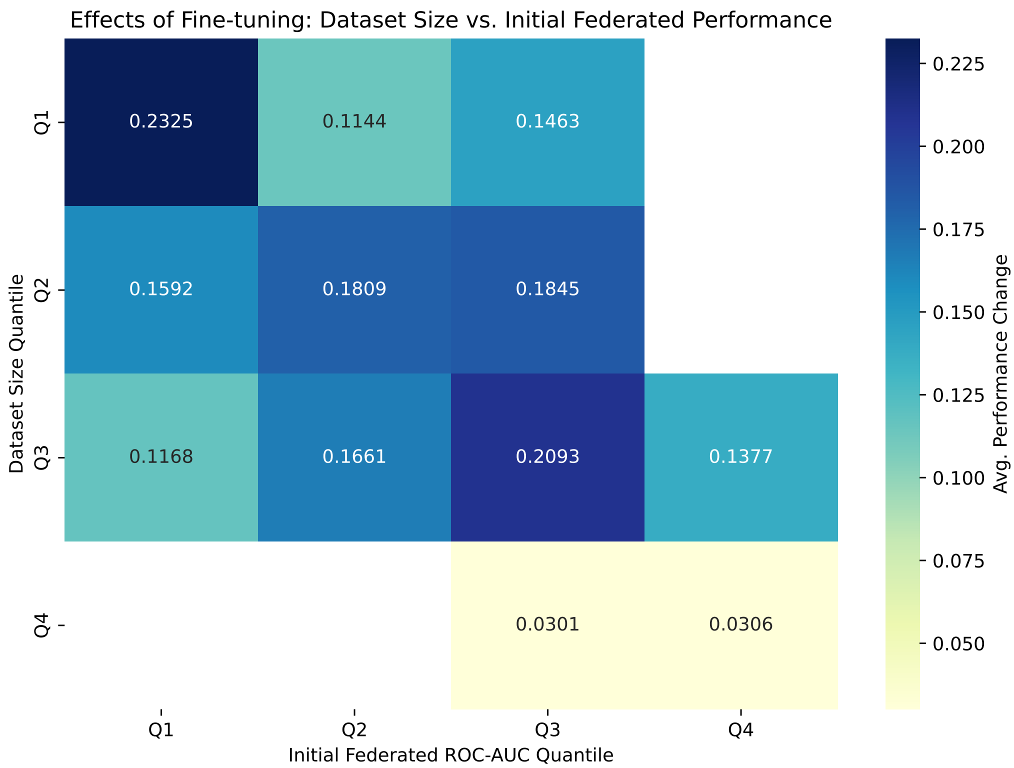
References

1. Li, T., Hu, S., Beirami, A. & Smith, V. Ditto: Fair and robust federated learning through personalization. In Meila, M. & Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, 6357–6368, DOI: 10.48550/arXiv.2012.04221 (PMLR, 2021).

2. Arivazhagan, M. G., Aggarwal, V., Singh, A. K. & Choudhary, S. Federated learning with personalization layers, DOI: 10.48550/arXiv.2003.00295 (2019). 1912.00818.

3. Beutel, D. J. *et al.* Flower: A friendly federated learning research framework, DOI: 10.48550/arXiv.2007.14390 (2020).

4. Alvarez-Melis, D. & Fusi, N. Geometric dataset distances via optimal transport. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, DOI: 10.5555/3495724.3497523 (Curran Associates Inc., Red Hook, NY, USA, 2020).



Supplementary Figure 2. Effects of Fine-tuning: Heatmap visualization of dataset size quantiles vs. initial federated performance quantiles, numbers within the quantiles highlight the average change of the performance after fine-tuning