SecDefender Appendix

Arnaldo Sgueglia, Sameera K M

August 19, 2024

A Model Architecture and Training Hyperparameters

Table 1 presents the model architectures employed in our experiments and the respective datasets. Additionally, we provide the hyperparameters utilized for each dataset.

Table 1: Model Architecture and Hyperparameter for six different datasets

Dataset	**Model Structure**
HAR	FC (561x1128) FC+Relu (128/6)
Fashion-MNIST	Conv+BN+Relu (5x1x16), MaxPool (2x2) Conv+BN+Relu (5x16x32), MaxPool (2x2), FC (1568/10)
FEDMNIST MNIST	Conv+BN+Relu (5x1x32), MaxPool (2x2) Conv+BN+Relu (5x32x64), MaxPool(2x2) FC (3136/10)
GTSR	Conv+Relu (3x3x32), MaxPool (2x2), Conv+Relu (3x32x64), MaxPool (2x2) FC+Relu (2304/128), FC (128/10)
CIC-Darknet2020	FC+Relu (79/64), FC+Relu (64/32), FC (32/4)

B ASR and TMR Attacks and Defense Performance

The tables from 2 to 7 show the attack success rates (ASR) and target misclassification rates (TMR) of the model performance on the selected datasets under various attack scenarios. Similarly, the tables from 8 to 13 show the attack success rates (ASR) and target misclassification rates (TMR) of the performance effectiveness of SecDefender on the selected datasets under different defense scenarios.

Table 2: Evaluation of the impact of full-round targeted label flipping attacks. The table compares attack success rates (ASR) for various LF scenarios.

$nLF \rightarrow$		Single-label				Double-label				Triple-label			
$Att_{ratio}\% \downarrow$	$m \downarrow$	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST
0	0	0.1656	2.442	0.309	0.526	0.1656	4.116	0.209	0.41	0.1656	7.407	0.308	0.314
10%	1	0.1633	2.941	0.372	0.526	0.1628	4.342	0.362	0.387	0.1624	7.6	0.351	0.354
	2	0.1644	2.812	0.397	0.626	0.1633	4.504	0.409	0.437	0.1627	7.941	0.371	0.407
	3	0.1644	3.098	0.477	0.724	0.1633	4.759	0.472	0.502	0.1629	7.801	0.403	0.442
20%	1	0.1633	3.215	0.395	0.603	0.1628	4.655	0.388	0.437	0.1629	7.771	0.369	0.399
	2	0.1644	3.206	0.474	0.748	0.1633	4.837	0.481	0.509	0.1631	7.939	0.421	0.469
	3	0.1644	3.705	0.531	0.839	0.1636	5.051	0.546	0.615	0.1631	8.265	0.496	0.583
30%	1	0.1633	3.25	0.419	0.606	0.1628	4.932	0.413	0.441	0.1631	8.013	0.388	0.386
	2	0.1644	3.581	0.493	0.847	0.1636	5.412	0.526	0.543	0.1633	8.425	0.431	0.546
	3	0.1644	4.16	0.609	1.065	0.1636	5.668	0.583	0.691	0.1633	8.272	0.517	0.748
40%	1	0.1633	3.49	0.433	0.71	0.1628	4.966	0.458	0.442	0.1632	8.187	0.401	0.421
	2	0.1644	4.382	0.524	0.968	0.1636	5.737	0.542	0.594	0.1635	8.939	0.4378	0.647
	3	0.1655	5.047	0.709	1.2	0.1636	6.549	0.659	0.72	0.1632	9.557	0.6196	1.013
50%	1	0.1633	3.797	0.423	0.755	0.1631	5.541	0.475	0.513	0.1634	8.498	0.419	0.445
	2	0.1644	4.886	0.611	1.007	0.1636	5.736	0.588	0.658	0.1632	9.286	0.529	0.711
	3	0.1655	5.825	0.838	1.245	0.1636	7.388	0.803	0.797	0.1632	10.049	0.674	1.103

Table 3: Evaluation of the impact of mid-round targeted label flipping attacks. The table compares attack success rates (ASR) for various LF scenarios.

