Supplementary Material

Contribution Statement

VEG, CW, JF, CK, CL, and DK collaboratively conceptualized the research question, designed the methodology and the principal experiments. VEG conducted the experiments on OoD-detection and FD and handled statistical evaluation. CW collected and defined aggregation strategies and devised spatial measures. JF co-prototyped and ablated the meta-aggregation strategy, led result reporting and visualization, and released the package. CK analyzed key properties and pitfalls of aggregation strategies and developed the aggregation code-base. VEG, CW, and CK collected datasets and evaluated pixelwise uncertainty maps. CL provided code and data for the CAR and LIDC datasets. MP and SG provided valuable feedback on the design of the GMM and the statistical evaluation. JLR contributed to the code and prepared the repository for publication. VEG, CW, JF, JLR, CK, CL and DK contributed to manuscript preparation. VEG, CK, CL, and DK supervised the project and KMH provided overall guidance. All authors reviewed and approved the final version of the manuscript.

A. Formal properties of intensity-based aggregation strategies

In this section we give formal definitions of relevant properties of intensity-based AggSs. The properties *Monotonicity* and *Proportion invariance* have already been mentioned in the main text where we showed how they impact downstream performance. Here, we additionally discuss *Parameter independence* and *Locality*.

When an intesity-based AggS exhibits a given property, we provide a formal proof or justification. Conversely, if an AggS does not satisfy a property, we present a counterexample. The counterexamples are derived from experiments on simplified toy datasets—designed to be idealized test cases. If an AggS fails under these conditions, it is likely to be even more unreliable in real-world applications.

	PF	M	PI	L
Global Average (AVG)	✓	✓	✗	✗
Above-Threshold Average (ATA)	✗	✗	✓	✗
Above-Quantile Average (AQA)	✗	✓	✗	✗
Patch-Level Maximum (PLM)	✗	✓	✓	✓
Weighted Class Average (WCA)	✗	✓	✗*	✗

Table A.1. **Overview of key properties of selected AggSs.** PF: Parameter-free, M: Monotonic, PI: Proportion invariant, L: Local. *WCA becomes proportion invariant when the background class is excluded in specific use cases where foreground classes exhibit high uncertainty and the background class has low uncertainty.

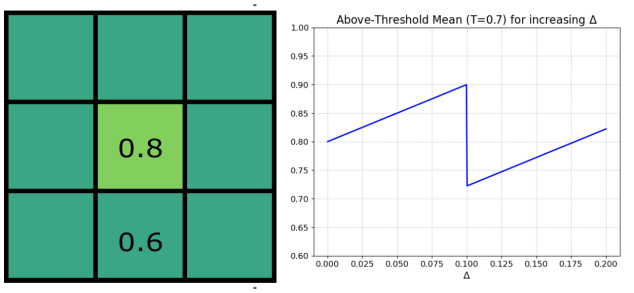

Figure A.1. **Non-monotonicity of ATA.** Left is shown a 3×3 uncertainty map. If we increase each pixel value by Δ the ATA score (threshold $T = 0.7$) will drop at a value of $\Delta = 0.1$, since the amount of pixels affecting the average suddenly increases as their values pass the threshold.

A.1. Parameter-fee property

AVG clearly is parameter-free. ATA depends on the fixed threshold $T > 0$ and is only applicable if T is less than the pixelwise maximum of the heatmap U.

AQA depends on the fixed portion $q \in (0, 1)$ of lower uncertainty values above which all uncertainties are averaged.

PLM depends on the choice of patch size. While too small patch sizes are prone to outliers (*e.g.* for a 1×1 patch size PLM is equivalent to the pixelwise maximum), too large patch sizes approximate AVG and thus might return lower averages than expected.

Weighted Class Average (WCA) technically depends on the choice of weights w_c for the classes. However, the most important examples Single-Class Average (SCA), Balanced Class Average (BCA) and Imbalanced Class Average (ICA) are parameter-free since the choice of weights is fixed.

A.2. Monotonicity

We call an AggS *monotonic* if an increase in all pixelwise uncertainty scores leads to an increase in the aggregated score. More formally, AggS f is monotonic if and only if $f(U) \leq f(V)$ for any uncertainty maps U, V satisfying $u_i \leq v_i$ for each pixel $i = 1, \ldots, mn$.

This property ensures that the aggregated scores reliably and intuitively track increases in pixelwise uncertainty. Suppose, *e.g.* that an OoD sample would lead to a slight increase in uncertainty distributed equally across all pixels of the uncertainty map. A non-monotonic AggS might assign similar aggregated scores to both the iD and the OoD sample making them impossible to distinguish despite a clear difference in pixelwise uncertainty.

AVG is monotonic since if we have two uncertainty maps U, V of shape $m \times n$ satisfying $u_i \leq v_i$ for all pixels $i =$

$1, \dots, mn$ then for the AVG holds

$$\text{AVG}(U) = \frac{1}{mn} \sum_i u_i \leq \frac{1}{mn} \sum_i v_i = \text{AVG}(V) \quad (1)$$

This implies that PLM is also monotonic, since it computes the average w.r.t. to patches of fixed size.

To prove that AQA is monotonic, consider the lists of pixelwise uncertainty values of U resp. V *sorted ascendingly*: $u^1 \leq \cdots \leq u^{mn}$ resp. $v^1 \leq \cdots \leq v^{mn}$. Our assumption that each pixelwise value of V is increased compared to U implies that for both lists we have $u^i \leq v^i$ for all $i = 1, \dots, mn$. Therefore, the q-quantile w.r.t. the values of V must be greater or equal to the q-quantile w.r.t. U. Furthermore, since both lists have the same number of uncertainty values lying above the q-quantile the average of those values for V will be greater or equal to the average for U.

For each choice of class weights w_c the WCA is monotonic as well: If $u_i \geq v_i$ for all pixels i for uncertainty maps U, V, then in particular for the class-wise average we have $\alpha_c^U \geq \alpha_c^V$ for each class c. For a fixed choice of weights w_c this implies

$$\text{WCA}(U) = \sum_c w_c \alpha_c^U \geq = \sum_c w_c \alpha_c^V = \text{WCA}(V) \quad (2)$$

In contrast, ATA is not monotonic, a counterexample is illustrated in Fig. 1b and Fig. A.1.

A.3. Proportion invariance

Consider idealized binary uncertainty maps having only low-uncertainty regions with uncertainty values approximately 0 (*e.g.* irrelevant background) and high-uncertainty regions with uncertainty values approximately 1 (*e.g.* relevant foreground).

An AggS f is *proportion invariant* if $f(U) = f(V)$ for any such uncertainty maps U, V only differing by the area proportions of the low- resp. high-uncertainty regions. A situation where this property would be desirable is the cropping of uncertainty maps: in this case we might want the aggregated uncertainty score to be unaffected when irrelevant low-uncertainty regions in the background are removed.

By this definition AVG is highly dependent on the present area-proportion, as it divides the sum over all pixelwise uncertainty by the total number of pixels.

AQA also depends on the proportion of low-uncertainty pixels as it considers the q highest uncertainty values in U If the proportion of high-uncertainty values drops below q, the selection of the top q values will inevitably include low-uncertainty pixels, resulting in a lower score. This effect is illustrated in Fig. 1 (b), where the presence of low-uncertainty pixels influences the final aggregation outcome. In contrast, PLM only depends on the uncertainty values within the maximal patch. Increasing or decreasing low-uncertainty proportion does not affect this patch and thus PLM remains unaffected.

Similarly, ATA only depends on high uncertainty values. Increasing or decreasing low-uncertainty proportion does not affect this set of high values (assuming that the threshold T is chosen sufficiently high to not capture low background uncertainty). As a result, ATA is proportion invariant.

The WCA is generally not proportion invariant, as changes in class proportions influence individual class averages, thereby affecting the overall weighted sum. However, in scenarios where high uncertainty is typically concentrated in foreground classes and low uncertainty is prevalent in the background the BCA and ICA remain proportion invariant in these settings, as they are unaffected by background proportion changes.

A.4. Locality

AVG is non-local as it is computed across all pixels. Similarly, ATA and AQA are non-local AggSs since the relevant uncertainty values above threshold T or q-quantile may occur at pixels across the whole image In contrast, PLM is local as its resulting score only depends on the uncertainty values of pixels within the maximal patch.

The WCA is generally not a local aggregation strategy, as it computes a weighted sum of class-wise uncertainty averages across all classes. Consequently, if all class weights are nonzero, every pixel in U contributes to the final score, making it inherently global rather than local. However, in the specific case of the SCA applied to a class that is a priori known to have spatially localized instances—meaning it occupies only a small proportion of the image—SCA can indeed exhibit local behavior.

Applicability to spatial measures We do not apply these properties to spatial measures because their values reflect local structural patterns in uncertainty maps and should not directly be compared to uncertainty intensities. Consequently, properties like monotonicity are irrelevant for spatial aggregators, as the increase of spatial measures is not always related to an increase of uncertainty. For instance, a global pixelwise increase of uncertainty can potentially decrease the entropy across an uncertainty map, in particular if the uncertainty values after the increase all lie within the same bin.

B. Details on spatial aggregation strategies

B.1. Selection of spatial measures

For our experiments, we use the following spatial measures: Moran's I [47], Edge Density Score [57], and Shannon Entropy [67].

Moran's I It measures spatial autocorrelation by comparing the similarity of values at neighboring pixels. A high

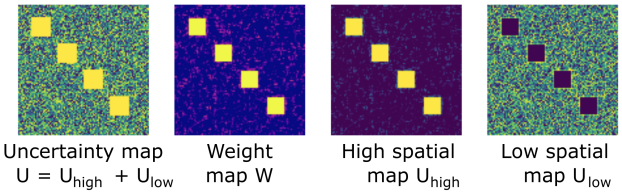

Figure B.1. **Spatial decomposition using Moran's I.** For an uncertainty map U we compute the pixelwise local spatial measure w.r.t. to Moran's I which captures noise ($I = 0$) and clusters ($I = 1$). The local spatial measures can be interpreted as weights of a weight matrix W where high weights correspond to high spatial auto-correlation. The pixelwise product of U and W is U_{high} while the pixelwise product of U and $1 - W$ is U_{low}, which yields a matrix decomposition of the original uncertainty map.

positive value indicates that similar uncertainty values cluster together, while a value near zero implies spatial randomness. In its original form, Moran's I ranges from -1 (negative correlation) to 1 (positive correlation). For our analysis we adapt Moran's I by capping negative values at 0, as negative spatial correlation (*e.g.* checkerboard patterns) is not to be expected in our data.

Edge Density Score It quantifies the presence of edges by computing the proportion of pixels within a local window whose gradient magnitude exceeds a specified threshold. It reflects how much spatial variation (*e.g.* sharp transitions or boundaries) is present in the uncertainty map.

