Post-Processing Matters: Additional Results

Anonymous Authors

I. Additional Results

Figures 1–9 report, for the explanation methods introduced in the main paper, the evaluation metrics considered in the main paper—namely $\mathtt{F}$, $\mathtt{R}^{n_g}$, $\mathtt{R}^{n_p}$, $\mathtt{R}^{n_b}$, and $\mathtt{R}^{a}$—as functions of the window size s. Results are shown across all datasets, with each figure corresponding to a different model. All metrics are computed using the same parameter settings adopted in the main paper. In the figures, red markers indicate values of s that maximize fidelity, while yellow markers—when present—denote configurations that allow for a 5% reduction in fidelity (in exchange for increased robustness).

Consistent with the trends discussed in the main paper, several patterns emerge. Applying post-processing ($s \neq 1$) leads to systematic trade-offs between fidelity and robustness: increasing s typically results in improvements in $\mathtt{R}^{n_g}$, $\mathtt{R}^{n_p}$, $\mathtt{R}^{n_b}$, and $\mathtt{R}^{a}$, as smoothing suppresses high-frequency, semantically irrelevant variations. At the same time, fidelity $\mathtt{F}$ generally improves for moderate values of s, often with an effect size comparable to—or larger than—that obtained by changing the underlying explanation method. For larger values of s, excessive smoothing progressively merges semantically distinct regions, yielding diminishing returns and eventually leading to saturation or degradation of $\mathtt{F}$.

We further analyze fidelity by varying the perturbation parameter p. Figures 10–12 illustrate the fidelity curves across different models and datasets, with each figure corresponding to a distinct model. For space reasons, we report only the top three explanation techniques ranked according to $\mathtt{F}$. The curves are shown for values of p ranging from 0 up to the value selected in the tables, following the criterion described in the main paper.

Consistent with the observations in the main paper, post-processing leads to improved fidelity in most settings. In cases where such improvements are limited, the fidelity of the original explanations typically degrades rapidly as p increases, leaving little room for further gains through post-processing.

While the fidelity curves in Figures 10–12 provide a detailed view of how post-processing affects individual explanation methods as the perturbation parameter p varies, we next move to a more compact, quantitative comparison across datasets and models. To this end, we summarize robustness results for the top-ranked explanation techniques in tabular form. Following the same criteria adopted in the main paper, Table I reports, for each dataset and model, the top three explanation techniques according to the $\mathtt{F}$ score. The parameter p in $\mathtt{F}$ is set as described in the main paper, while the parameter s is optimized following the same criterion used for the tables in the main paper.

For each technique, we report the method name, the corresponding $\mathtt{R}^{n_g}$ obtained using the brightness increase factor defined in the main paper, and the $\mathtt{R}^{n_g}$ obtained using the resizing pixel factor parameter selected as in the main paper. Results are provided for both the best-performing raw explanation and its post-processed counterpart for each explanation technique.

Table I illustrates that applying post-processing yields consistent improvements in robustness to natural noise across all evaluated configurations. In all settings, explanation methods that are less competitive in their raw form achieve higher robustness scores than the strongest non–post-processed explanations once post-processing is applied. This observation suggests that post-processing can be a practical and computationally lightweight alternative to replacing the underlying explanation method.

Beyond individual explanations, we also investigate whether combining multiple explanation methods can further improve robustness. Table II reports results for explanation ensembling, both with and without post-processing. Table II summarizes the results for the best-performing individual explanation ("best expl"), selected according to the procedure described in the main paper, together with those obtained using the ensembling strategies W1 and W2. For the W2 ensemble, the top-k explanations are selected following the same protocol adopted in the main paper. We additionally report results obtained by applying the post-processing strategy that achieves the highest fidelity, with the parameter s tuned according to the criteria specified in the main paper. Table II, report$\mathtt{R}^{n_b}$ and $\mathtt{R}^{n_p}$ metrics, computed with the same parameter settings as in the main paper.

Table II indicates that, in some cases, the best individual explanation attains higher robustness than the corresponding ensemble, suggesting that ensembling is not uniformly beneficial. Consistent with the trends observed in the main paper, performance improvements are more often driven by post-processing than by ensembling alone. Moreover, ensembles constructed from post-processed explanations generally outperform those based on raw explanations, further underscoring the importance of post-processing for improving explanation quality.

Finally, Figure 13 provides a qualitative illustration of the effect of post-processing on explanation maps. As shown in the figure, post-processing leads to more spatially coherent explanations by more clearly delineating semantically related regions of interest. In particular, it helps suppress isolated, high-importance regions that are not associated with meaningful image structures, thereby improving the overall visual interpretability of the explanations.

TABLE I: $\mathtt{R}^{n_b}$, and $\mathtt{R}^{n_p}$ for the top three explanation techniques according to $\mathtt{F}$, with and without post-processing, across all datasets and models considered in the paper.

	best expl			**best expl post**			**2nd best expl**			**2nd best expl post**			**3rd best expl**			**3rd best expl post**		
Dataset	$\mathscr{E}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$	$\mathscr{E}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$	$\mathscr{E}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$	$\mathscr{E}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$	$\mathscr{E}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$	$\mathscr{E}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$
ResNet50																		
TissueMNIST	GGC	0.06	0.108	GGC	0.021	0.039	DL	0.205	0.353	DL	0.042	0.075	GB	0.151	0.263	GB	0.048	0.079
RetinaMNIST	SA	0.456	0.56	SA	0.032	0.038	IG	0.246	0.376	IG	0.02	0.03	SG	0.366	0.355	SG	0.234	0.217
DermaMNIST	GB	0.085	0.15	GB	0.022	0.032	GGC	0.033	0.063	GGC	0.01	0.017	SG	0.409	0.41	SG	0.228	0.231
PneumoniaMNIST	DL	0.162	0.279	DL	0.067	0.114	SG	0.509	0.509	SG	0.189	0.19	IG	0.263	0.422	IG	0.102	0.164
OctMNIST	DL	0.052	0.154	DL	0.006	0.02	IG	0.081	0.211	IG	0.012	0.03	GS	0.208	0.277	GS	0.034	0.042
OrgancMNIST	GB	0.123	0.219	GB	0.044	0.081	SA	0.296	0.344	SA	0.111	0.131	GGC	0.054	0.093	GGC	0.022	0.041
BreastMNIST	GGC	0.03	0.069	GGC	0.013	0.03	GB	0.073	0.197	GB	0.008	0.019	SG	0.34	0.338	SG	0.256	0.254
PathMNIST	GGC	0.059	0.119	GGC	0.011	0.02	GB	0.101	0.238	GB	0.014	0.023	GS	0.506	0.559	GS	0.039	0.041
BloodMNIST	GB	0.118	0.235	GB	0.043	0.079	SA	0.507	0.566	SA	0.176	0.193	GGC	0.053	0.099	GGC	0.026	0.045
DenseNet121																		
TissueMNIST	GGC	0.05	0.088	GGC	0.012	0.021	GB	0.184	0.309	GB	0.048	0.074	DL	0.272	0.419	DL	0.017	0.026
RetinaMNIST	GGC	0.028	0.059	GGC	0.004	0.009	DL	0.148	0.279	DL	0.022	0.042	IG	0.209	0.357	IG	0.018	0.03
DermaMNIST	GGC	0.068	0.117	GGC	0.015	0.021	GB	0.111	0.203	GB	0.023	0.034	SG	0.626	0.588	SG	0.342	0.323
PneumoniaMNIST	GGC	0.112	0.234	GGC	0.014	0.026	GB	0.162	0.35	GB	0.024	0.05	SG	0.546	0.545	SG	0.136	0.135
OctMNIST	DL	0.048	0.139	DL	0.006	0.017	IG	0.069	0.194	IG	0.01	0.026	GS	0.19	0.253	GS	0.032	0.038
OrgancMNIST	SA	0.387	0.501	SA	0.07	0.092	SG	0.517	0.516	SG	0.343	0.342	GB	0.132	0.216	GB	0.038	0.065
BreastMNIST	GGC	0.024	0.053	GGC	0.01	0.021	GB	0.058	0.141	GB	0.011	0.026	IG	0.123	0.261	IG	0.02	0.044
PathMNIST	GGC	0.058	0.11	GGC	0.019	0.022	GB	0.117	0.235	GB	0.05	0.065	DL	0.232	0.393	DL	0.09	0.141
BloodMNIST	GGC	0.06	0.118	GGC	0.025	0.045	GB	0.126	0.246	GB	0.013	0.018	SG	0.568	0.566	SG	0.219	0.204
RegNet																		
TissueMNIST	GGC	0.056	0.111	GGC	0.013	0.027	SG	0.456	0.456	SG	0.267	0.267	GB	0.114	0.218	GB	0.039	0.071
RetinaMNIST	SG	0.448	0.447	SG	0.288	0.287	SA	0.308	0.482	SA	0.067	0.102	DL	0.143	0.325	DL	0.042	0.091
DermaMNIST	GGC	0.077	0.113	GGC	0.036	0.042	GB	0.115	0.163	GB	0.052	0.055	SA	0.452	0.517	SA	0.033	0.035
PneumoniaMNIST	GGC	0.091	0.181	GGC	0.034	0.062	GB	0.11	0.219	GB	0.039	0.067	SG	0.278	0.278	SG	0.174	0.174
OctMNIST	GGC	0.006	0.017	GGC	0.001	0.004	GB	0.029	0.083	GB	0.003	0.008	DL	0.053	0.174	DL	0.011	0.038
OrgancMNIST	SG	0.358	0.359	SG	0.225	0.225	GB	0.122	0.212	GB	0.038	0.06	SA	0.327	0.455	SA	0.06	0.085
BreastMNIST	GGC	0.027	0.062	GGC	0.004	0.008	GB	0.055	0.128	GB	0.009	0.017	DL	0.102	0.24	DL	0.01	0.023
PathMNIST	SG	0.441	0.441	SG	0.231	0.229	GGC	0.063	0.119	GGC	0.018	0.016	SA	0.378	0.506	SA	0.031	0.037
BloodMNIST	SG	0.444	0.442	SG	0.266	0.264	GGC	0.078	0.156	GGC	0.029	0.054	GB	0.123	0.24	GB	0.041	0.073