$nLF \rightarrow$		Single-label				Double-label				Triple-label			
$Att_{ratio}\% \downarrow$	$m \downarrow$	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST
0	0	0.1656	2.442	0.309	0.526	0.1656	4.116	0.209	0.41	0.1656	7.407	0.308	0.314
10%	1	0.1622	2.458	0.335	0.526	0.1628	4.11	0.327	0.409	0.1624	7.275	0.324	0.377
	2	0.1622	2.729	0.373	0.521	0.1628	4.201	0.351	0.407	0.1624	7.364	0.344	0.371
	3	0.1622	2.519	0.399	0.553	0.1628	4.241	0.369	0.4562	0.1624	7.371	0.37	0.389
20%	1	0.1622	2.552	0.350	0.537	0.1628	4.21	0.335	0.412	0.1624	7.381	0.331	0.368
	2	0.1622	2.608	0.390	0.554	0.1628	4.241	0.365	0.420	0.1624	7.304	0.361	0.375
	3	0.1622	2.4345	0.438	0.6495	0.1628	4.175	0.401	0.495	0.1624	7.372	0.397	0.407
30%	1	0.1622	2.644	0.360	0.562	0.1628	4.089	0.342	0.412	0.1624	7.280	0.339	0.362
	2	0.1622	2.714	0.402	0.579	0.1628	4.241	0.385	0.459	0.1624	7.408	0.377	0.385
	3	0.1622	2.598	0.463	0.708	0.1628	4.098	0.435	0.5102	0.1624	7.372	0.4102	0.436
40%	1	0.1622	2.593	0.368	0.536	0.1628	4.211	0.347	0.418	0.1624	7.202	0.348	0.368
	2	0.1622	2.574	0.424	0.655	0.1628	4.151	0.406	0.473	0.1624	7.338	0.393	0.410
	3	0.1622	2.581	0.504	0.755	0.1628	4.272	0.454	0.566	0.1624	7.357	0.448	0.481
50%	1	0.1622	2.626	0.374	0.555	0.1628	4.197	0.361	0.429	0.1624	7.376	0.354	0.381
	2	0.1622	2.583	0.449	0.707	0.1628	4.166	0.415	0.663	0.1624	7.199	0.406	0.436
	3	0.1622	2.653	0.565	0.867	0.1628	4.085	0.473	0.615	0.1624	7.378	0.481	0.513

Table 4: Evaluation of the impact of end-round targeted label flipping attacks. The table compares attack success rates (ASR) for various LF scenarios.

$nLF \rightarrow$		Single-label				Double-label				Triple-label			
$Att_{ratio}\% \downarrow$	$m \downarrow$	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST
0	0	0.1656	2.442	0.309	0.526	0.1656	4.116	0.209	0.41	0.1656	7.407	0.308	0.314
10%	1	0.1622	2.601	0.343	0.529	0.1628	4.098	0.323	0.43	0.1624	7.293	0.322	0.378
	2	0.1622	2.741	0.371	0.582	0.1628	4.158	0.354	0.458	0.1624	7.41	0.345	0.396
	3	0.1622	2.669	0.399	0.663	0.1628	4.078	0.381	0.522	0.1624	7.353	0.370	0.395
20%	1	0.1622	2.739	0.351	0.526	0.1628	4.203	0.334	0.438	0.1624	7.195	0.333	0.379
	2	0.1622	2.593	0.337	0.659	0.1628	4.115	0.374	0.475	0.1624	7.321	0.365	0.425
	3	0.1622	2.5975	0.429	0.743	0.1628	4.244	0.415	0.611	0.1624	7.35	0.397	0.460
30%	1	0.1622	2.599	0.354	0.534	0.1628	4.396	0.343	0.444	0.1624	7.199	0.340	0.383
	2	0.1622	2.568	0.406	0.681	0.1628	4.037	0.392	0.547	0.1624	7.342	0.383	0.442
	3	0.1622	2.716	0.471	0.864	0.1628	4.211	0.436	0.635	0.1624	7.243	0.428	0.487
40%	1	0.1622	2.493	0.358	0.582	0.1628	4.014	0.349	0.459	0.1624	7.279	0.345	0.394
	2	0.1622	2.501	0.422	0.764	0.1628	4.182	0.411	0.554	0.1624	7.303	0.399	0.489
	3	0.1622	2.75	0.504	0.955	0.1628	4.144	0.469	0.711	0.1624	7.247	0.452	0.585
50%	1	0.1622	2.598	0.366	0.619	0.1628	4.186	0.364	0.464	0.1624	7.389	0.355	0.396
	2	0.1622	2.589	0.458	0.828	0.1628	4.112	0.435	0.481	0.1624	7.264	0.418	0.512
	3	0.1622	2.754	0.575	1.083	0.1628	4.192	0.47	0.82	0.1624	7.391	0.489	1.013

Table 5: Evaluation of the impact of full-round targeted label flipping attacks. The table compares target misclassification rate (TMR) for various LF scenarios.