Shannon Entropy It measures the local variability of uncertainty values by computing the Shannon entropy over discretized uncertainty levels within a local window. Higher entropy indicates more heterogeneous or noisy uncertainty, while lower values correspond to more uniform regions.

Both the Edge Density Score and entropy involve additional hyperparameters. The Edge Density Score uses a gradient threshold τ to classify pixels as edge-like; we set $\tau = 0.2$. Entropy is computed using $b = 4$ bins to discretize local uncertainty values. Another viable spatial measure is Geary's C [23], which captures local dissimilarity, with values close to 1 indicating randomness, below 1 indicating positive autocorrelation, and above 1 indicating negative autocorrelation. However, we omit it, as it provides information similar to Moran's I.

B.2. Spatial mass ratio

Given an uncertainty map U and a spatial measure with values between 0 and 1, we propose to calculate a *Spatial Mass Ratio (SMR)* which reflects how much of the total uncertainty is concentrated in spatially structured regions of the image.

For each pixel u of the (1-padded) uncertainty map U we first compute the pixelwise local spatial measure w_u within a sliding window of size 3×3 centered at u. The resulting matrix $W \in [0,1]^{m \times n}$ contains the local spatial measures for all pixels $u_i \in U$.

Using W we then compute a filtered uncertainty map U^{high} by pixelwise multiplying U with W, retaining only the uncertainty mass in regions of high local spatial measure. Similarly, multiplying U with $1 - W$ yields U^{low}, capturing the mass in regions of low local spatial measure. This results in a spatial decomposition of the uncertainty (see Figure B.1):

$$U = U^{\text{high}} + U^{\text{low}}.$$

Finally, we define the *Spatial Mass Ratio (SMR)* as the portion of uncertainty mass located in pixels with high spatial measure:

$$\text{SMR} := \frac{\sum_i^{nm} u_i w_i}{\sum_i^{nm} u_i} \in [0,1].$$

The behavior of the SMR can be intuitively understood by considering its extreme cases: For Moran's I, an SMR of 0 indicates that all uncertainty is located in noise-like, uncorrelated regions, while a value of 1 reflects fully clustered uncertainty. For Edge Density Score, SMR equals 0 when uncertainty lies entirely in flat areas and reaches 1 when it is concentrated along edges. For Entropy, an SMR of 0 means local uncertainty values fall within a single bin (i.e., low variability), whereas a value of 1 corresponds to a uniform distribution across bins, indicating high local randomness.

C. Details on Benchmarking Results

Table C.1 provides a more detailed analysis of the results shown in Figures 3 and 5b, reporting mean scores and standard deviations computed over 500 bootstrap samples of the evaluation data. This approach reduces the risk that observed performance is driven by sample variability in the test sets. As expected, standard deviations are larger when fewer evaluation samples are available (cf. Supp. D.7); however, the relative ranking of AggSs remains stable across datasets.

To assess the statistical significance of performance differences, we conducted a one-sided Wilcoxon signed-rank test [78] on the AUROC and E-AURC scores obtained from the bootstrapped samples across the 10 datasets. The null hypothesis ($\mathscr{H}_0$) assumes that aggregator f_A does not outperform f_B, while the alternative hypothesis ($\mathscr{H}_1$) asserts that f_A performs significantly better (i.e., higher AUROC or lower E-AURC), with a significance level of $\alpha = 0.05$. Full p-value matrices are omitted here for clarity but are available for AggSs on MCD heatmaps at https://github.com/Kainmueller-Lab/aggrigator_experiments.

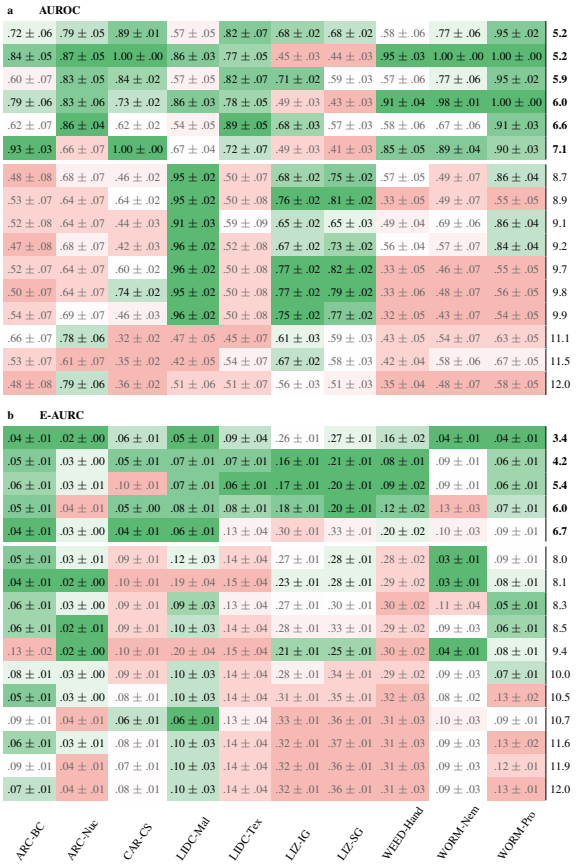

Table C.1. **Performance of AggSs on uncertainty maps generated with MC Dropout (MCD) in OoD and failure detection.** Columns are color-coded with a red-to-green gradient, where white represents the mean (for AUROC this is equivalent to random guessing). (a) Higher AUROC values (greener cells, 0.6–1.0) indicate better iD vs. OoD separation, while (b) lower E-AURC values (greener) indicate better alignment between uncertainty estimates and prediction errors. AggSs are ranked from best (top) to worst (bottom) based on their average metric, first computed across 500 bootstrap samples per dataset, and then averaged across datasets to ensure stable rankings (rightmost column). Each value is reported with the standard deviation across bootstrap samples. These tables complement Figures 3 and 5b.

To substantiate our claim that the variability in the performance of different AggSs arises from the task and dataset properties rather than the UQ methods used to generate the pixelwise uncertainties we have reproduced our main benchmarking results for additional UQ methods beyond Monte Carlo Dropout (MCD, [21]): Test Time Augmentation (TTA, [49]), Maximum Softmax Probability (MSP, [28]) and two ensembling approaches: Deep Ensembles (DE, [35], for CAR-CS, LIDC-Mal, and LIDC-Tex) and the computationally lighter Checkpoint Ensemble (CE, [11] for ARC-BC and ARC-Nuc.)

TTA selects augmentations at test time that best suit each dataset, avoiding to affect the model's predictive capacity and improving its ability to generalize. Tables C.3 and C.4 follow the same statistical protocol as Table C.1. Consistent with previous observations, standard deviations increase when fewer evaluation samples are available. Additionally, *GMM-All* and its ablated variants exhibit reduced performance on MSP heatmaps of LIZ-SG and LIZ-IG, suggesting that the underlying data modes are insufficiently separable in high-dimensional feature space.

While CE, DE, and TTA are available only for a subset of datasets, MSP provides a comprehensive overview that allows fair comparison with MCD. Table C.2 shows that even when the uncertainty estimate is heuristic (as in MSP, which relies solely on softmax scores and thus cannot capture OoD uncertainty), the relative performance of AggSs remains largely consistent with that observed under MCD. This suggests, at least empirically, that best practices for aggregation are better identified by evaluating AggSs across diverse datasets rather than across uncertainty methods.

C.1. On the AggSs performance in OoD Detection

As shown in Table C.2a, the top performers, based on their average rank, are the *prediction-based* aggregators, as well as the *GMM-All*, *GMM-Int* and *GMM-Spa* applied to MSP uncertainty maps. The conclusions drawn in Section 4.3 remain consistent, with the notable exception of the default aggregation strategy AVG: its mean rank improves from 12th to 7th due to stronger performance on ARC-BC, ARC-Nuc, WEED-Pro, and WEED-Nem. This behavior is largely driven by the synthetic nature of the ARC datasets, where MSP is more sensitive to induced and controlled noise, and by the label shift present in the WEED OoD variants, which likely reduces model confidence. In LIDC-Tex, MSP fails to detect OoD instances: the increased transparency of tumors should create elevated uncertainty along borders within the predicted masks, but the overconfident MSP does not capture this, leading to missed detections (with a consequent increase in E-AURC; cf. Table C.2b).

When the analysis is repeated for CE and DE, as reported in Table C.3a, the ranking aligns closely with that observed for MCD, reflecting the greater suitability of these techniques for capturing epistemic uncertainty. Results for TTA are available only for a smaller subset of datasets (cf. Table C.4a); therefore, it is not possible to determine whether uncertainty increases in OoD samples extend to the extreme tail of the uncertainty distribution, potentially affecting the performance of AQA 0.90. Nonetheless, we continue to observe the dominance of *prediction-based* aggregators and the *GMM-All* AggS, while *GMM-Spa* performs poorly in scenarios where spatial augmentations produce similar uncertainty maps for both iD and OoD samples. This is expected, for instance, in the three-label task learned on ARC-

Nuc, where removing nuclei intensity leads to fragmented border predictions; for both iD and OoD samples, the resulting uncertainty structure resembles the effects of rotations or crops.

Across UQ methods, the p-value matrices show that no single aggregator consistently dominates in AUROC, supporting our hypothesis that dataset-specific factors (*e.g.* object count, class imbalance) strongly affect the choice of AggS and no universally optimal method can be determined. Instead, a statistically significant top tier emerges: the *GMM*-based strategies and the weighted averaging approaches (BCA, ICA) significantly outperform all others ($p < 0.05$), although no method within this group is uniformly superior. While *GMM-All* often leads, its advantage over BCA or other GMM variants is not always significant. Lower-tier methods, such as threshold-based approaches (ATA, AQA), patch-based PLM, and the baseline AVG, are significantly outperformed in most comparisons, confirming their status as suboptimal choices.

Extending the analysis to MSP, TTA, and CE/DE largely reinforces these conclusions. The MSP baseline mirrors the MCD results, showing a clear separation ($p \ll 0.01$) between top-tier AggSs (*GMM*-based ones, BCA, ICA) and lower-tier methods. Interestingly, QFR, while dominant in Failure Detection, ranks lower here but still outperforms most lower-tier methods ($p < 0.05$), and under ensemble- and TTA-based uncertainty it becomes statistically indistinguishable from (or even superior to) *GMM-All* and BCA, suggesting that its foreground–background ratio strategy benefits from targeted augmentations and increased model diversity. Across MSP, TTA, and ensembles, GMM-based AggSs (*GMM-All, GMM-Int, GMM-Spa*) remain consistently strong, rarely showing significant disadvantages. Overall, this analysis supports our core hypothesis that, as a general rule, a stable top group of aggregators exists (the *prediction-aware* strategies and the meta-aggregators family), but that the optimal choice for unexplored cases strongly depends on the specific dataset properties and OoD perturbation.