TABLE II: $\mathtt{R}^{n_b}$, and $\mathtt{R}^{n_p}$ for the best individual explanation ("best expl") according to $\mathtt{F}$, together with the results obtained using W1 or W2 with top-k explanations. Results are also reported for the case in which the best post-processing technique is applied.

	best expl		**W1**		**W2-top2**		**W2-top3**		**best expl post**		**W1 post**		**W2-top2 post**		**W2-top3 post**	
Dataset	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$	$\mathtt{R}^{n_b}$	$\mathtt{R}^{n_p}$
ResNet50																
TissueMNIST	0.060	0.108	0.146	0.205	0.104	0.181	0.151	0.181	0.021	0.039	0.043	0.057	0.014	0.027	0.014	0.026
RetinaMNIST	0.456	0.560	0.145	0.189	0.181	0.227	0.142	0.195	0.018	0.030	0.047	0.056	0.125	0.130	0.089	0.099
DermaMNIST	0.085	0.150	0.197	0.224	0.057	0.102	0.179	0.205	0.010	0.017	0.056	0.062	0.056	0.085	0.039	0.058
PneumoniaMNIST	0.162	0.279	0.142	0.183	0.149	0.254	0.118	0.203	0.127	0.148	0.039	0.050	0.141	0.224	0.098	0.156
OctMNIST	0.052	0.154	0.096	0.157	0.016	0.063	0.023	0.073	0.006	0.020	0.035	0.052	0.002	0.008	0.037	0.039
OrganMNIST	0.123	0.219	0.122	0.148	0.166	0.211	0.120	0.161	0.111	0.131	0.050	0.055	0.048	0.057	0.061	0.072
BreastMNIST	0.030	0.069	0.081	0.109	0.049	0.128	0.054	0.119	0.013	0.030	0.035	0.040	0.020	0.045	0.015	0.033
PathMNIST	0.059	0.119	0.179	0.218	0.075	0.169	0.103	0.183	0.037	0.053	0.069	0.072	0.020	0.030	0.030	0.044
BloodMNIST	0.118	0.235	0.186	0.218	0.215	0.195	0.156	0.171	0.176	0.193	0.067	0.069	0.212	0.185	0.145	0.135
DenseNet121																
TissueMNIST	0.050	0.088	0.186	0.244	0.115	0.194	0.189	0.217	0.012	0.021	0.061	0.073	0.009	0.017	0.011	0.019
RetinaMNIST	0.028	0.059	0.140	0.196	0.045	0.098	0.081	0.144	0.018	0.030	0.045	0.059	0.073	0.116	0.085	0.104
DermaMNIST	0.068	0.117	0.201	0.238	0.087	0.159	0.127	0.187	0.015	0.021	0.061	0.068	0.057	0.055	0.090	0.083
PneumoniaMNIST	0.112	0.234	0.156	0.230	0.131	0.238	0.124	0.242	0.136	0.135	0.042	0.061	0.056	0.118	0.040	0.082
OctMNIST	0.048	0.139	0.090	0.149	0.037	0.103	0.033	0.091	0.006	0.017	0.033	0.050	0.047	0.050	0.076	0.077
OrganMNIST	0.387	0.501	0.167	nan	0.274	0.290	0.195	0.215	0.070	0.092	0.054	nan	0.051	0.068	0.038	0.050
BreastMNIST	0.024	0.053	0.096	0.131	0.039	0.093	0.086	0.125	0.010	0.021	0.048	0.055	0.004	0.008	0.005	0.011
PathMNIST	0.058	0.110	0.167	0.232	0.083	0.166	0.102	0.183	0.019	0.022	0.057	0.069	0.022	0.026	0.030	0.047
BloodMNIST	0.060	0.118	0.175	0.226	0.089	0.175	0.203	0.231	0.025	0.045	0.058	0.068	0.034	0.063	0.029	0.053
RegNet																
TissueMNIST	0.056	0.111	0.137	0.203	0.240	0.260	0.169	0.194	0.013	0.027	0.043	0.057	0.134	0.134	0.090	0.091
RetinaMNIST	0.448	0.447	0.124	0.198	0.098	0.198	0.134	0.232	0.288	0.287	0.045	0.063	0.214	0.213	0.223	0.222
DermaMNIST	0.077	0.113	0.181	0.226	0.092	0.135	0.127	0.181	0.052	0.055	0.058	0.067	0.044	0.038	0.101	0.097
PneumoniaMNIST	0.091	0.181	0.118	0.190	0.107	0.221	0.095	0.190	0.112	0.156	0.036	0.054	0.107	0.221	0.085	0.179
OctMNIST	0.006	0.017	0.075	0.118	0.211	0.212	0.141	0.148	0.001	0.004	0.033	0.042	0.195	0.198	0.203	0.206
OrganMNIST	0.358	0.359	0.136	0.198	0.195	0.216	0.148	0.181	0.060	0.085	0.041	0.055	0.268	0.269	0.180	0.182
BreastMNIST	0.027	0.062	0.075	0.126	0.040	0.092	0.047	0.107	0.004	0.008	0.027	0.039	0.004	0.009	0.006	0.011
PathMNIST	0.441	0.441	0.150	0.221	0.225	0.233	0.201	0.232	0.231	0.229	0.048	0.063	0.218	0.216	0.229	0.227
BloodMNIST	0.444	0.442	0.149	0.230	0.229	0.241	0.168	0.205	0.134	0.184	0.049	0.066	0.101	0.100	0.123	0.122