$nLF \rightarrow$		Single-label				Double-label				Triple-label			
$Att_{ratio}\% \downarrow$	$m \downarrow$	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST
0	0	0.8252	3.009	0.89	01.19	0.8252	10.558	0.949	1.06	0.8252	18.437	0.92	0.824
10%	1	0.6517	3.643	0.953	1.161	0.7135	10.87	1.009	1.017	0.7327	18.674	0.984	0.914
	2	0.6517	3.564	0.984	1.371	0.7146	10.911	1.076	1.294	0.7331	18.894	1.018	0.995
	3	0.6529	3.853	1.096	1.448	0.7144	11.204	1.155	1.161	0.7331	18.974	1.061	1.037
20%	1	0.6517	3.973	0.989	1.313	0.7133	11.144	1.055	1.078	0.7336	18.802	1.013	0.964
	2	0.6529	3.965	1.095	1.456	0.7143	11.328	1.102	1.141	0.7325	18.946	1.089	1.136
	3	0.654	4.472	1.173	1.629	0.7149	11.704	1.256	1.274	0.7332	19.554	1.172	1.279
30%	1	0.6514	4.022	1.021	1.313	0.7129	11.557	1.094	1.078	0.7338	19.281	1.042	1.033
	2	0.6555	4.344	1.129	1.548	0.7143	11.948	1.243	1.182	0.7319	19.592	1.1	1.252
	3	0.6547	5.124	1.28	1.806	0.715	12.435	1.306	1.373	0.7324	19.604	1.219	1.508
40%	1	0.6536	4.25	1.051	1.433	0.7138	11.643	1.143	1.087	0.7332	19.432	1.062	1.04
	2	0.6569	5.157	1.172	1.747	0.7136	12.391	1.267	1.271	0.7372	20.205	1.166	1.393
	3	0.658	6.018	1.41	2.005	0.713	13.583	1.407	1.461	0.73	20.563	1.335	1.835
50%	1	0.6551	4.661	1.066	1.465	0.7151	12.492	1.185	1.165	0.7323	19.618	1.089	1.107
	2	0.6577	5.938	1.284	1.713	0.7138	12.391	1.337	1.361	0.7298	20.576	1.23	1.466
	3	0.6599	6.859	1.567	2.073	0.7135	14.673	1.571	1.522	0.7274	21.375	1.407	1.813

Table 6: Evaluation of the impact of mid-round targeted label flipping attacks. The table compares target misclassification rate (TMR) for various LF scenarios.

$nLF \rightarrow$		Single-label				Double-label				Triple-label			
$Att_{ratio}\% \downarrow$	$m \downarrow$	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST
0	0	0.8252	3.009	0.89	01.19	0.8252	10.558	0.949	1.06	0.8252	18.437	0.92	0.824
10%	1	0.6506	3.015	0.908	1.195	0.7136	10.34	0.975	1.019	0.7329	18.387	0.943	0.963
	2	0.6506	3.345	0.944	1.216	0.7136	10.403	1.002	1.036	0.7329	18.414	0.971	0.953
	3	0.6506	3.083	0.973	1.273	0.7136	10.570	1.028	1.127	0.7329	18.374	1.001	0.976
20%	1	0.6506	3.181	0.923	1.234	0.7136	10.615	0.986	1.053	0.7329	18.441	0.954	0.958
	2	0.6506	3.213	0.964	1.259	0.7136	10.57	1.027	1.058	0.7329	18.369	0.996	0.974
	3	0.6506	3.321	1.018	1.398	0.7136	10.594	1.077	1.139	0.7329	18.401	1.042	0.979
30%	1	0.6506	3.222	0.934	1.252	0.7136	10.478	0.999	1.05	0.7329	18.330	0.966	0.932
	2	0.6506	3.306	0.975	1.304	0.7136	10.57	1.06	1.111	0.7329	18.548	1.019	0.964
	3	0.6506	3.159	1.05	1.405	0.7142	10.365	1.121	1.171	0.7329	18.401	1.064	1.021
40%	1	0.6506	3.152	0.945	1.223	0.7136	10.485	1.008	1.055	0.7329	18.353	0.979	0.9765
	2	0.6506	3.147	1.009	1.363	0.7147	10.526	1.089	1.12	0.7329	18.376	1.04	1.015
	3	0.6506	3.012	1.09	1.462	0.7142	10.542	1.156	1.212	0.7323	18.328	1.11	1.077
50%	1	0.6506	3.19	0.946	1.215	0.7136	10.561	1.031	1.058	0.7329	18.53	0.990	0.989
	2	0.6506	3.197	1.032	1.442	0.7152	10.544	1.106	1.368	0.7327	18.146	1.06	1.037
	3	0.6506	3.221	1.169	1.597	0.7149	10.427	1.176	1.273	0.7314	18.529	1.152	1.114