C.2. On the AggS performance in Failure Detection

In line with the numerical and qualitative results discussed in Section 4.4, Table C.2b shows similar trends for AggSs applied on MSP uncertainty heatmaps. The top-performing predictors remain QFR and BCA, both *prediction-aware*, and *GMM-* meta-aggregators, while ICA continues to underperform due to repeated segmentation errors in small objects. Table C.3b and Table C.4b reveal no major deviations in their restricted analysis subsets, with one exception: the patch-based PLM20 ranks among the top five, owing to the localized uncertainty maps from test-time augmentations that closely match segmentation error regions. However, given the limited subset of data analyzed for TTA heatmaps, no general conclusions can be drawn regarding the UQ method.

For TTA, AggSs perform better in the three-label classification setting (*e.g.* ARC-Nuc). For MSP, however, performance is also negatively impacted when uncertainty maps show increased uncertainty starting from the borders and extending within the predicted mask (*e.g.* in LIDC-Tex), where the OoD variant exhibits inconsistent patterns. This demonstrates that in this case augmentations alone are insufficient to capture the uncertainty nuances between iD and OoD samples, while MSP tends to be overconfident in these regions.

Analysis of the p-value matrices reveals that QFR emerges as the top-performing aggregator for Failure Detection, showing statistically significant improvements over all other methods ($p < 0.05$ in all pairwise comparisons,

a AUROC

	ARC-BC	ARC-Nuc	CAR-CS	LIDC-Mal	LIDC-Tex	LIZ-IG	LIZ-SG	WEED-Hand	WORM-Nem	WORM-Pro	
GMM-All	.81 ± .05	.85 ± .05	1.00 ± .00	.92 ± .03	.55 ± .08	.55 ± .03	.45 ± .03	.93 ± .03	.99 ± .01	1.00 ± .00	**5.4**
BCA	.87 ± .04	.75 ± .06	.90 ± .01	.60 ± .05	.64 ± .09	.66 ± .02	.73 ± .02	.62 ± .06	.69 ± .06	.86 ± .04	**5.6**
ICA	.83 ± .05	.78 ± .06	.80 ± .02	.61 ± .05	.64 ± .09	.70 ± .02	.66 ± .02	.61 ± .06	.68 ± .06	.86 ± .03	**5.9**
QFR	.85 ± .05	.81 ± .06	.56 ± .02	.60 ± .05	.62 ± .08	.73 ± .02	.71 ± .02	.60 ± .06	.66 ± .06	.89 ± .03	**5.9**
GMM-Int	.81 ± .05	.76 ± .06	.80 ± .02	.91 ± .03	.58 ± .07	.52 ± .03	.46 ± .03	.94 ± .03	.99 ± .01	1.00 ± .00	**6.3**
GMM-Spa	.84 ± .06	.73 ± .07	1.00 ± .00	.70 ± .05	.58 ± .07	.51 ± .02	.52 ± .03	.89 ± .03	.82 ± .05	.87 ± .04	**6.7**
AVG	.62 ± .08	.60 ± .08	.74 ± .02	.96 ± .02	.40 ± .07	.77 ± .02	.78 ± .02	.32 ± .06	.50 ± .07	.65 ± .05	8.7
AQA .6	.68 ± .07	.60 ± .07	.60 ± .02	.96 ± .02	.40 ± .07	.77 ± .02	.77 ± .02	.33 ± .05	.47 ± .07	.63 ± .05	9.1
PLM 20	.66 ± .06	.62 ± .07	.37 ± .02	.96 ± .02	.40 ± .08	.68 ± .02	.68 ± .02	.57 ± .04	.62 ± .07	.80 ± .05	9.2
AQA .9	.72 ± .06	.61 ± .07	.44 ± .03	.96 ± .02	.41 ± .08	.76 ± .02	.79 ± .02	.31 ± .05	.42 ± .07	.60 ± .05	9.3
PLM 50	.55 ± .08	.61 ± .07	.42 ± .02	.96 ± .02	.41 ± .07	.69 ± .02	.69 ± .03	.56 ± .05	.48 ± .07	.78 ± .04	9.3
AQA .75	.67 ± .07	.61 ± .07	.56 ± .02	.95 ± .02	.41 ± .07	.77 ± .02	.78 ± .02	.32 ± .05	.45 ± .07	.62 ± .05	9.4
PLM 10	.50 ± .07	.58 ± .08	.35 ± .02	.90 ± .03	.45 ± .08	.66 ± .02	.59 ± .03	.50 ± .04	.69 ± .06	.81 ± .04	10.5
ATA .5	.88 ± .04	.82 ± .05	.31 ± .02	.50 ± .00	.50 ± .00	.52 ± .03	.63 ± .02	.46 ± .04	.43 ± .06	.35 ± .05	10.5
ATA .3	.87 ± .04	.72 ± .06	.52 ± .02	.65 ± .05	.34 ± .06	.60 ± .02	.61 ± .02	.44 ± .04	.48 ± .07	.42 ± .05	10.9
ATA .7	.56 ± .08	.50 ± .00	.28 ± .02	.50 ± .00	.50 ± .00	.50 ± .00	.64 ± .03	.39 ± .04	.50 ± .00	.50 ± .00	12.9

b E-AURC

	ARC-BC	ARC-Nuc	CAR-CS	LIDC-Mal	LIDC-Tex	LIZ-IG	LIZ-SG	WEED-Hand	WORM-Nem	WORM-Pro	
GMM-All	.06 ± .01	.03 ± .00	.05 ± .01	.07 ± .01	.10 ± .02	.19 ± .01	.21 ± .01	.12 ± .02	.08 ± .01	.04 ± .01	**4.4**
QFR	.05 ± .01	.02 ± .00	.07 ± .01	.06 ± .01	.13 ± .04	.28 ± .01	.32 ± .01	.16 ± .02	.04 ± .01	.04 ± .01	**4.5**
GMM-Int	.07 ± .01	.04 ± .01	.08 ± .01	.07 ± .01	.11 ± .02	.16 ± .01	.22 ± .01	.12 ± .02	.08 ± .01	.04 ± .01	**5.1**
BCA	.05 ± .01	.02 ± .00	.03 ± .01	.07 ± .01	.14 ± .04	.26 ± .01	.35 ± .01	.23 ± .02	.07 ± .02	.07 ± .01	**5.7**
GMM-Spa	.05 ± .01	.04 ± .01	.04 ± .01	.09 ± .01	.09 ± .01	.17 ± .01	.24 ± .01	.12 ± .02	.09 ± .02	.07 ± .01	**6.1**
ATA .5	.04 ± .01	.02 ± .00	.10 ± .01	.15 ± .01	.11 ± .01	.21 ± .01	.29 ± .01	.34 ± .02	.05 ± .01	.08 ± .01	7.1
ATA .3	.04 ± .01	.03 ± .01	.10 ± .01	.11 ± .01	.15 ± .03	.24 ± .01	.30 ± .01	.34 ± .02	.03 ± .01	.08 ± .01	8.1
PLM 20	.08 ± .01	.03 ± .01	.10 ± .01	.07 ± .01	.15 ± .04	.28 ± .01	.31 ± .01	.34 ± .02	.09 ± .03	.05 ± .01	8.8
PLM 10	.12 ± .01	.03 ± .00	.10 ± .01	.06 ± .01	.15 ± .04	.27 ± .01	.29 ± .01	.34 ± .02	.10 ± .04	.05 ± .01	9.0
ICA	.10 ± .01	.04 ± .01	.06 ± .01	.07 ± .01	.14 ± .04	.33 ± .01	.36 ± .01	.34 ± .02	.07 ± .02	.07 ± .01	9.6
PLM 50	.10 ± .01	.03 ± .00	.10 ± .01	.07 ± .01	.16 ± .04	.29 ± .01	.32 ± .01	.34 ± .02	.08 ± .03	.06 ± .01	10.0
AQA .6	.08 ± .01	.04 ± .01	.08 ± .01	.07 ± .01	.15 ± .04	.33 ± .01	.35 ± .01	.34 ± .02	.06 ± .02	.08 ± .01	10.7
AQA .9	.08 ± .01	.03 ± .00	.10 ± .01	.07 ± .01	.16 ± .04	.32 ± .01	.36 ± .01	.35 ± .02	.05 ± .02	.09 ± .01	10.8
AVG	.10 ± .02	.04 ± .01	.06 ± .01	.07 ± .01	.15 ± .04	.33 ± .01	.36 ± .01	.33 ± .02	.06 ± .02	.08 ± .01	10.9
AQA .75	.08 ± .01	.04 ± .01	.09 ± .01	.07 ± .01	.15 ± .04	.33 ± .01	.36 ± .01	.32 ± .02	.06 ± .02	.09 ± .01	11.1
ATA .7	.10 ± .02	.05 ± .00	.11 ± .01	.15 ± .01	.11 ± .01	.17 ± .01	.29 ± .01	.35 ± .02	.11 ± .02	.09 ± .01	11.1

Table C.2. **AggSs performance on uncertainty maps generated with Maximum Softmax Probability (MSP) in OoD and failure detection.** Color coding and metric interpretation are as detailed in Table A.1. (a)-(b) AggSs are ranked from best (top) to worst (bottom) based on their average metric, first computed across 500 bootstrap samples per dataset, and then averaged across datasets to ensure stable rankings. Each value is reported with the std. deviation across bootstrap samples.

with $p < 0.001$ in most cases). The next best group includes the GMM-based scores and BCA, which consistently outperform intensity based aggregators but are statistically indistinguishable from one another, indicating a robust, if slightly weaker, alternative to QFR. By contrast, threshold-based methods (ATA, AQA) and patch-based approaches (PLM) exhibit significantly lower performance in most tests, while the global average (AVG) consistently ranks lowest, confirming it should not be used as a default strategy.

These patterns are corroborated when examining AggSs on MSP, TTA, and ensemble-based uncertainty heatmaps. Across all three, QFR's dominance is consistently confirmed, with pairwise p-values frequently approaching machine precision ($p \ll 0.001$), establishing it as the unequivocal top-tier AggS for FD across all tested uncertainty frameworks. The second tier under MCD, consisting of *GMM*-variants and BCA, remains stable and consistently outperforms lower-tier aggregators - thresholding (AQA, ATA), patch-based (PLM), and averaging (AVG) - which are statistically inferior across nearly all datasets and tests. Collectively, these results validate the rankings reported in Figure 5b and underscore the importance of *prediction-aware*, structure-sensitive aggregators for effective FD.