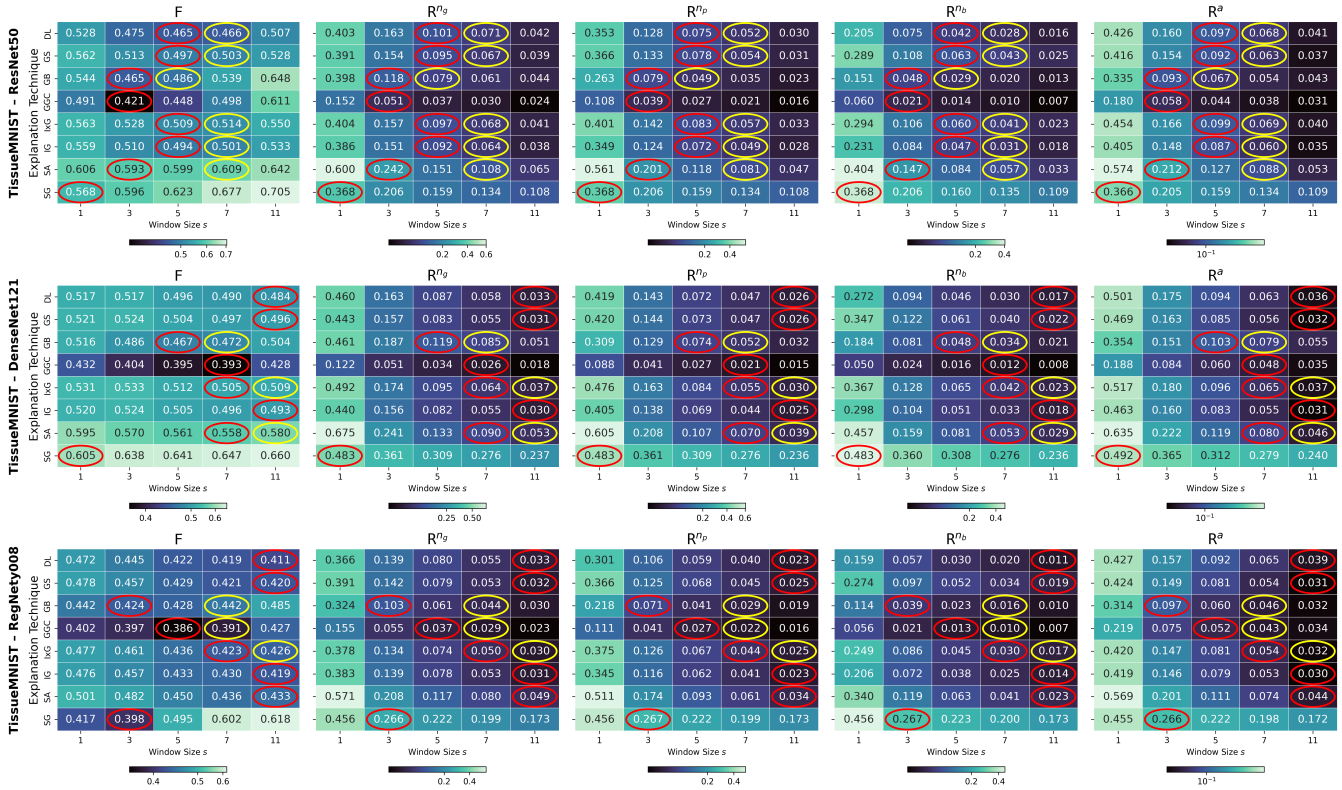

Fig. 1: *TissueMNIST* dataset and all models: impact of s on F, R^{n_g}, R^{n_p}, R^{n_b}, and R^a for the different explanation techniques.

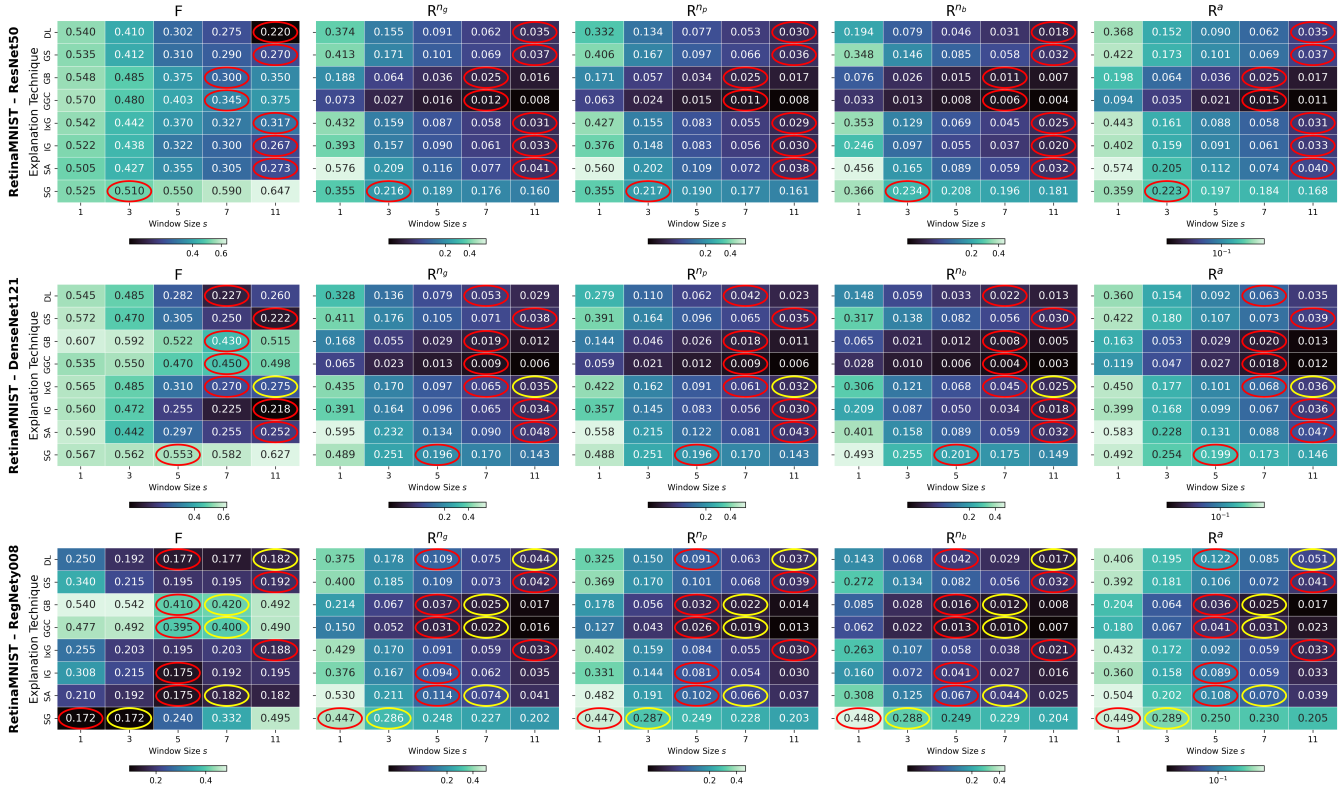

Fig. 2: *RetinaMNIST* dataset and all models: impact of s on F, R^{n_g}, R^{n_p}, R^{n_b}, and R^a for the different explanation techniques.