Table 7: Evaluation of the impact of end-round targeted label flipping attacks. The table compares target misclassification rate (TMR) for various LF scenarios.

$nLF \rightarrow$		Single-label				Double-label				Triple-label			
$Att_{ratio}\% \downarrow$	$m \downarrow$	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST
0	0	0.8252	3.009	0.89	01.19	0.8252	10.558	0.949	1.06	0.8252	18.437	0.92	0.824
10%	1	0.6506	3.162	0.919	1.269	0.7136	10.399	0.973	1.001	0.7329	18.441	0.943	0.967
	2	0.6506	3.379	0.945	1.263	0.7136	10.564	1.008	1.106	0.7329	18.373	0.977	0.985
	3	0.6506	3.216	0.976	1.407	0.7136	10.391	1.042	1.219	0.7329	18.427	1.101	0.996
20%	1	0.6506	3.302	0.928	1.254	0.7136	10.553	0.990	1.102	0.7329	18.262	0.957	1.006
	2	0.6506	3.153	0.955	1.389	0.7136	10.505	1.037	1.155	0.7329	18.348	1.004	1.018
	3	0.6506	3.159	1.013	1.496	0.7136	10.601	1.091	1.288	0.7329	18.292	1.044	1.065
30%	1	0.6506	3.245	0.931	1.26	0.7136	10.372	1.006	1.0937	0.7329	18.292	0.971	0.974
	2	0.6506	3.13	0.988	1.442	0.7136	10.322	1.068	1.234	0.7327	18.444	1.031	1.046
	3	0.6506	3.282	1.068	1.644	0.7103	10.631	1.13	1.323	0.7316	18.388	1.082	1.100
40%	1	0.6506	3.052	0.935	1.308	0.7136	10.381	1.015	1.118	0.7329	18.299	0.981	1.009
	2	0.6506	3.102	1.009	1.523	0.735	10.464	1.096	1.212	0.7305	18.391	0.995	1.016
	3	0.6506	3.306	1.103	1.75	0.7103	10.506	1.176	1.398	0.7305	18.391	1.11	1.23
50%	1	0.6506	3.201	0.943	1.327	0.7136	10.588	1.040	1.131	0.7329	18.307	0.995	1.016
	2	0.6506	3.14	1.052	1.589	0.7124	10.502	1.135	1.151	0.7306	18.411	1.072	1.142
	3	0.6506	3.307	1.188	1.881	0.7082	10.592	1.18	1.528	0.7286	18.445	1.156	1.803

Table 8: Evaluation of the effectiveness of SecDefender against full-round targeted label flipping attacks. The table compares attack success rates (ASR) for various LF scenarios.

