Supplementary Material

Model Selection and Post Hoc Benchmarking

The TriPath model was initially developed using GigaPath-derived features, prior to large-scale benchmarking of alternative foundation models. To contextualize this design choice, we subsequently evaluated multiple feature extractors and aggregation strategies across biopsy and aspirate WSIs. Specifically, we compared four foundation models (GigaPath [7], H-Optimus-0 [4], Virchow2-CLS [8], and MUSK [6]) combined with multiple-instance learning (MIL) aggregators (ABMIL [1], DSMIL [3], and TransMIL [5]) for feature extraction from whole-slide images (WSIs) (Table S1). For GigaPath, we additionally evaluated its native slide-level encoder alongside MIL-based tile feature aggregation. Performance was assessed on biopsy and aspirate modalities using 5-fold cross-validation (CV), held-out test sets, and external validation.

On biopsy samples, H-Optimus-0 with ABMIL demonstrated higher cross-validation performance compared to the GigaPath slide encoder (0.82 vs. 0.77), although this difference was not reflected in the held-out test set, where performance was comparable (both 0.79). Across both specimen types, performance differences between models were generally modest and varied across evaluation settings, with no single model consistently outperforming others. Notably, the external validation cohort (n=95) was relatively small and enriched for high-grade MDS, which may influence performance estimates and limit its suitability as the primary criterion for model selection. Given these considerations, together with the advantages of maintaining a unified feature extraction framework across modalities, we retained GigaPath as the primary encoder for all subsequent experiments.

Tile Feature Extractor	Tile Feature Aggregator	Biopsy WSI			Aspirate WSI		
		5-fold CV AUC	Held-Out Test AUC	External	5-fold CV AUC	Held-Out Test AUC	External
GigaPath	ABMIL	0.80±0.05	0.82	0.90	0.77±0.06	0.66	0.65
	DSMIL	0.73±0.05	0.73	0.84	0.73±0.07	0.67	0.58
	TransMIL	0.74±0.05	0.80	0.92	0.76±0.05	0.66	0.60
	Slide Encoder	0.77±0.04	0.79	0.90	0.77±0.04	0.67	0.68
H-Optimus-0	ABMIL	0.82±0.04	0.79	0.93	0.77±0.07	0.68	0.66
	DSMIL	0.77±0.05	0.70	0.80	0.66±0.03	0.72	0.66
	TransMIL	0.77±0.05	0.78	0.95	0.75±0.08	0.67	0.62
Virchow2-CLS	ABMIL	0.81±0.04	0.76	0.80	0.78±0.05	0.74	0.59
	DSMIL	0.73±0.04	0.70	0.72	0.66±0.05	0.64	0.60
	TransMIL	0.78±0.05	0.74	0.94	0.75±0.05	0.67	0.68
MUSK	ABMIL	0.77±0.05	0.81	0.83	0.76±0.04	0.72	0.67
	DSMIL	0.70±0.05	0.77	0.82	0.68±0.04	0.66	0.68
	TransMIL	0.73±0.03	0.70	0.84	0.76±0.04	0.69	0.54

Table S1: Comparison of foundation models and aggregation strategies for feature extraction from biopsy and aspirate whole-slide images (WSIs). We evaluated four foundation models (GigaPath, H-Optimus-0, Virchow2-CLS, and MUSK) in combination with multiple-instance learning (MIL) aggregators (ABMIL, DSMIL, and TransMIL), along with the native slide-level encoder for GigaPath. Performance is reported as mean ± standard deviation of area under the receiver operating characteristic curve (AUC) for 5-fold cross-validation (CV), held-out test, and external validation cohorts. While certain model–aggregator combinations achieved higher performance on specific evaluation settings, differences were generally modest and varied across datasets, with no single approach consistently outperforming others across both biopsy and aspirate modalities. The external cohort (n=95) was relatively small and enriched for high-grade MDS, and thus results should be interpreted with caution.

TriPath model training and objective formulation

The TriPath model combines three modality-specific encoders—(1) BMB WSI encoder (Bx WSI), (2) BMA bag-of-cells hybrid ViT (Asp BoC), and (3) holistic BMA WSI encoder (Asp WSI)—using a gated fusion mechanism to generate a unified multimodal representation.

Fusion mechanism Each encoder produces a feature embedding $\mathbf{z}_m \in \mathbb{R}^{d_m}$ for modality $m \in \{\text{Asp WSI}, \text{Bx WSI}, \text{Asp BoC}\}$. These embeddings are linearly projected into a shared latent dimension $d = 512$:

$$\tilde{\mathbf{z}}_m = \mathbf{W}_m \mathbf{z}_m + \mathbf{b}_m.$$

The projected embeddings are concatenated and passed through a shallow gating network that outputs modality-specific weights α_m, normalized via a temperature-scaled softmax:

$$\alpha_m = \frac{\exp(h_m/T)}{\sum_{j\in\{\text{Asp WSI,Bx WSI,Asp BoC}\}} \exp(h_j/T)}, \quad T = 2.0,$$

where h_m denotes the raw gating logits. The fused embedding is obtained as a weighted combination of modality representations:

$$\mathbf{z}_{\text{fused}} = \sum_m \alpha_m \tilde{\mathbf{z}}_m.$$

The fused embedding is followed by layer normalization and dropout ($p = 0.3$) before classification.