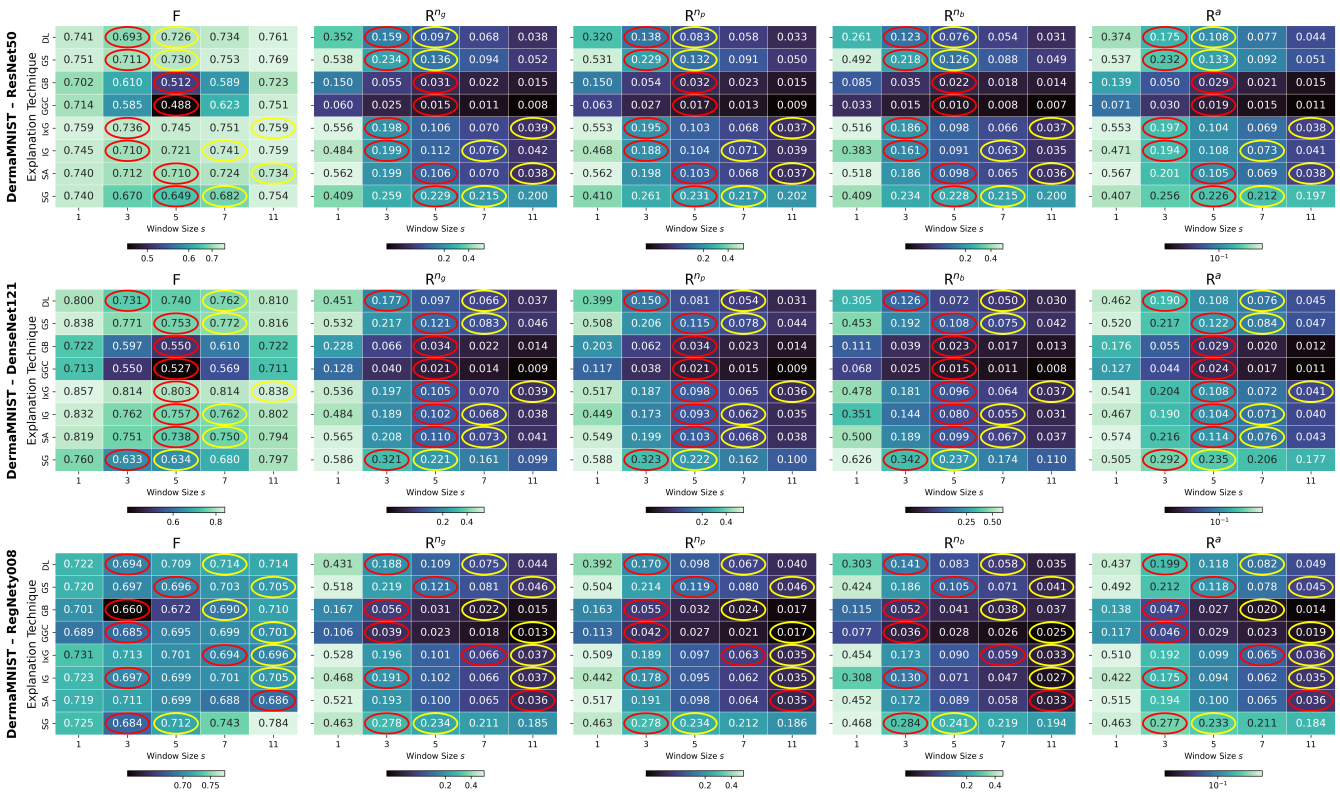

Fig. 3: *DermaMNIST* dataset and all models: impact of s on $\mathtt{F}$, $\mathtt{R}^{n_g}$, $\mathtt{R}^{n_p}$, $\mathtt{R}^{n_b}$, and $\mathtt{R}^{a}$ for the different explanation techniques.

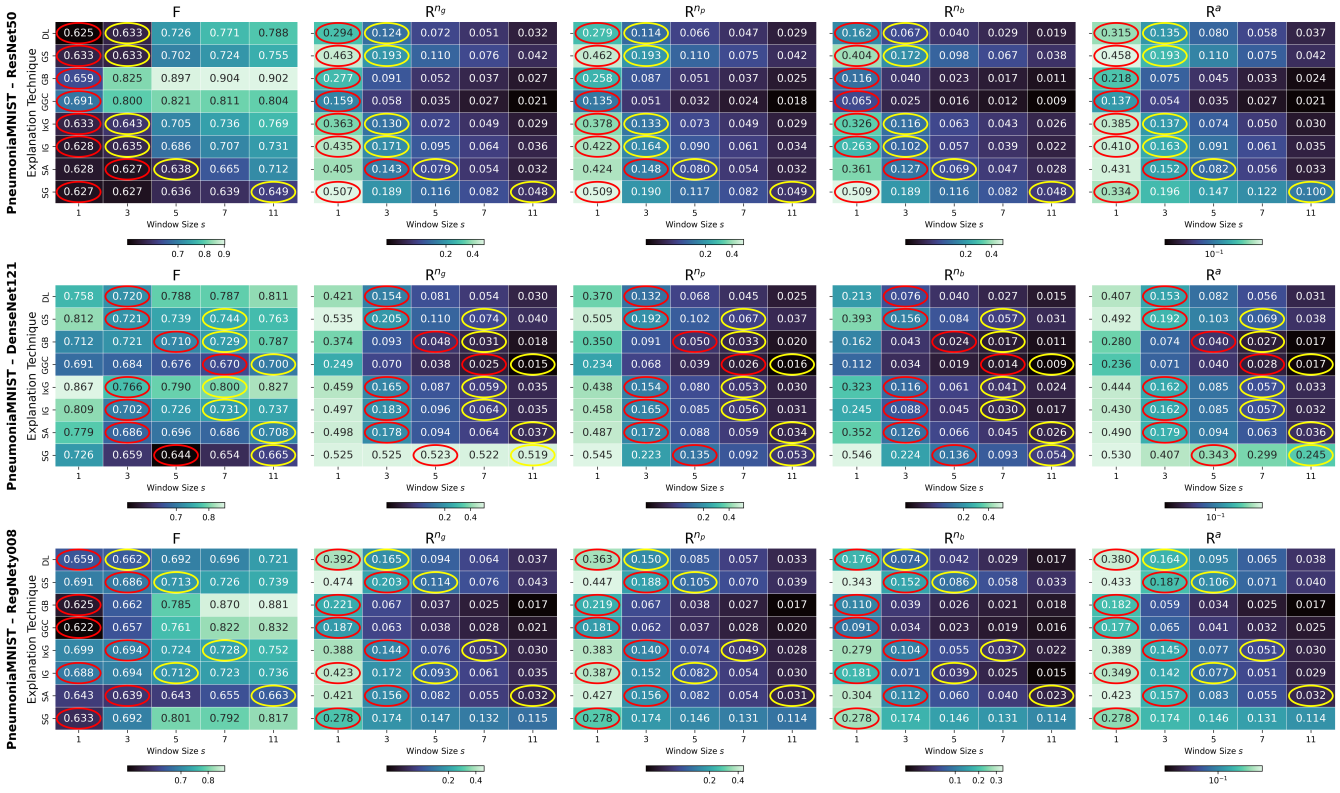

Fig. 4: *PneumoniaMNIST* dataset and all models: impact of s on $\mathtt{F}$, $\mathtt{R}^{n_g}$, $\mathtt{R}^{n_p}$, $\mathtt{R}^{n_b}$, and $\mathtt{R}^{a}$ for the different explanation techniques.