$nLF \rightarrow$		Single-label				Double-label				Triple-label			
$Att_{ratio}\% \downarrow$	$m \downarrow$	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST
0	0	0.1656	2.442	0.309	0.526	0.1656	4.116	0.209	0.41	0.1656	7.407	0.308	0.314
10%	1	0.1572	2.482	0.295	0.479	0.1541	4.209	0.286	0.369	0.1508	7.6	0.308	0.336
	2	0.1577	2.508	0.330	0.532	0.1541	4.221	0.310	0.378	0.1514	7.435	0.316	0.35
	3	0.1578	2.758	0.346	0.571	0.1544	4.467	0.342	0.388	0.1515	7.891	0.354	0.343
20%	1	0.1572	2.631	0.295	0.538	0.1541	4.098	0.279	0.368	0.1510	7.473	0.310	0.329
	2	0.1578	2.651	0.327	0.58	0.1548	4.270	0.307	0.396	0.1516	7.722	0.319	0.355
	3	0.1577	2.697	0.358	0.549	0.1546	4.284	0.359	0.372	0.1523	7.702	0.355	0.354
30%	1	0.1572	2.492	0.292	0.563	0.1541	4.137	0.284	0.386	0.1511	7.45	0.309	0.353
	2	0.1578	2.686	0.349	0.603	0.1548	4.087	0.309	0.407	0.1516	7.797	0.322	0.344
	3	0.1578	2.875	0.392	0.537	0.1548	4.389	0.362	0.427	0.1525	7.721	0.338	0.362
40%	1	0.1574	2.764	0.300	0.542	0.1543	4.409	0.291	0.376	0.1488	7.41	0.311	0.351
	2	0.1593	2.692	0.334	0.602	0.1549	4.087	0.314	0.41	0.1491	7.496	0.316	0.358
	3	0.1584	2.904	0.324	0.699	0.1550	4.395	0.346	0.386	0.1494	7.958	0.323	0.354
50%	1	0.1579	2.692	0.299	0.512	0.1538	4.287	0.291	0.379	0.1488	7.526	0.310	0.358
	2	0.1584	2.620	0.338	0.561	0.1549	4.154	0.311	0.377	0.1486	7.352	0.329	0.360
	3	0.1583	3.04	0.363	0.619	0.1547	4.460	0.347	0.429	0.1493	7.999	0.315	0.377

Table 9: Evaluation of the effectiveness of SecDefender against mid-round targeted label flipping attacks. The table compares attack success rates (ASR) for various LF scenarios.

$nLF \rightarrow$		Single-label			Double-label			Triple-label		
$Att_{ratio}\% \downarrow$	$m \downarrow$	HAR	Fashion-MNIST	FEDMNIST	HAR	Fashion-MNIST	FEDMNIST	HAR	Fashion-MNIST	FEDMNIST
0	0	0.1656	2.442	0.309	0.1656	4.116	0.209	0.1656	7.407	0.308
10%	1	0.156	2.391	0.29	0.1538	4.112	0.280	0.1507	7.28	0.3
	2	0.156	2.413	0.297	0.1538	4.099	0.288	0.1507	5.411	0.305
	3	0.1559	2.497	0.313	0.1539	4.114	0.292	0.1507	7.114	0.305
20%	1	0.1559	2.413	0.289	0.1538	4.067	0.280	0.1506	7.291	0.3
	2	0.1558	2.338	0.294	0.1539	4.064	0.284	0.1506	7.278	0.302
	3	0.1558	2.364	0.325	0.1539	4.159	0.296	0.1506	7.319	0.309
30%	1	0.156	2.408	0.288	0.1538	1.195	0.28	0.1506	7.222	0.3
	2	0.1559	2.387	0.299	0.1539	4.136	0.285	0.1506	7.264	0.303
	3	0.1558	2.422	0.32	0.1539	1.235	0.298	0.1507	7.319	0.301
40%	1	0.1559	2.27	0.287	0.1538	4.108	0.280	0.1508	7.266	0.309
	2	0.1558	2.390	0.291	0.1539	4.136	0.284	0.1508	7.362	0.304
	3	0.1558	2.326	0.324	0.1539	4.086	0.289	0.1508	7.364	0.309
50%	1	0.1559	2.416	0.287	0.1538	4.106	0.280	0.1508	7.305	0.301
	2	0.1558	2.413	0.293	0.1539	4.114	0.289	0.1508	7.173	0.303
	3	0.1558	2.384	0.314	0.154	4.145	0.293	0.1508	7.286	0.303

Table 10: Evaluation of the effectiveness of SecDefender against end-round targeted label flipping attacks. The table compares attack success rates (ASR) for various LF scenarios.