C.3. On the AggS performance across downstream tasks

Mean ranks, metric values, and standard deviations reported in Table C.1, Table C.2, Table C.3 and Table C.4 imply that, in the absence of structural knowledge about the OoD dataset, aggregators possessing the theoretical property of *proportion invariance* (cf. Supp. A) and defined as *prediction-aware* can be reliably selected for both OoD Detection and Failure Detection tasks (with the exception of ICA in FD). This conclusion also applies to our proposed *GMM-All* and its ablated variants, *GMM-Spa* and *GMM-Int*, which provide a robust alternative across datasets when a dataset- or task-specific choice is uncertain.

D. Details on Experimental Setup

The datasets were selected to cover a broad range of application domains for image segmentation (autonomous driving, agricultural monitoring, biomedical imaging) as well as maximum diversity in terms of image structure (*e.g.* number, size, and shape of objects, as well as the number of semantic classes). This diversity is reflected in the distribution of aggregated uncertainty scores, as illustrated in Figure 2. Figure D.1 further complements the illustration of the datasets diversity by showing the distribution of aggregated scores along additional pairs of AggSs.

a AUROC

	ARC-BC	ARC-Nuc	CAR-CS	LIDC-Mal	LIDC-Tex	
GMM-All	.83 ± .05	.84 ± .06	.99 ± .00	.86 ± .03	.70 ± .07	**4.8**
QFR	.90 ± .04	.81 ± .06	.67 ± .02	.77 ± .04	.88 ± .04	**4.8**
GMM-Spa	.93 ± .03	.72 ± .06	1.00 ± .00	.83 ± .03	.70 ± .06	**5.0**
BCA	.89 ± .04	.76 ± .06	.93 ± .01	.71 ± .04	.82 ± .06	**5.2**
ICA	.88 ± .05	.79 ± .06	.89 ± .01	.70 ± .04	.82 ± .07	**5.4**
GMM-Int	.79 ± .05	.76 ± .06	.79 ± .02	.89 ± .03	.68 ± .06	**6.4**
AQA .9	.86 ± .05	.64 ± .07	.42 ± .03	.96 ± .01	.57 ± .07	8.6
PLM 20	.78 ± .05	.63 ± .07	.39 ± .02	.96 ± .01	.58 ± .07	8.8
AQA .75	.75 ± .06	.61 ± .07	.60 ± .02	.96 ± .01	.58 ± .07	9.0
AQA .6	.71 ± .06	.61 ± .07	.67 ± .02	.96 ± .01	.57 ± .07	9.4
PLM 50	.65 ± .08	.63 ± .07	.44 ± .02	.96 ± .01	.57 ± .07	9.8
AVG	.62 ± .06	.59 ± .07	.78 ± .02	.96 ± .01	.57 ± .07	10.2
PLM 10	.66 ± .06	.58 ± .08	.38 ± .03	.93 ± .02	.65 ± .07	11.2
ATA .5	.84 ± .05	.76 ± .06	.25 ± .02	.51 ± .06	.40 ± .08	11.6
ATA .3	.87 ± .05	.62 ± .07	.30 ± .02	.60 ± .05	.55 ± .08	11.8
ATA .7	.42 ± .07	.74 ± .06	.28 ± .02	.42 ± .05	.32 ± .06	14.0

b E-AURC

	ARC-BC	ARC-Nuc	CAR-CS	LIDC-Mal	LIDC-Tex	
QFR	.04 ± .01	.02 ± .00	.06 ± .01	.04 ± .00	.04 ± .01	**2.9**
GMM-Spa	.05 ± .01	.04 ± .01	.05 ± .00	.05 ± .01	.06 ± .01	**4.4**
GMM-All	.05 ± .01	.03 ± .00	.06 ± .00	.07 ± .01	.06 ± .01	**5.7**
BCA	.05 ± .01	.03 ± .00	.04 ± .00	.05 ± .01	.08 ± .01	**6.4**
GMM-Int	.06 ± .01	.04 ± .01	.08 ± .01	.06 ± .01	.06 ± .01	**7.0**
PLM 10	.10 ± .02	.02 ± .00	.09 ± .01	.05 ± .01	.07 ± .01	7.6
PLM 20	.07 ± .01	.03 ± .00	.09 ± .01	.05 ± .01	.08 ± .01	8.1
ATA .3	.04 ± .01	.03 ± .00	.10 ± .01	.12 ± .02	.12 ± .03	9.1
AQA .9	.05 ± .01	.03 ± .00	.09 ± .01	.06 ± .01	.08 ± .01	9.1
ATA .5	.05 ± .01	.02 ± .00	.11 ± .01	.16 ± .03	.17 ± .04	9.3
PLM 50	.07 ± .01	.03 ± .00	.09 ± .01	.06 ± .01	.08 ± .01	9.4
ATA .7	.14 ± .02	.01 ± .00	.10 ± .01	.23 ± .04	.18 ± .04	10.3
AQA .75	.07 ± .01	.03 ± .01	.08 ± .01	.06 ± .01	.08 ± .01	10.6
AQA .6	.08 ± .01	.04 ± .01	.08 ± .01	.06 ± .01	.08 ± .01	10.9
AVG	.10 ± .02	.04 ± .01	.07 ± .01	.06 ± .01	.08 ± .01	11.3
ICA	.10 ± .02	.04 ± .00	.06 ± .01	.05 ± .01	.08 ± .01	11.3

Table C.3. **AggSs performance on ensembling heatmaps in OoD and failure detection.** Color coding and metric interpretation are as detailed in Table A.1. (a)-(b) AggSs are ranked from best (top) to worst (bottom) based on their average metric, first computed across 500 bootstrap samples per dataset, and then averaged across datasets to ensure stable rankings. Each value is reported with the std. deviation across bootstrap samples.

D.1. Segmentation of nuclei in pathology images

LIZ (iD) The Lizard dataset [26] is a large-scale histopathology (H&E) dataset with annotations for instance- and semantic segmentation of cell nuclei in histopathology images of colon tissue. The dataset consists of 238 images of varying sizes, from which we extract patches of size 256 × 256 pixels. Patches from the DigestPath, TCGA, PanNuke, CRAG and CoNSeP subsets are used as iD.

LIZ-G (OoD) Based on established baselines [6, 64], we use the GlaS subset as the OoD dataset, consisting of 61 images, where we again extract patches of size 256 × 256 pixels. We evaluate both semantic and instance segmentation, referring to the resulting OoD sets as LIZ-SG (se-

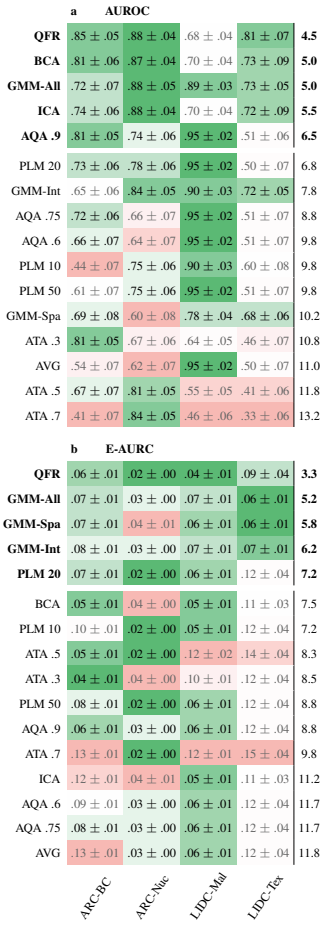

Table C.4. **AggSs performance on TTA heatmaps AggSs in OoD and failure detection.** Color coding and metric interpretation are as detailed in Table A.1. (a)-(b) AggSs are ranked from best (top) to worst (bottom) based on their average metric, first computed across 500 bootstrap samples per dataset, and then averaged across datasets to ensure stable rankings. Each value is reported with the std. deviation across bootstrap samples.

mantic) and LIZ-IG (instance). Here, the distribution shift arises from variations in acquisition and recording conditions (*e.g.*, lighting, temperature, and focal plane), which affect tissue appearance and increase overall object uncertainty.

Segmentation We train the HoVer-NeXt (HN) model [6]. HN builds upon HoVer-Net (HRNet, [75]) by simplifying the pipeline: it replaces the binary nuclei segmentation map with a 3-class center-background (BCB) prediction and merges the instance segmentation decoders into a single branch. The semantic arm predicts class labels for each pixel, while the instance arm outputs center-point vectors and BCB maps for individual nuclei. The architecture follows a U-Net [63] with a ConvNeXt-v2 encoder [79] (Tiny variant).

HN is trained from scratch for 200,000 steps with a batch size of 12 using AdamW (weight decay 0.0001) and a cosine-annealing learning rate schedule (1e-4 to 1e-8). The encoders use 50% dropout; the decoder does not. Data augmentations include: HED color, hue/saturation/brightness, random noise, Gaussian blur, rotation, flipping, mirroring, zoom, scale, shear, translation, and elastic transforms. The training loss combines semantic and instance losses, summed and weighted by 0.02. The instance arm uses MSELoss for center-point vectors and cross-entropy for the BCB map, while the semantic arm employs a standard focal loss [38] with $\gamma = 2.0$. Hyperparameter search and performance monitoring are conducted on the validation set, optimizing micro- and macro-F1 scores rather than panoptic quality, which has been shown to be suboptimal for histopathology data [18].

D.2. Synthetic Histopathological Images

ARC (iD) The Arctique dataset [19] is a procedurally generated dataset modeled after histopathological colon tissue sections (inspired by the LIZ dataset), providing precise ground-truth masks for both semantic and instance segmentation. We use the standard training set which comprises 1,500 samples of 512×512-pixel RGB images.

ARC-Nuc/ARC-BC (OoD) Arctique includes controlled variations in targeted parameters, such as the presence of red blood cells (ARC-BC) and the intensity of nuclei staining (ARC-Nuc). In the ARC-BC setting, additional red blood cells are misclassified as eosinophil cells, leading to local increases in uncertainty at the locations of the blood cells for semantic segmentation. In the ARC-Nuc setting, treduced nuclei staining intensity makes it more difficult for the model to identify individual cells, thus increasing the uncertainty at object boundaries in instance segmentation.

Segmentation We use the same training setup as described for the LIZ dataset.

D.3. Binary segmentation of lung nodules in CT volumes

LIDC (iD) We use the LIDC-IDRI dataset [4], with the `pylidc` library [27], following the pre-processing strategy of Kahl et al. [31]. in Kahl et al. [31]. In particular, we use only lung nodules $\geq$ 3mm; each annotated by up to four raters, and a consensus mask is computed as the union of all annotations. Images are cropped to 64×64×64 voxels and resampled to 1×1×1 mm resolution, yielding 901 samples. Finally, for better comparability with the other datasets, we convert the images and masks volumes to 2d by restricting evaluation to the central 50% of the z-axis.