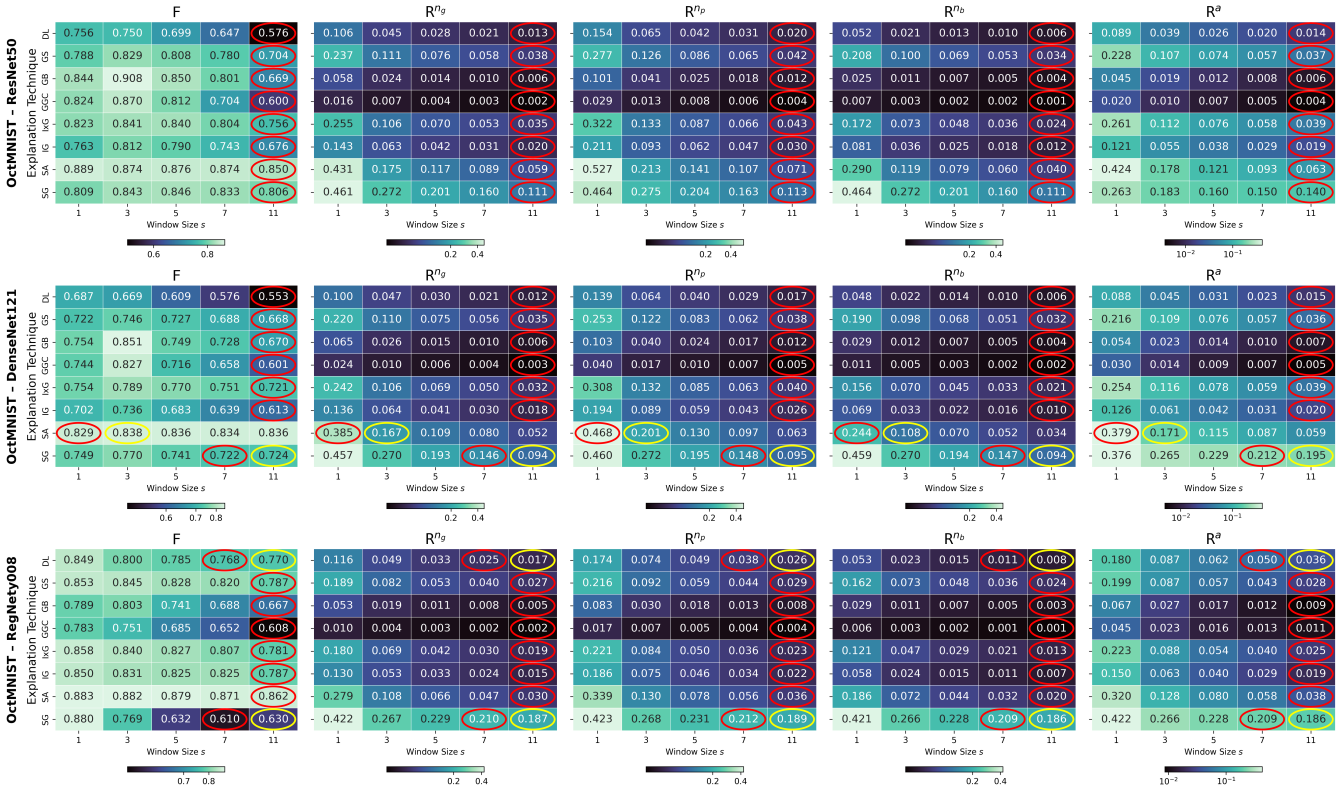

Fig. 5: *OctMNIST* dataset and all models: impact of s on F, R^{n_g}, R^{n_p}, R^{n_b}, and R^a for the different explanation techniques.

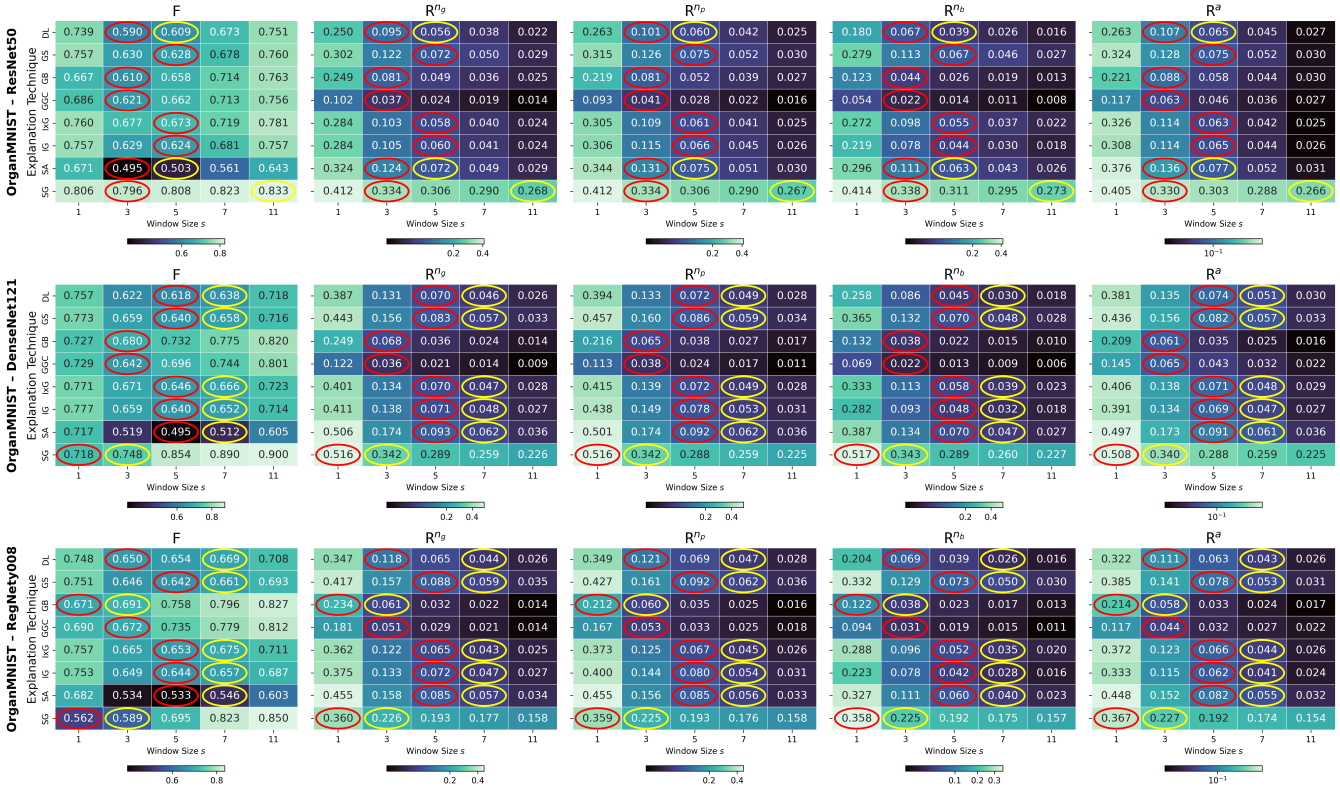

Fig. 6: *OrgancMNIST* dataset and all models: impact of s on F, R^{n_g}, R^{n_p}, R^{n_b}, and R^a for the different explanation techniques.