$nLF \rightarrow$		Single-label			Double-label			Triple-label		
$Att_{ratio}\% \downarrow$	$m \downarrow$	HAR	Fashion-MNIST	FEDMNIST	HAR	Fashion-MNIST	FEDMNIST	HAR	Fashion-MNIST	FEDMNIST
0	0	0.1656	2.442	0.309	0.1656	4.116	0.209	0.1656	7.407	0.308
10%	1	0.1571	2.414	0.292	0.1534	4.112	0.293	0.1507	5.411	0.301
	2	0.157	2.404	0.305	0.1534	4.075	0.296	0.1507	7.362	0.307
	3	0.157	2.414	0.324	0.1535	4.15	0.298	0.1508	7.287	0.331
20%	1	0.157	2.414	0.295	0.1534	4.069	0.292	0.1507	7.145	0.303
	2	0.157	2.390	0.293	0.1535	4.036	0.295	0.1506	7.307	0.305
	3	0.157	2.414	0.325	0.1536	4.054	0.297	0.1508	7.312	0.337
30%	1	0.157	2.346	0.293	0.1534	4.058	0.293	0.1506	7.196	0.304
	2	0.157	2.398	0.309	0.1535	4.14	0.295	0.1506	7.286	0.303
	3	0.157	2.436	0.311	0.1537	4.087	0.298	0.1507	7.3	0.307
40%	1	0.156	2.397	0.239	0.1534	4.082	0.285	0.149	7.228	0.303
	2	0.1558	2.385	0.303	0.1535	4.19	0.299	0.1507	7.3	0.303
	3	0.1558	2.403	0.303	0.1537	4.095	0.33	0.1508	7.3	0.309
50%	1	0.1559	2.354	0.294	0.1534	4.11	0.291	0.1506	7.318	0.302
	2	0.1558	2.402	0.305	0.1536	4.080	0.298	0.1487	7.3	0.304
	3	0.1557	2.413	0.309	0.1538	4.067	0.3	0.1495	7.322	0.302

Table 11: Evaluation of the effectiveness of SecDefender against full-round targeted label flipping attacks. The table compares target misclassification rates (TMR) for various LF scenarios.

$nLF \rightarrow$		Single-label				Double-label				Triple-label			
$Att_{ratio}\% \downarrow$	$m \downarrow$	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST	HAR	Fashion-MNIST	FEDMNIST	MNIST
0	0	0.8252	3.009	0.89	1.19	0.8252	10.558	0.949	1.06	0.8252	18.437	0.92	0.824
10%	1	0.7617	3.148	0.868	1.24	0.7582	10.381	0.923	0.1.003	0.762	18.639	0.916	0.889
	2	0.7643	3.1	0.874	1.28	0.7577	10.439	0.959	1.013	0.763	18.238	0.933	0.959
	3	0.7643	3.512	0.851	1.376	0.7574	10.696	0.990	1.011	0.7615	18.538	0.985	0.909
20%	1	0.7621	3.405	0.865	1.3	0.7580	10.654	0.918	1.004	0.7628	18.802	0.924	0.881
	2	0.7645	3.501	0.853	1.328	0.7578	10.709	0.966	1.039	0.763	18.196	0.937	0.989
	3	0.7625	3.554	0.946	1.307	0.7563	10.787	0.98	0.975	0.7617	19.152	0.990	0.97
30%	1	0.7629	3.247	0.860	1.3	0.7565	10.622	0.922	1.023	0.7629	18.744	0.928	0.897
	2	0.7657	3.547	0.905	1.371	0.7584	10.448	0.964	1.032	0.7616	18.964	0.942	0.969
	3	0.7623	3.139	0.974	1.371	0.7559	10.795	1.030	1.075	0.7610	18.9505	0.961	0.978
40%	1	0.7634	3.581	0.868	1.311	0.7567	10.618	0.929	0.987	0.7505	18.634	0.932	0.884
	2	0.7621	3.367	0.896	1.331	0.7577	10.191	0.969	1.035	0.75	18.531	0.936	0.998
	3	0.7646	3.751	0.888	1.402	0.7564	10.69	1.012	1.005	0.7466	19.043	0.942	0.965
50%	1	0.7635	3.367	0.864	1.242	0.7549	10.401	0.934	0.990	0.7503	18.51	0.942	0.894
	2	0.7662	3.310	0.900	1.308	0.7581	10.134	0.961	1.015	0.7471	18.960	0.943	0.992
	3	0.7639	3.825	0.935	1.322	0.7560	10.727	1.02	1.067	0.7472	19.229	0.908	0.978

Table 12: Evaluation of the effectiveness of SecDefender against mid-round targeted label flipping attacks. The table compares target misclassification rates (TMR) for various LF scenarios.