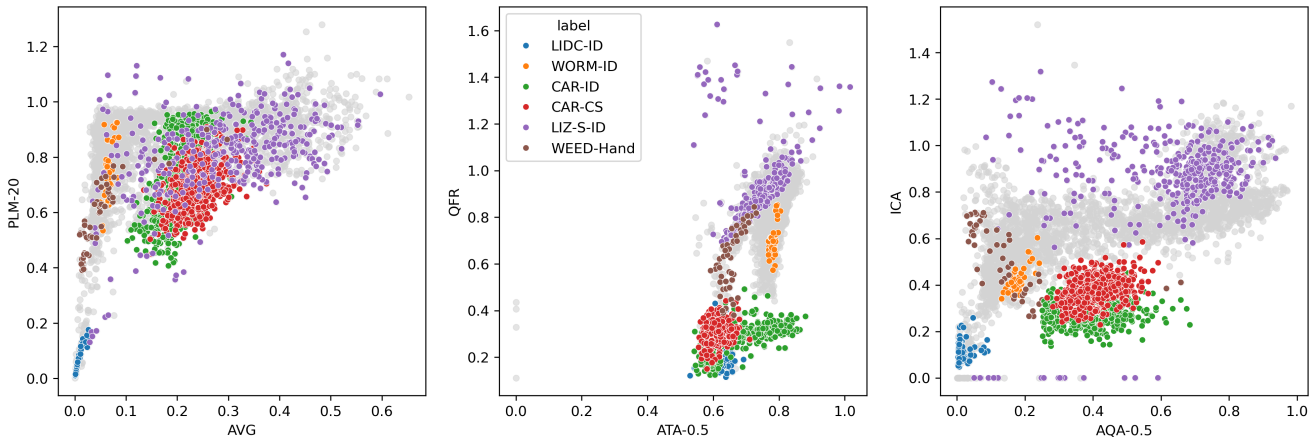

Figure D.1. **Illustration of dataset diversity as captured by different aggregation measures.** Unlike the projection of uncertainty maps into EDS–MOR space (cf. Fig. 2, with non-highlighted datasets shown in gray), projections onto spaces defined by randomly chosen pairs of *pixel-intensities* or *prediction-aware* AggSs do not reveal a clear separation of the datasets. Nevertheless, the two-dimensional configuration of CAR-ID and CAR-CS points suggests that a separation might emerge in a multidimensional space incorporating all AggSs.

LIDC-Tex/LIDC-Mal (OoD) The LIDC-IDRI dataset includes nine metadata features, each with 4–6 categorical values assigned by the raters, which we use to define the iD and OoD splits. Specifically, we consider textured (iD) nodules vs. non-textured (OoD) nodules for the texture shift (LIDC-Tex) and benign (iD) nodules vs. malignant (OoD) nodules for the malignancy shift (LIDC-Mal). Since malignant nodules are typically larger in size than benign nodules, the uncertainty in the LIDC-Mal shift is expected to increase along the object boundaries. In contrast, in the LIDC-Tex shift uncertainty will increase within the nodule, reflecting the change in texture.

Segmentation We use a 3D U-Net as the segmentation backbone to detect lung nodules, with an initial filter size of 16 and four encoder–decoder blocks. The loss combines Dice and cross-entropy terms. Data augmentations include random flipping and Gaussian noise. Training uses the Adam optimizer (learning rate 3e-4 and weight decay 1e-5) with a batch size of 8 and 50% dropout after each convolutional block. During training, validation performance is measured as the average Dice score between the predictions (or a single prediction for MSP) and each of the four reference annotations. Further details are provided in the Supp. of Kahl et al. [31].

D.4. Binary segmentation of microorganisms in microscopy images

WORM (iD) The C.elegans live/dead assay from Broad Bioimage Benchmark Collection (BBBC010) contains 100 brightfield microscopy images of multiple C. elegans worms with pixelwise foreground and instance segmentation masks [40]. Images are single-channel and and cropped from 696×520 pixels to 512×512 pixels to match the network input constraints.

WORM-Nem/WORM-Pro (OoD) We use the Protist (1003 samples) and Nematode (468 samples) datasets from the SinfNet Microorganism Image Classifier dataset [65]. Both contain microscope images of various microorganisms along with polygon annotations of the instances. For comparability with the iD data, images are converted to grayscale, polygon annotations are converted to pixelwise masks, and both images and masks are cropped/resized to 512×512 pixels. In this binary segmentation setting, the main sources of uncertainty are the unfamiliar shapes of the other microorganisms and the background noise (*e.g.* small particles misclassified as foreground).

Segmentation We train a standard U-Net with 5 convolutional layers, ReLU activations and 20% dropout set after each convolutional layer. Input images are augmented with random resized crops, flips, Gaussian blur, sharpness, gamma adjustment, and intensity inversion. The network predicts binary foreground/background, trained with a combined Dice and cross-entropy loss. Optimization is performed with Adam (learning rate 1e-5, weight decay 1e-5), in batches of 16 for 20 epochs. Validation performance is monitored using the Dice score.

D.5. Semantic segmentation of urban street scenes

CAR-ID (iD) The GTA dataset [62] contains 24,966 (1914×1052) of urban street scenes with dense semantic segmentation masks covering 19 classes from the computer

game GTA V. Semantic masks are constructed by intercepting information between the game engine and graphics processor, and by reconstructing and annotating individual objects in scenes. These masks distinguish between 19 semantic classes. To address ethical concerns, we verified that the dataset contains only urban street scenes with the 19 semantic classes matching standard driving benchmarks; no violent or sexualized content is present, and pedestrians appear incidentally while walking on sidewalks (0.36% of pixels). We thank the anonymous reviewer for bringing this issue to our attention.

CAR-CS (OoD) The Cityscapes dataset [13] includes 5,000 densely annotated urban street scenes from 27 cities and 20,000 coarsely annotated images from 23 cities. Images were captured via stereo cameras, thus the resolution being 1280×720, and annotations include instance and semantic segmentation masks (differientating 30 classes), which we map to the 19 GTA classes. The OoD-ness stems from a sim-to-real shift, affecting the entire image and resulting in more "blurry" uncertainty maps.

Segmentation We perform semantic segmentation using HRNet [75] as the backbone, pre-trained on ImageNet. The model is trained with cross-entropy loss and optimized using SGD (learning rate 0.01, weight decay 5e-4, momentum 0.9) with a batch size of 6. Data augmentations include random horizontal flipping, rotations, scaling, cropping, and Gaussian noise. 50% dropout is applied at the end of each branch. Validation performance is monitored using the Dice score. Further training details are provided in the Supplement of Kahl et al. [31].

D.6. Crop and Weed

WEED (iD) The Weedsgalore datase is a high-resolution drone-based multispectral imaging dataset with dense annotations for crop and weed segmentation in maize fields [8]. We use the training set comprising 104 images, each measuring 600×600 pixels, and only the RGB channels for prediction.

WEED-Hand (OoD) The "Crop and Weed" dataset [68] contains 8,034 manually captured RGB images covering 16 species of crop and 58 species of weed over a variety of locations, soil types and lightning conditions. For comparability with Weedsgalore we collapse annotations into two classes, by considering maize as "crop" and all other plants as "weed". We further crop images to a a quadratic size of 600×600 pixels.

Segmentation We use publicly available checkpoints of a probabilistic DeepLabv3+ model [12] trained on Weedsgalore for 3-class semantic segmentation (background, crops, weeds).

D.7. Training and Hardware

When training was performed from scratch, it was carried out on single NVIDIA H100 and A40 GPUs using PyTorch [55].

D.8. Evaluation sets

Dataset	iD Samples	OoD Samples
ARC-BC	25	50
ARC-Nuc	25	50
LIDC-Mal	53	93
LIDC-Tex	84	20
WORM-Nem	25	47
WORM-Pro	25	82
CAR-CS	300	300
WEED-HAND	26	159
LIZ-SG	193	356
LIZ-IG	193	356

Table D.1. Sample counts for iD and OoD splits across evaluation datasets.

As evident from the dataset descriptions in this Supp. Section, dataset sizes vary widely, from roughly 100 training samples in WORM to tens of thousands in CAR. For comparability between iD and OoD sets, as well as for processing efficiency, we use randomly selected subsets of the larger datasets for evaluation experiments. Final sample counts are summarized in Table D.1.

E. Details on Downstream Tasks

E.1. Out-of-Distribution Detection

To empirically evaluate the theoretical properties of AggSs, we focus on Out-of-Distribution (OoD) detection at the image level rather than the pixel level. Intuitively, humans perceive an image as OoD as a whole rather than based on individual pixels. Moreover, when part of an image is OoD, it can affect all predictions, making them unreliable.

Metric: Area Under the Receiver Operating Characteristic Curve (AUROC) We use AUROC to assess an AggS's ability to detect OoD images. Each image is labeled as 1 for OoD and 0 for iD. The True Positive Rate (TPR) is the proportion of correctly identified OoD images, i.e., the fraction of OoD samples whose aggregated uncertainty score exceeds a given threshold. The False Positive Rate (FPR) reflects the proportion of iD images incorrectly classified as OoD, i.e., the fraction of iD samples whose aggregated uncertainty score surpasses the same threshold. By

varying this threshold, we calculate the TPR and FPR at different levels to construct the ROC curve and subsequently compute the AUROC. We use the `sklearn` library [56] to first compute the TPRs and FPRs of the ROC curve with the ground truth input (0 or 1) and the aggregated uncertainty scores as target values. From this we then calculate the AUC (area under the curve).

E.2. Failure Detection

To assess the impact of AggSs on model performance in real-world applications, we define a continuous failure signal based on two segmentation accuracy metrics: (1) micro-Dice, which captures the accuracy of object and boundary detection in the instance-3-label segmentation task (*e.g.* nuclei and boundary detection), and (2) macro-Dice, which reflects class-wise performance in semantic segmentation tasks. The motivation behind this approach is that, while automated decision-making requires a holistic view at image level, the performance of a panoptic, instance-based three-label, or semantic segmentation model must ultimately be evaluated at the instance level.

Metric: Excess-Area Under the Risk-Coverage Curve (E-AURC) As analyzed by Jaeger et al. [30] and originally proposed by Geifman et al. [25], we use the Excess-Area Under the Risk-Coverage Curve (E-AURC) as the evaluation metric for the SC experiment. In this downstream task, E-AURC measures the quality of AggSs-based ranking while remaining independent of the underlying model's absolute performance. Moreover, we compute E-AURC for each employed uncertainty-based predictive model, making the segmentation model's absolute performance irrelevant and further confirming the suitability of E-AURC for our benchmarking. Like AURC [24], E-AURC balances two objectives: minimizing risk (i.e., ensuring strong classifier performance) while maximizing coverage (i.e., reducing the fraction of cases requiring manual review). Both AURC and E-AURC are computed following the implementation of Jaeger et al. [30], adapted here for multi-class segmentation as detailed below.