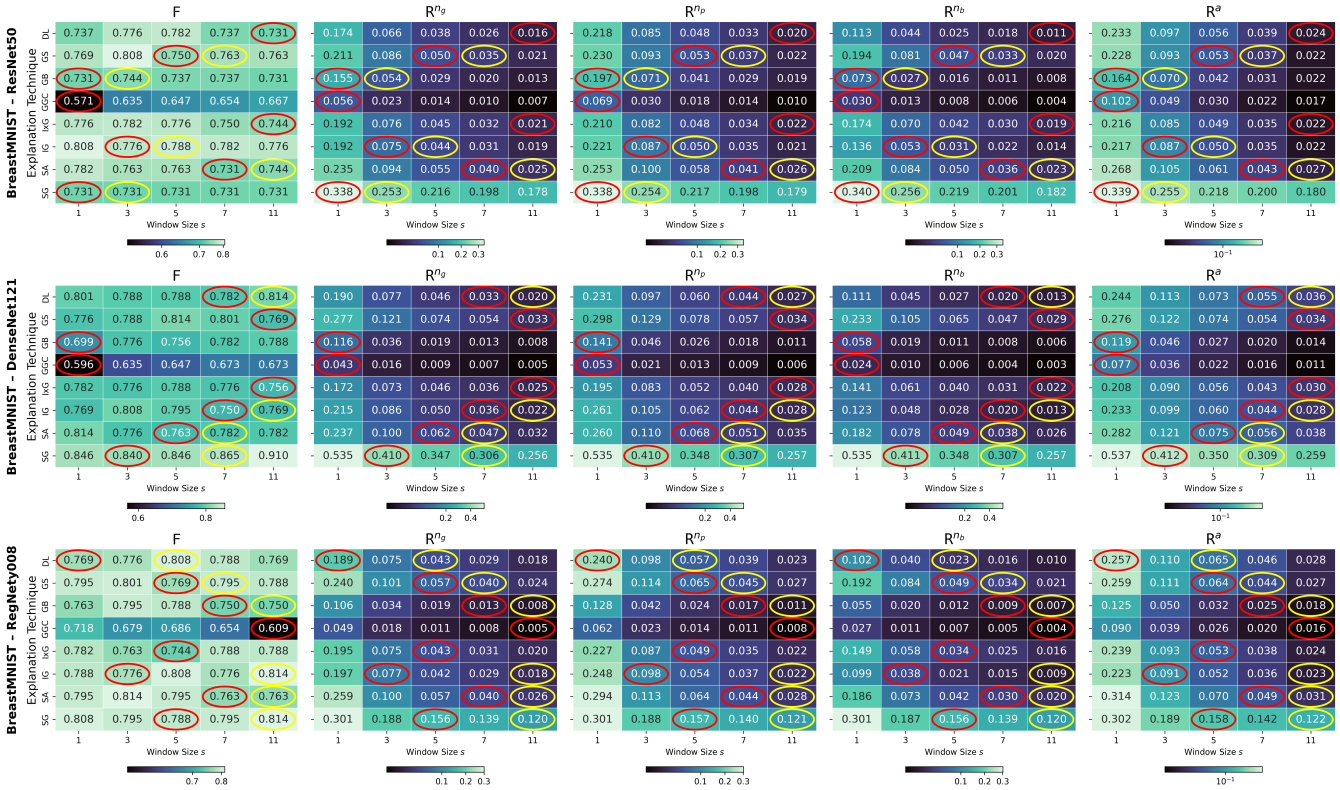

Fig. 7: *BreastMNIST* dataset and all models: impact of s on $\mathtt{F}$, $\mathtt{R}^{n_g}$, $\mathtt{R}^{n_p}$, $\mathtt{R}^{n_b}$, and $\mathtt{R}^{a}$ for the different explanation techniques.

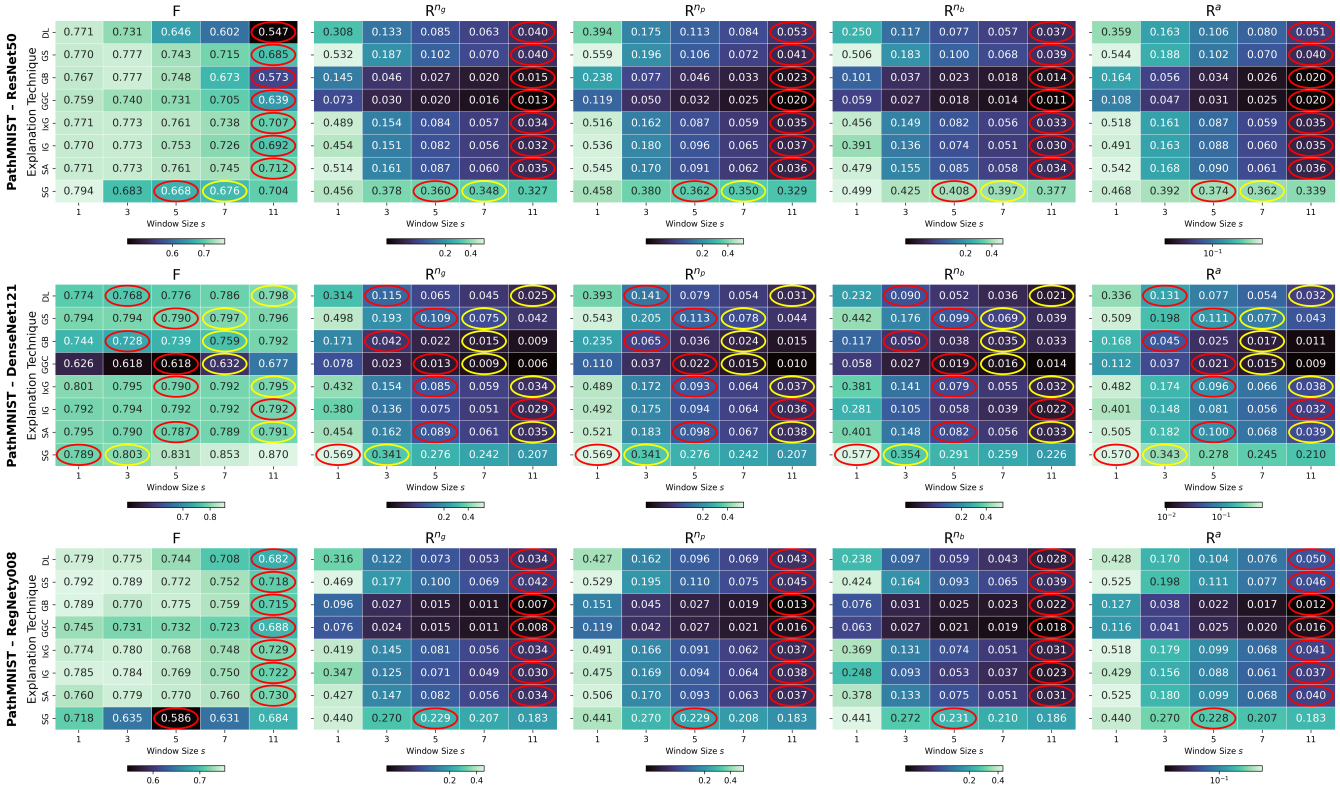

Fig. 8: *PathMNIST* dataset and all models: impact of s on $\mathtt{F}$, $\mathtt{R}^{n_g}$, $\mathtt{R}^{n_p}$, $\mathtt{R}^{n_b}$, and $\mathtt{R}^{a}$ for the different explanation techniques.