$nLF \rightarrow$		Single-label			Double-label			Triple-label		
$Att_{ratio}\% \downarrow$	$m \downarrow$	HAR	Fashion-MNIST	FEDMNIST	HAR	Fashion-MNIST	FEDMNIST	HAR	Fashion-MNIST	FEDMNIST
0	0	0.8252	3.009	0.89	0.8252	10.558	0.949	0.8252	18.437	0.92
10%	1	0.7622	3.017	0.864	0.7568	10.582	0.915	0.762	18.208	0.910
	2	0.762	3.04	0.868	0.7566	10.355	0.926	0.7619	18.405	0.92
	3	0.7621	3.1	0.876	0.7569	10.378	0.907	0.7617	18.378	0.899
20%	1	0.7621	3.04	0.862	0.7568	10.287	0.916	0.7618	18.415	0.91
	2	0.7619	2.931	0.863	0.7569	10.347	0.921	0.7616	18.372	0.913
	3	0.762	3.084	0.89	0.7569	10.317	0.91	0.7616	18.485	0.909
30%	1	0.7622	3.09	0.863	0.7569	10.365	0.913	0.7617	18.297	0.91
	2	0.7619	2.962	0.866	0.7568	10.334	0.922	0.7615	18.362	0.914
	3	0.7619	2.958	0.887	0.757	10.438	0.936	0.7615	18.288	0.893
40%	1	0.7622	2.851	0.86	0.7569	10.403	0.915	0.7603	18.313	0.91
	2	0.7621	2.995	0.86	0.7567	10.334	0.299	0.7602	18.125	0.915
	3	0.7622	2.904	0.888	0.757	10.255	0.926	0.7601	18.405	0.92
50%	1	0.762	2.993	0.859	0.7569	10.219	0.916	0.7306	18.463	0.910
	2	0.762	3.04	0.859	0.7568	10.300	0.928	0.7602	18.136	0.916
	3	0.7622	2.956	0.88	0.7517	10.407	0.928	0.76	18.333	0.897

Table 13: Evaluation of the effectiveness of SecDefender against end-round targeted label flipping attacks. The table compares target misclassification rates (TMR) for various LF scenarios.

$nLF \rightarrow$		Single-label			Double-label			Triple-label		
$Att_{ratio}\% \downarrow$	$m \downarrow$	HAR	Fashion-MNIST	FEDMNIST	HAR	Fashion-MNIST	FEDMNIST	HAR	Fashion-MNIST	FEDMNIST
0	0	0.8252	3.009	0.89	0.8252	10.558	0.949	0.8252	18.437	0.92
10%	1	0.7617	3.04	0.861	0.7618	10.621	0.912	0.7619	18.4	0.916
	2	0.7617	3.01	0.871	0.7619	10.586	0.909	0.7618	18.409	0.894
	3	0.7616	3.04	0.862	0.7619	10.719	0.905	0.7618	18.402	0.891
20%	1	0.7621	3.04	0.863	0.7617	10.638	0.911	0.7618	18.119	0.91
	2	0.7615	2.995	0.827	0.7618	10.497	0.926	0.7616	18.258	0.901
	3	0.7615	3.04	0.864	0.7618	10.510	0.918	0.7616	18.3	0.921
30%	1	0.7616	2.981	0.864	0.7616	10.5	0.913	0.7617	18.193	0.909
	2	0.7614	2.973	0.870	0.7618	10.622	0.933	0.7616	18.4	0.929
	3	0.7616	3.133	0.843	0.7616	10.597	0.902	0.7614	18.4	0.9
40%	1	0.7625	3.067	0.862	0.7616	10.57	0.919	0.7526	18.44	0.901
	2	0.7621	3.1	0.864	0.7618	10.59	0.909	0.7613	18.4	0.911
	3	0.7621	2.987	0.838	0.7616	10.644	0.951	0.7609	18.4	0.926
50%	1	0.7624	2.924	0.862	0.7616	10.607	0.934	0.7613	18.3	0.91
	2	0.7622	3.06	0.87	0.7617	10.634	0.961	0.7523	18.4	0.914
	3	0.7622	2.99	0.855	0.7614	10.5	0.970	0.7535	18.4	0.903