Let ϕ denote a predictive model that maps an image to a multi-class segmentation mask. We define the evaluation dataset of images I and segmentation masks M as $\widetilde{\mathscr{D}} = \{I_\ell, M_\ell\}_{\ell=1}^{L}$, and the *confidence scoring function* (CSF) $g(I_\ell)$ as the negative aggregated uncertainty score, $-f(U_\ell)$. Furthermore, we choose the inverted micro-Dice (macro-Dice) as the *risk* s associated with a prediction,

$$s(I, M, \phi) = 1 - \mathrm{Dice}(\phi(I), M). \tag{3}$$

The risk-coverage curve is obtained by introducing a confidence threshold ρ, which leads to the selective risk

$$\text{Sel. Risk}(\rho|\phi, g, \widetilde{\mathscr{D}}) = \frac{\sum_\ell s(I_\ell, M_\ell, \phi) \cdot \mathbb{I}(g(I_\ell) \geq \rho)}{\sum_\ell \mathbb{I}(g(I_\ell) \geq \rho)} \tag{4}$$

and coverage, defined as the ratio of cases remaining after selection,

$$\mathrm{Coverage}(\rho|g, \widetilde{\mathscr{D}}) = \frac{\sum_\ell \mathbb{I}(g(I_\ell) \geq \rho)}{L}. \tag{5}$$

The AURC based on a threshold list $\{\rho_r\}_{r=1}^{R}$ with R values of a CSF that are sorted ascending can then be computed as,

$$\mathrm{AURC}(f, g, \widetilde{\mathscr{D}}) = \sum_r (\mathrm{Coverage}(\rho_r) - \mathrm{Coverage}(\rho_{r-1}) \cdot \tag{6}$$
$$\cdot \frac{(\text{Sel. Risk}(\rho_r) + \text{Sel. Risk}(\rho_{r-1}))}{2}$$

where we omit the conditioning on ϕ, g, $\widetilde{\mathscr{D}}$ on the RHS for brevity. We now derive the E-AURC as follows,

$$\text{E-AURC} = \mathrm{AURC}(\phi, g, \widetilde{\mathscr{D}}) - \mathrm{AURC}(\phi, g^*, \widetilde{\mathscr{D}}) \tag{7}$$

where the second term represents the optimal AURC achievable. This optimal CSF can be defined, for example, by an oracle that assigns to each prediction a confidence equal to the negative risk: $g^*(x) = -s(x, y, \phi)$. In practice, this corresponds to perfectly ranking the predictions by their risk, in our case, in ascending order of Dice score. For a formal discussion of the applicability and limitations of this metric, please refer to Jaeger et al. [30].

F. Details on UQ map generation

For UQ, we use Monte Carlo Dropout for all results shown in the main text. All segmentation models employ dropout during training, and the dropout layers remain active at test time. By passing each input image through the model for $L = 10$ times with different dropout masks we generate L samples of the pixel-wise class probabilities $p_i^{(l)}(c)$, where $l = 1, \ldots, L$ indexes the dropout samples and $i = 1, \ldots, mn$ refers to the pixel index. We then compute the mean class probability across samples as:

$$\bar{p}_i(c) = \frac{1}{L} \sum_{l=1}^{L} p_i^{(l)}(c). \tag{8}$$

Finally, pixelwise uncertainty is calculated using Shannon entropy over the averaged probabilities:

$$u_i = -\sum_{c=1}^{K} \bar{p}_i(c) \log \bar{p}_i(c). \tag{9}$$

Shannon Entropy is maximal when all K classes have equal probability i.e $p(c) = 1/K$ for $c = 1 \dots K$. In that case $-\sum_{c=1}^{K} p(c)\log(p(c)) = \log(K)$ and thus the pixelwise uncertainty scores u_i fall in $[0, \log(K)]$ where K is the total number of (semantic) classes.

The same generation process applies to the pixel-wise class probabilities produced by other UQ methods when replicating the AggSs benchmarking for OoD and FD detection. What differs is how and how many samples $p_i^{(l)}(c)$ are generated:

- TTA: $L = 16$ samples are generated by applying various combinations of the augmentations used during training to the test input, then inverting them before applying Softmax.
- DE: $L = 5$ samples are obtained from predictions of five segmentation models with identical architecture and training setup, but initialized with different random seeds.
- CE: $L = 10$ samples are generated using nine model checkpoints selected near convergence in the loss basin.

we generate $L = 10$ by using 9 checkpoints of the model registered near convergence to loss basin.

To ensure comparability across aggregated uncertainty maps, we divide the pixelwise scores from all UQ methods except MSP by $\log(K)$ to make sure all maps are normalized such that $U \in [0, 1]^{m \times n}$.

G. Details on Meta-Aggregation via GMM

G.1. Gaussian Mixture Models

The meta-aggregation approach is motivated by the task of OoD detection as illustrated in Fig. 1. Here, we assume access to a representative collection of uncertainty maps generated from iD data, i.e. data drawn from the same distribution the model was trained on. To summarize the overall uncertainty present in these maps, we can apply one of the various aggregation strategies (AggSs) introduced in Sec. 3.1, 3.3 and 3.4.

When deploying the same trained model on a new, potentially out-of-distribution sample our experimental results in Fig. 3 and Fig. 5 reveal a key challenge: it remains unclear which specific AggS is most suitable for determining whether a given sample is iD or OoD for any particular task. Therefore, instead of relying exclusively on any single AggS we aim to combine multiple AggSs each focusing on different aspects of the uncertainty map. To this end, we represent each uncertainty map U by its aggregated feature vector $\mathbf{f}_U \in \mathbb{R}^d$, defined as $\mathbf{f}_U = (f_1(U), \dots, f_d(U))$ where d denotes the number of aggregation strategies. We then model the distribution of these features using a Gaussian Mixture Mode Gaussian Mixture Model (GMM) [16]:

$$p_{\text{GMM}}(\mathbf{f}_U; \Theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{f}_U \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{10}$$

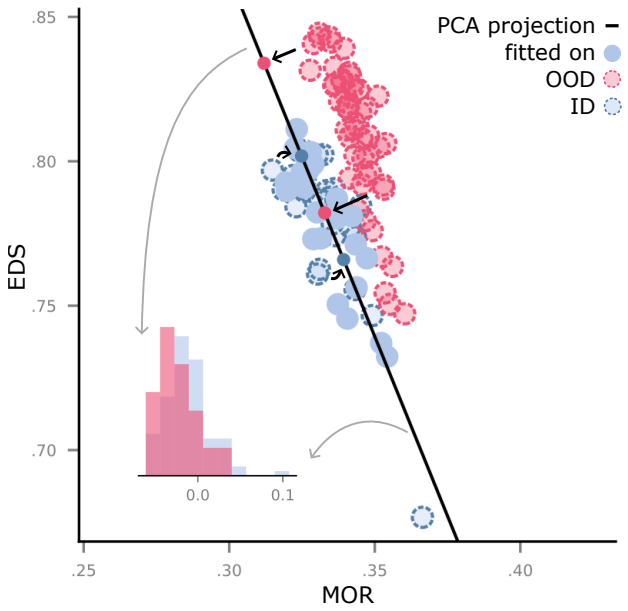

Figure G.1. **Effect of PCA.** Example of feature collapse resulting from computing a 1D PCA projection in 2D space using only iD samples, and subsequently applying this projection to both iD and OoD samples. The resulting histogram (inset) illustrates that this approach yields poor separability between iD and OoD data.

Here, $\Theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ are the parameters of the mixture, with π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ representing the mixing coefficient, mean vector, and covariance matrix of the k-th component, respectively. The number of components K is selected via the Bayesian Information Criterion (BIC) [66]. Each Gaussian component is defined by

$$\mathcal{N}(\mathbf{f}_U \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}\,|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{f}_U - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{f}_U - \boldsymbol{\mu})},$$

And the mixture weights π_k are constrained by:

$$\sum_{k=1}^{K} \pi_k = 1, \ \pi_k \geq 0. \tag{11}$$

We estimate the mode-parameters Θ on a held-out iD subset and evaluate on unseen iD and OoD data points.

The final meta-aggregator score is defined as the negative log-likelihood (NLL):

$$f_{\text{meta}}(U) = -\log p_{\text{GMM}}(\mathbf{f}_U), \tag{12}$$

which serves as a unified uncertainty score: larger values indicate greater deviation from the learned iD distribution and thus flag the corresponding sample as likely OoD.

An example of the effectiveness of this approach is illustrated in Fig. 4d-e, where EDS and MOR individually are unable to separate between iD and OoD data. Yet, fitting a GMM on both allows for an improved separation.

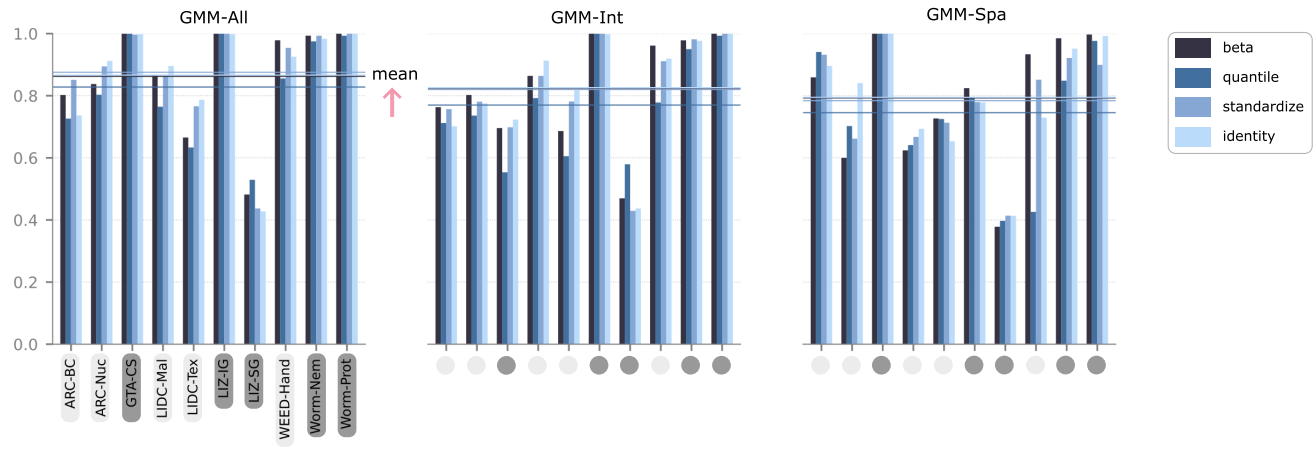

Figure G.2. **Data Normalization.** AUROC scores from our OOD detection experiment are shown for four different normalization methods: Beta CDF Gaussianizer, Standard Scaler, Quantile Transformer, and no normalization. The three bar plots represent the performance of the methods *GMM-All*, *GMM-Int*, and *GMM-Spa* under each normalization setting.