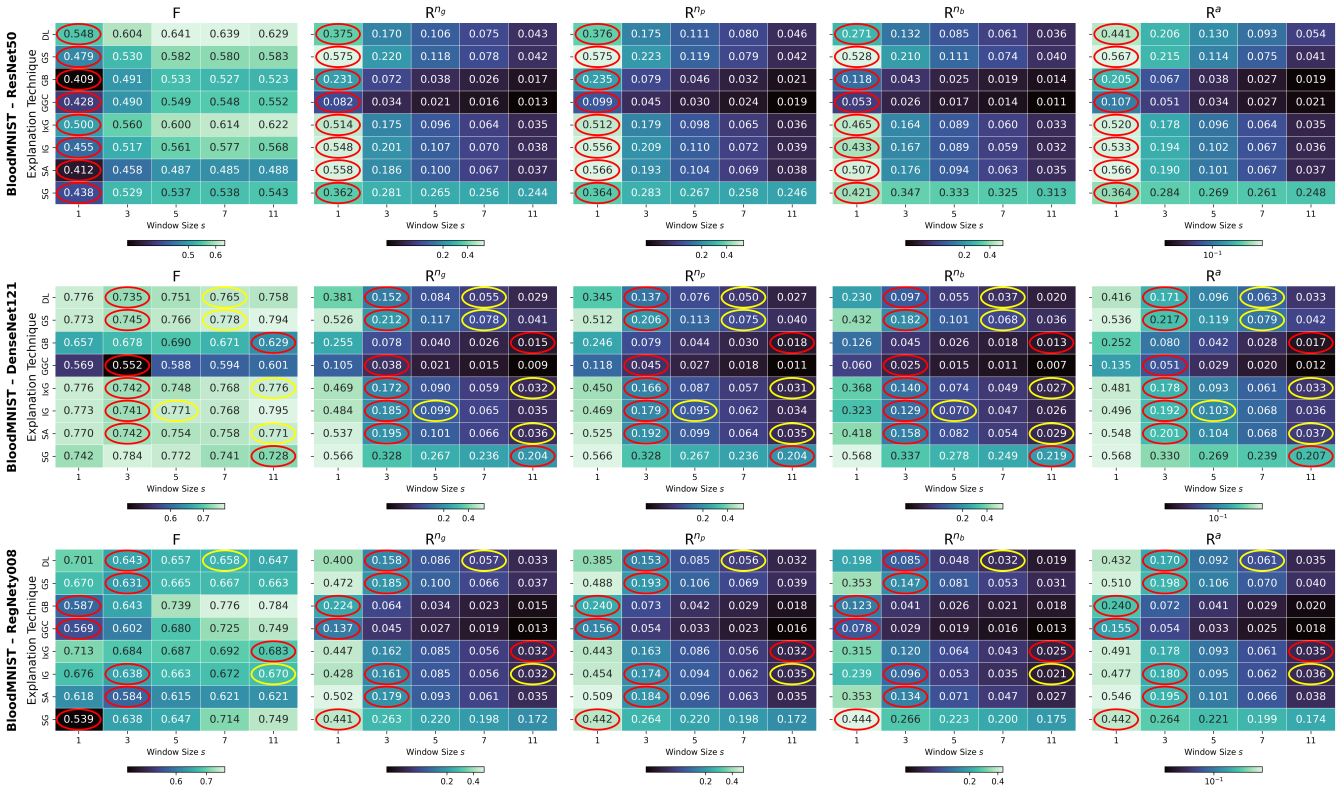

Fig. 9: *BloodMNIST* dataset and all models: impact of s on $\mathtt{F}$, $\mathtt{R}^{n_g}$, $\mathtt{R}^{n_p}$, $\mathtt{R}^{n_b}$, and $\mathtt{R}^{a}$ for the different explanation techniques.

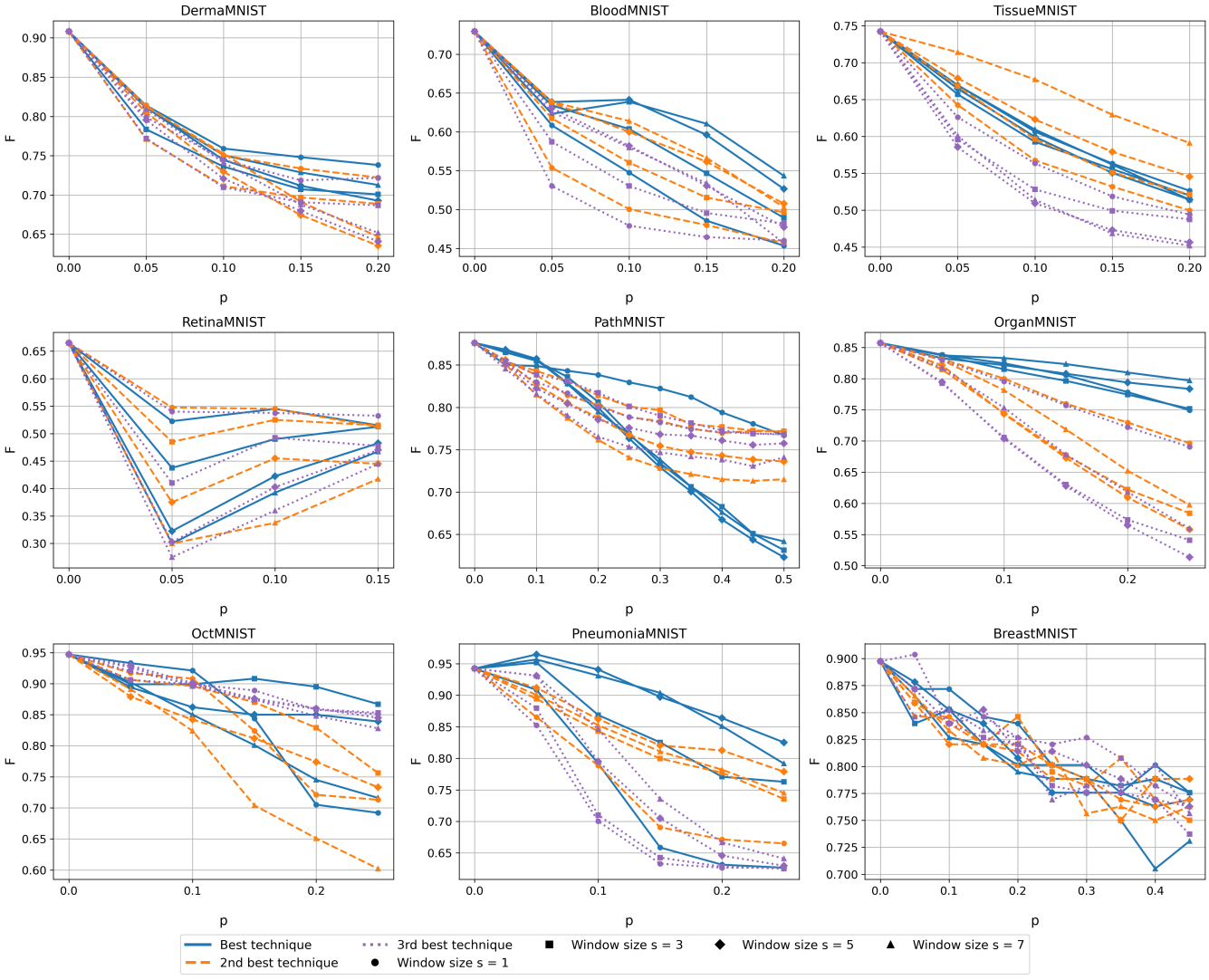

Fig. 10: ResNet50, all datasets: impact of s, p for F for the top-3 (according to F) explanation techniques.

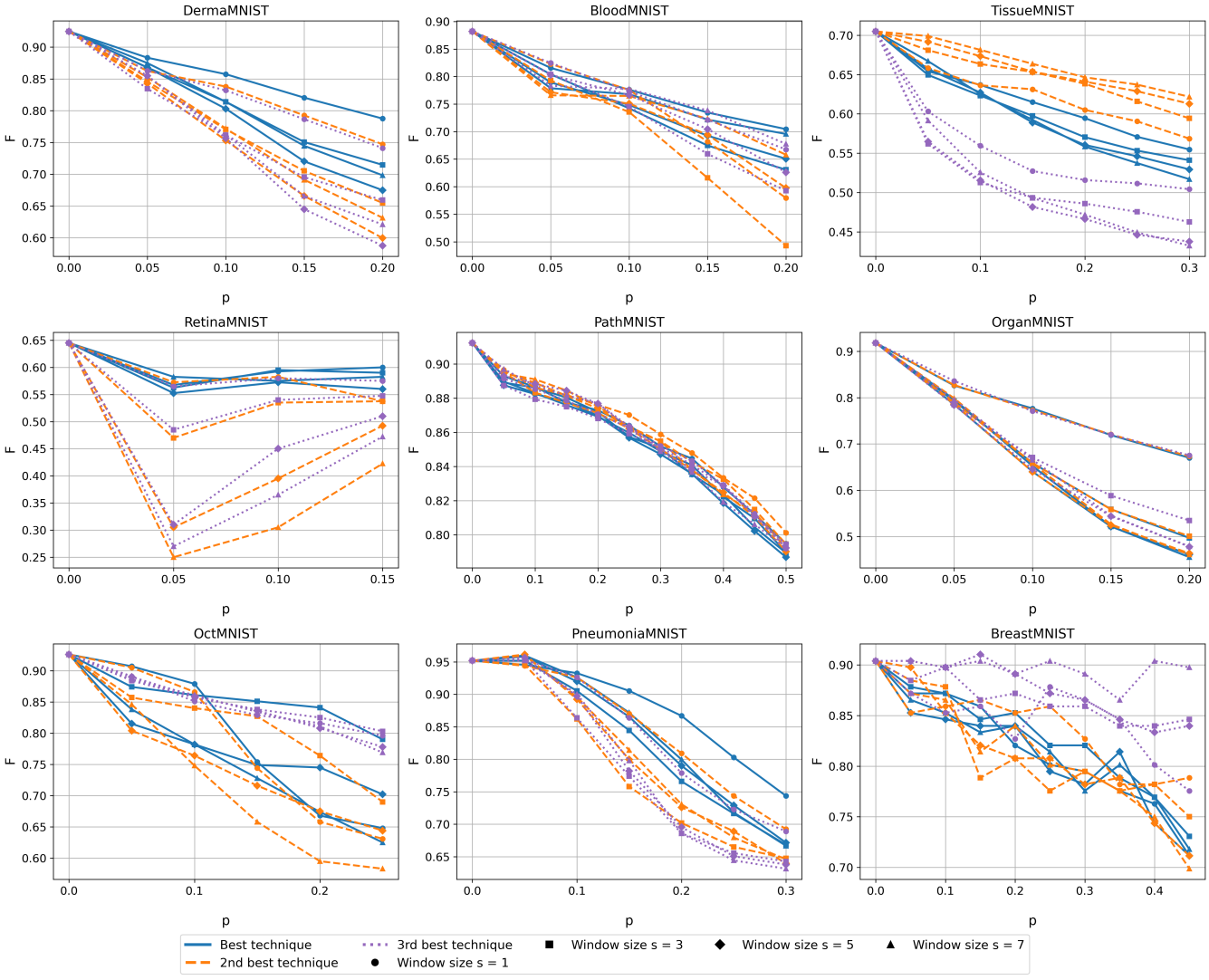

Fig. 11: DenseNet121, all datasets: impact of s, p for F for the top-3 (according to F) explanation techniques.

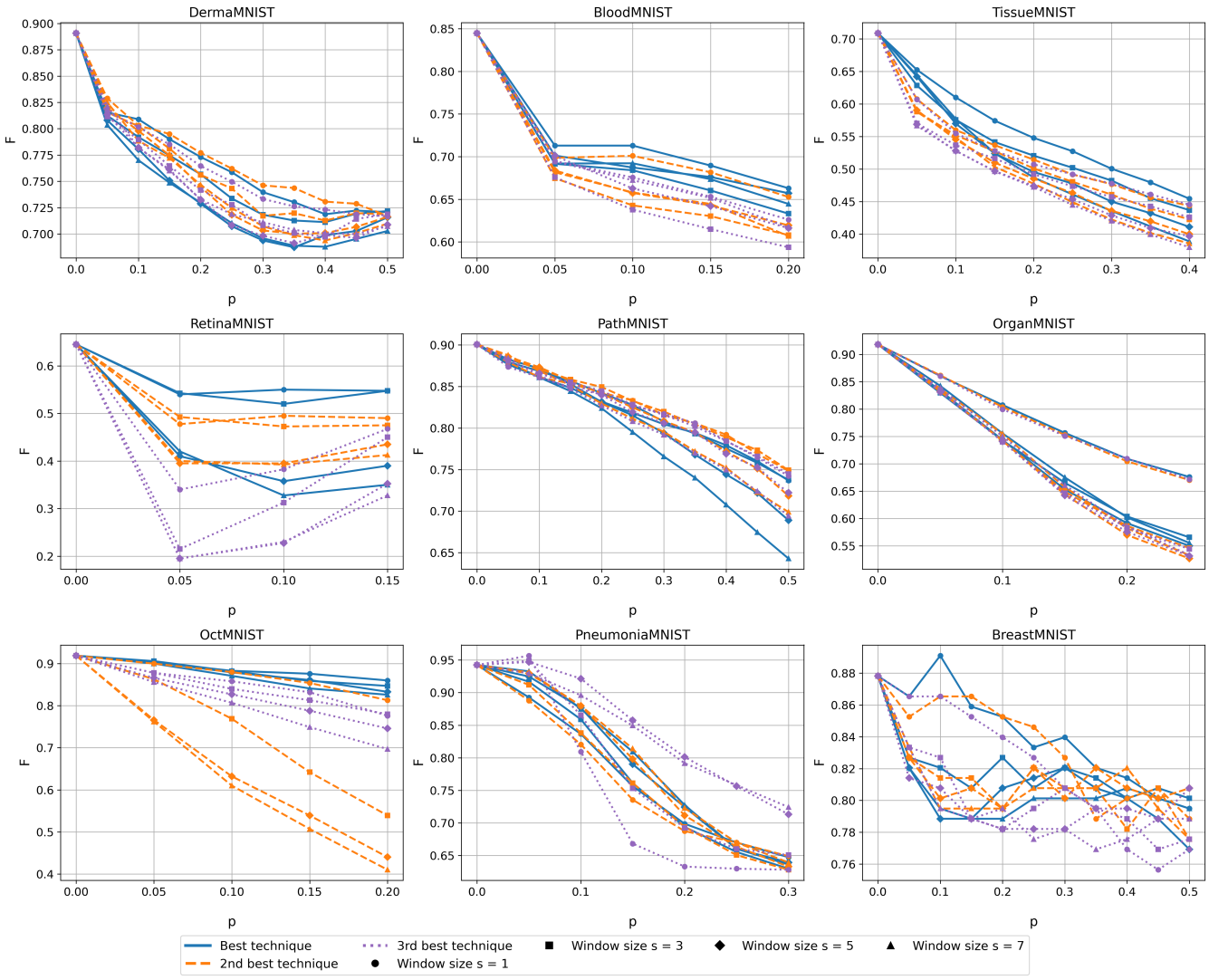

Fig. 12: RegNety008, all datasets: impact of s, p for F for the top-3 (according to F) explanation techniques.

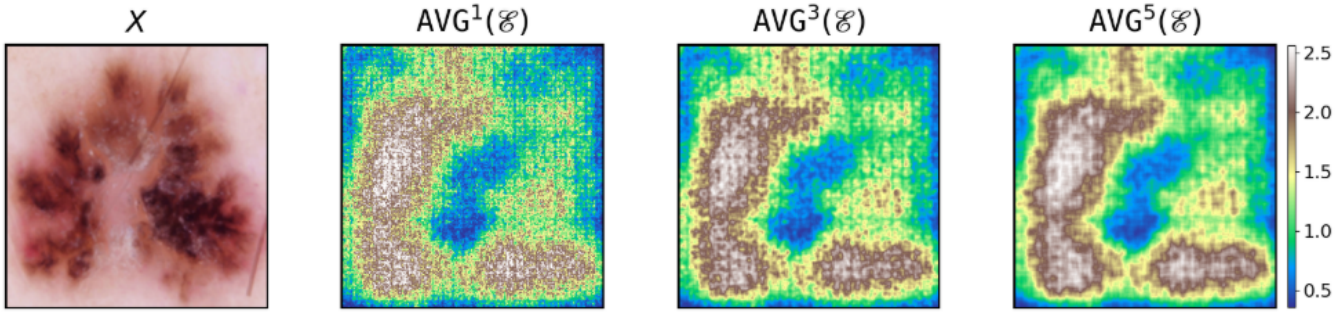

Fig. 13: Example of smoothed explanation for *DermaMNIST* dataset and ResNet50 model with IntegratedGradients.