G.2. Data preprocessing

We compare three different aggregation sets to fit a GMM on: (1) *GMM-Int*, based on 13 classical intensity-based AggSs; *GMM-Spa*, which includes our 3 novel spatial AggSs; and (3) *GMM-All*, which combines all 16 aggregated scores. The relatively high dimensionality of the aggregated uncertainty feature vectors would typically motivate dimensionality reduction prior to fitting the GMM. However, for the task of OoD detection, such an approach can actually be detrimental. As illustrated in Fig. G.1 for the Principled-Component Analysis (PCA) technique: projecting the OoD samples onto the two lower-dimensional PCA subspace leads to a strong overlap between iD and OoD samples, making their separation more challenging.

A second preprocessing issue concerns the bounded range of the individual AggSs. These scores lie in $[0, 1]$, whereas GMMs are defined over $[-\infty, \infty]$. However, we find that non-linear transformations can lead to distortions, by exaggerating differences between data points that are close to 0 or 1, potentially affecting the GMM's ability to accurately detect OoD samples. Hence, following common practice, we standardize our features $\mathbf{f}_U$ by $\mathbf{f}'_U = (\mathbf{f}_U - \mu)/\sigma$, where μ and σ denote the mean and standard deviation, respectively. This preprocessing is used for the results shown in Figure 3 and 5.

To evaluate the impact of the normalization methods on the GMM performance, in Fig. G.2 we test four different strategies (including no normalization) for the OoD detection benchmark (note that prior to normalization, we rescale scores as $\mathbf{f}_U = (1 - 2\epsilon)(\mathbf{f}_U - 0.5) + 0.5$ with a small ϵ to avoid values at 0 or 1). Two key takeaways emerge regardless of the normalization method: (1) the mean AUROC consistently increases when using all the features (*GMM-All*); and (2) the meta-aggregators outperform in most cases (7 out of 10) the individual AggSs (cf. Figure 3). Furthermore, while performance varies across individual datasets, the mean AUROC indicates that the Quantile Transformer is the least effective among the tested preprocessing methods.

G.3. Meta-aggregation ablations

Figure G.4 provides a comprehensive overview of the ablation studies conducted for the meta-aggregation *GMM-All* in OoD detection. We focus on this downstream task, as its evaluation metric is more intuitively interpretable than that of FD. The top panel of Figure G.4a compares the AUROC of *GMM-All* (shown as a horizontal black line) to leave-one-out variants, in which one AggS is omitted at a time during fitting. Results are visualized as jittered points obtained by evaluating across test samples spanning the 10 benchmarking datasets and following the bootstrapping protocol described in Supp. G.2. The iD and OoD samples used to compute the AUROC are the ones reported in Supp. D.8 and obtained via a fixed split; the remaining iD samples are those used for fitting the GMM. The bottom panel of Figure G.4a, in turn, shows the performance of GMMs fitted on each individual AggS relative to the combined approach.

The two panels indicate that fitting *GMM-All* on all AggSs either outperforms or matches the performance of individual-feature fits in 6 of 10 datasets, with generally minimal sensitivity to the removal of specific features. Exceptions arise when: a particular feature dominates (*e.g.*, EDS for CAR-CS and QFR for WEED-Hand, marked as Tukey outliers [72]), or when most features lack clear discriminatory power between iD and OoD samples (*e.g.*, in LIZ-SG). Notably, the most discriminative AggSs for ARC-BC, ARC-Nuc, CAR-CS, LIDC-MAL, and WEED-

Hand are either *spatial mass ratios* or *prediction-based* (proportion-invariant), consistent with the findings reported in Section 4.5.

Figure G.3 provides distributional insights into the individual fitting for the CAR-ID and CAR-CS datasets by illustrating the distinct iD and OoD distributions of the AggSs. It further depicts the *GMM-All* density when fitted on each individual AggS (solid line), alongside its corresponding diagonal component (dotted line) when fitted on all AggSs. While the OoD shift induced by real CAR-CS images relative to synthetic CAR-ID images is clearly reflected in the EDS, this is not the case for MOR, where the support and likelihood of the iD and OoD distributions do not exhibit a clear separation, nor for ATA .3, whose distribution is broader and bi-modal, making an individual GMM fit insufficient. In contrast, ENT highlights a key advantage of the meta-aggregator: using a density-based OoD detector allows us to identify OoD samples, even when their aggregate uncertainty for that score is lower than that of iD samples. Overall, this illustrates that spatially aware scores can be incorporated as AggSs in a GMM-based OoD detector, potentially allowing us to capture distributional structure beyond simple shifts in aggregate uncertainty.

To complement our ablations and provide a mechanistic explanation for the results in Figure 3, we first fit the *GMM-All* on all aggregated values using the protocol and sample splits described above. We then compute NLL scores for iD and OoD test samples, followed by SHAP values [41] to assess how individual features contribute to the *GMM-All* predictions (cf. Supp. D.8). These SHAP values quantify the contribution of each AggS to the *GMM-All*'s performance in detecting OoD samples.

Figure G.4b reports these SHAP values averaged across samples from the 10 benchmarking datasets (absolute SHAP), highlighting that, for datasets where *GMM-All* achieved a particularly high AUROC, all individual AggS components contributed positively, producing a clear separation between OoD samples and the estimated GMM modes. The positive contributions are especially pronounced for datasets exhibiting semantic shifts (WEED-Hand, WORM-Nem, WORM-Pro). By contrast, for synthetic datasets with covariate shifts (ARC-BC and ARC-Nuc), uncertainty concentrated at cell borders reduces the effectiveness of intensity-based AggSs such as PLM and AQA. These methods rely on hyperparameters tuned to localized uncertainty distributions in the simulated tissue slices, slightly limiting overall *GMM-All* performance.

The performance of *GMM-All* on CAR-CS is entirely explained by EDS, which alone suffices as a discriminator, followed by ENT. For lung-nodule detection, even if prediction-unaware AggSs are favored by the larger malignant nodules in LIDC-Mal, a few positive outliers from *prediction-aware* and other intensity-based AggSs still contribute meaningfully within *GMM-All*, despite generally offset contributions from the remaining AggSs. A similar pattern occurs for partly solid tumors in LIDC-Tex, although the effect of these outliers is milder. Consequently, the *GMM-All* performance is somewhat lower in LIDC-Tex than in LIDC-Mal, as individual *prediction-aware* AggSs achieve high performance when fitted with a single GMM (bottom panel of Figure G.4a). Finally, the absolute SHAP values help explain the relatively poor performance of the meta-aggregator on the LIZ dataset. For both OoD variants, LIZ-IG and LIZ-SG, we observe an offset in AggS contributions, resulting from the presence of both high and low values for certain features in iD and OoD samples. This variability arises from tissue folds, background noise, and variable cell positioning, independently of the distribution shift introduced by the different recording technique.

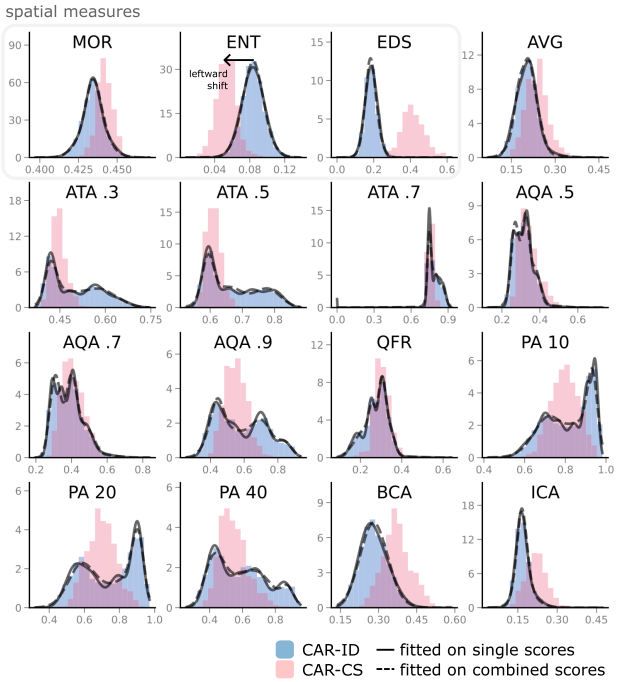

Figure G.3. **Utilizing spatial measures as AggSs.** Histograms of individual AggSs for CAR-ID and CAR-CS are shown along with: 1. a GMM fitted on an an iD subset of the individual score (solid line), and 2. the diagonal component of a GMM fitted on all scores (dotted line). This empirically shows that spatial measures can be leveraged by a GMM to detect OoD samples, even when the shift from iD to OoD is toward lower aggregated values.

G.4. Gradual distribution shifts

In this paper, we demonstrated that fitting a Gaussian Mixture Model (GMM) on multiple aggregation scores yields a robust OoD detection method. Our experiments were designed by pairing an iD dataset with an OoD counterpart, showing that the GMM effectively distinguishes between

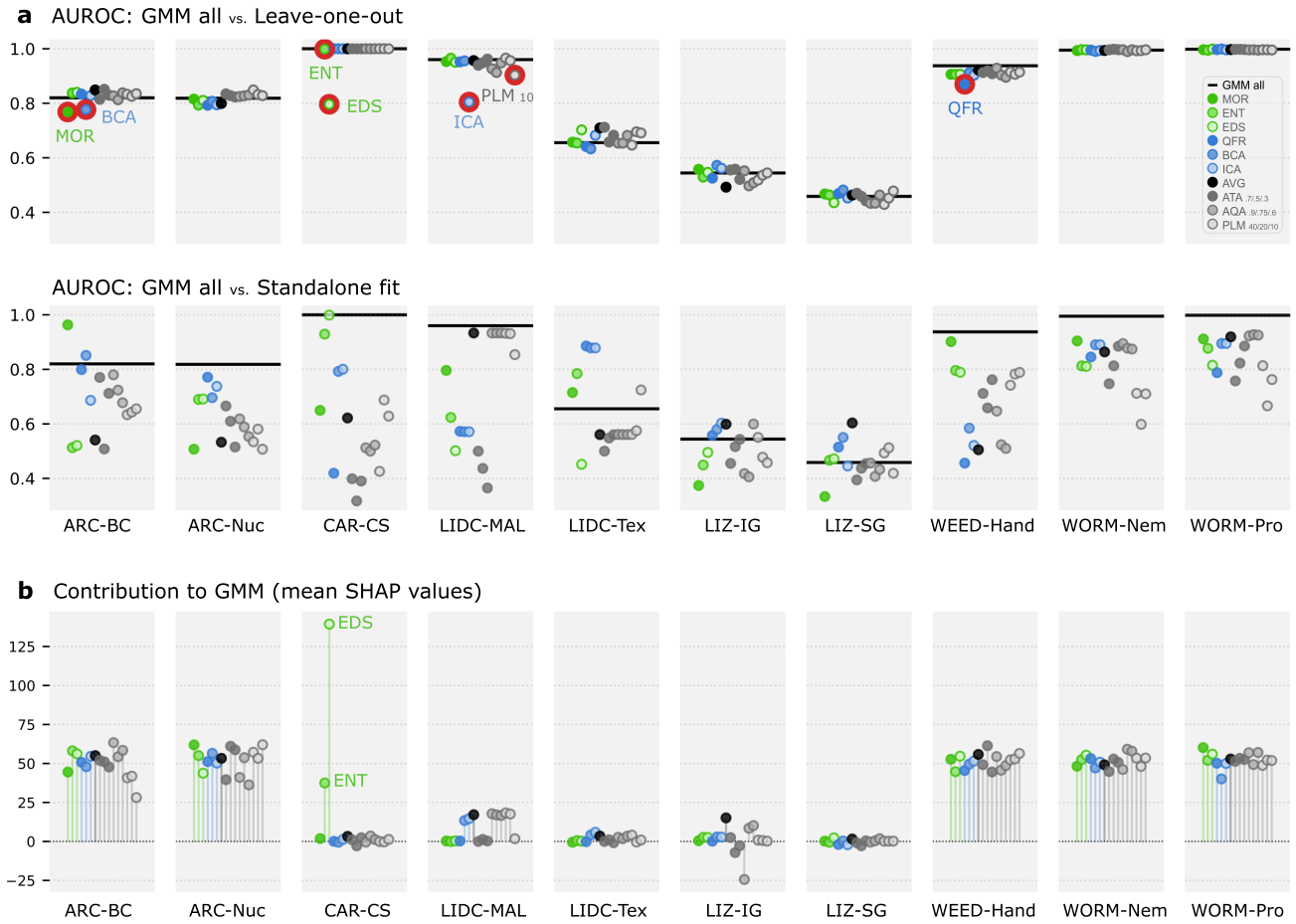

Figure G.4. **GMM Robustness for OoD Detection for all datasets.** (a) ***GMM-All* AUROC: leave-one-out (top) and individual fitting (bottom).** Using all AggSs generally matches or outperforms individual ones in 6 of 10 cases, with minimal impact from removing specific AggSs. Exceptions occur when a feature dominates (*e.g.*, EDS for CAR-CS or ICA in LIDC-MAL, marked in red as Tukey outliers) or when features lack discriminative power between iD and OoD samples (*e.g.*, LIZ-SG). The most discriminative AggSs for ARC-BC, ARC-Nuc, CAR-CS, LIDC-MAL, and WEED-Hand are typically *spatial mass ratios* or *prediction-based* (proportion-invariant). (b) ***GMM-All* AUROC: absolute SHAP values.** Positive values indicate that an AggS helps *GMM-All* separate iD–OoD, confirming the results observed in (a); bi-directional or null contributions reduce performance (e.g., in LIZ-IG).

the two in this binary setting. However, an ideal score should not only support binary classification but also capture gradual distribution shifts.

To explore this, we used the synthetic ARC dataset, which enables the study of such progressive changes: ARC-Nuc involves perturbations of nuclei intensity, while ARC-BC involves an increased presence of blood cells (see Fig. G.5). Note, in our main analysis, we only focused on the ARC-Nuc 0.5 and ARC-BC 0.75 variations.

Fig. G.5 illustrates whether the aggregated scores increase or decrease in response to the respective perturbations. Notably, the AVG score shows almost strictly monotonic increases. However, its separation between perturbed and unperturbed samples is less pronounced than that of the GMM score for instance. While the GMM score does not increase perfectly monotonically, it still exhibits a high fraction of increases—0.94 for ARC-Nuc and 0.84 for ARC-BC. Also *e.g.* the AggSs exhibits promising behaviour in that regard.

H. Details on Limitations

Uncertainty Types While disentangling epistemic and aleatoric uncertainty is important—particularly for OoD detection—such a separation would primarily affect pixel-wise scores rather than the aggregated outputs. Moreover, the impact on failure detection (FD) performance is expected to be limited, as this task may be conducted on iD test data only, depending on the scenario under consideration. Nevertheless, investigating how uncertainty type dis-

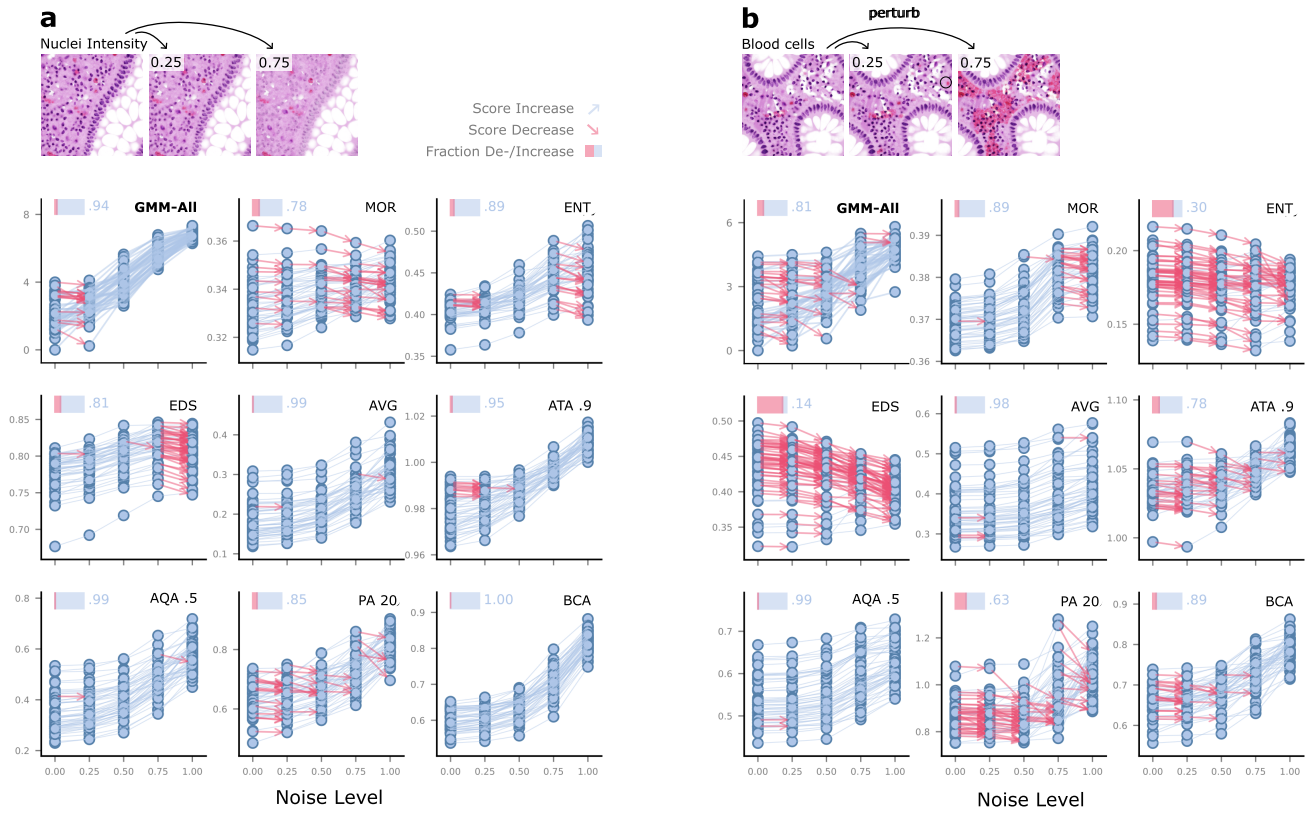

Figure G.5. **AggSs behavior under gradual perturbations.** For varying perturbation levels of a) ARC-Nuc and b) ARC-BC we display the sample-wise behavior of selected AggSs, i.e. an increase with increasing perturbation parameter manipulation is marked as light blue connection, a decrease as red arrow. The overall fraction between de- and increases is inset in the upper left corner of each AggSs plot.

entanglement influences aggregation remains an interesting direction for future work.

Diversity of datasets and OoD scenarios We aimed for maximal diversity in our datasets, focusing on realistic shifts following Taori et al. [70]. Nevertheless, our selection is naturally non-exhaustive. Other types of OoD shifts, such as image augmentations (e.g., blurring) or new-class shifts in label space, could also be explored. Furthermore, our study focuses on the aggregation of uncertainty in 2D. While we expect many of our findings to generalize to 3D, additional challenges may arise in volumetric settings, such as when uncertainty is concentrated at object surfaces.

Diversity of AggSs Our work emphasizes the most common intensity-based AggSs. Future investigations could incorporate additional sources of information beyond prediction masks, such as ground truth annotations, error maps for calibration tasks, or auxiliary modalities (*e.g.* depth). Another direction is the extension of the pipeline to unbounded logit-based scores, such as DDU [50, 51] or energy scores [39]. However, this extension is non-trivial, as extreme pixel values can dominate aggregation and may require specifically tailored AggSs. For the AggSs explored in this study, hyperparameters were chosen based on standard practices in the literature; a more comprehensive parameter sweep could reveal more optimal settings.

Gaussian Mixture Models The GMM approach demonstrates the effectiveness of combining individual AggSs for OoD and Failure Detection. In Supp. G.2, we already discuss the potential issue arising from data transformations. Another prominent limitation occurs in low-data, high-dimensional settings, where a GMM may struggle to accurately fit the sample distribution. This highlights the need for more robust OoD detection methods or a reduction in the number of aggregation measures in these particular settings. Alternatives to GMMs could be explored in future work, such as Vine Copulas [14], which can handle arbitrary distributions within the [0,1] range without requiring transformations and may provide greater robustness in a higher dimensional settings.
Additionally, further studies could investigate ablations with fewer or better-chosen parameterized AggSs (*e.g.*

PLM), potentially revealing an optimal dimensionality for the GMM and more stable performance in certain datasets (e.g., LIZ). For completeness, it would be interesting to compare our non-learned, interpretable meta-aggregator with OoD detectors based on learned representations, such as those obtained through self-supervised learning for modeling sample distributions. Our approach retains interpretability, as it relies on spatial and intensity-based scores whose individual meanings directly reflect the structure and magnitude of uncertainty.