Auxiliary supervision Each encoder includes an auxiliary classification head to preserve modality-specific discriminative capacity. Given label y and model predictions $\hat{y}$ (fused) and $\hat{y}_m$ (modality-specific), the total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{main}} + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}},$$

where

$$\mathcal{L}_{\text{main}} = \text{CE}(\hat{y}, y), \quad \mathcal{L}_{\text{aux}} = \frac{1}{3} \sum_{m} \text{CE}(\hat{y}_m, y),$$

and $\lambda_{\text{aux}} = 0.2$ controls the contribution of the auxiliary terms.

Optimization Training was performed using stochastic gradient descent (momentum 0.9, weight decay 0.05). Learning rates were set to 10^{-3} for the fusion and classification heads, and 10^{-4} for the unfrozen encoder layers. Layer-wise learning rate decay (0.95) was applied across the encoder depth. Models were trained for 20 epochs with gradient accumulation over 32 steps and dropout 0.3.

During training, the final two layers of each pretrained encoder were unfrozen, enabling the model to adapt higher-level representations while maintaining the integrity of pretrained features.

Implementation All models were implemented in PyTorch 2.2 and trained on NVIDIA A100 GPUs (80GB). For cross-validation experiments, each fold was trained independently, and the checkpoint achieving the highest validation accuracy within that fold was used for evaluation. Held-out and external cohorts were evaluated using a model trained on the entire cross-validation dataset (all folds combined) without additional fine-tuning. Performance metrics reported for cross-validation represent the mean and standard deviation across folds.

Model	**CV AUC (mean ± SD)**	**Held-out AUC (95% CI, n=129)**	**External AUC (95% CI, n=95)**	**Held-out ACC (95% CI, n=129)**	**External ACC (95% CI, n=95)**
Asp WSI	0.78 ± 0.06	0.72 (0.63-0.81)	0.65 (0.54-0.76)	0.66 (0.57-0.74)	0.63 (0.52-0.72)
Bx WSI	0.75 ± 0.04	0.81 (0.73-0.89)	0.90 (0.84-0.95)	0.76 (0.68-0.83)	0.78 (0.68-0.86)
Asp BoC (Bag-of-Cells)	0.78 ± 0.03	0.76 (0.68-0.84)	0.90 (0.82-0.96)	0.68 (0.60-0.76)	0.86 (0.77-0.92)
TriPath	**0.81 ± 0.01**	**0.84 (0.76-0.90)**	**0.92 (0.83-0.98)**	**0.81 (0.73-0.87)**	**0.84 (0.75-0.90)**

Table S2: Performance comparison across internal cross-validation (CV), held-out test, and external validation cohorts. CV AUC is reported as mean ± standard deviation across five folds. Held-out and external AUCs are reported with 95% confidence intervals (CIs) estimated via 1,000 bootstrap resamples. Held-out and external accuracies (ACC) are reported with 95% Clopper–Pearson exact CIs. Bold values indicate the best-performing model for each metric. TriPath integrates aspirate WSI (Asp WSI), biopsy WSI (Bx WSI), and aspirate bag-of-cells features (Asp BoC) via gated fusion.

Modality Contribution and Synergy

Model confidence was defined as the normalized absolute distance of the positive-class probability from the dataset-specific optimal threshold τ (computed on training/validation only). The thresholds used were: fold1=0.5, fold2=0.5, fold3=0.5, fold4=0.5, fold5=0.5, heldout=0.626, external=0.880.

Table S3: Per-fold 4×4 correlation matrices between fusion gate weights and model confidence across cross-validation (CV) folds, held-out test, and external datasets.

(a) Fold 1

	Asp WSI	Bx WSI	Asp BoC	Confidence
Asp WSI	1.000	0.885	-0.976	-0.043
Bx WSI	0.885	1.000	-0.966	-0.126
Asp BoC	-0.976	-0.966	1.000	0.083
Confidence	-0.043	-0.126	0.083	1.000

(b) Fold 2

	Asp WSI	Bx WSI	Asp BoC	Confidence
Asp WSI	1.000	-0.551	-0.512	0.057
Bx WSI	-0.551	1.000	-0.434	0.053
Asp BoC	-0.512	-0.434	1.000	-0.116
Confidence	0.057	0.053	-0.116	1.000

(c) Fold 3

	Asp WSI	Bx WSI	Asp BoC	Confidence
Asp WSI	1.000	-0.361	-0.039	0.153
Bx WSI	-0.361	1.000	-0.918	0.117
Asp BoC	-0.039	-0.918	1.000	-0.190
Confidence	0.153	0.117	-0.190	1.000

(d) Fold 4

	Asp WSI	Bx WSI	Asp BoC	Confidence
Asp WSI	1.000	-0.698	0.075	0.033
Bx WSI	-0.698	1.000	-0.766	-0.141
Asp BoC	0.075	-0.766	1.000	0.167
Confidence	0.033	-0.141	0.167	1.000

(e) Fold 5

	Asp WSI	Bx WSI	Asp BoC	Confidence
Asp WSI	1.000	-0.438	-0.048	-0.066
Bx WSI	-0.438	1.000	-0.877	-0.012
Asp BoC	-0.048	-0.877	1.000	0.049
Confidence	-0.066	-0.012	0.049	1.000

(f) Held-out (final model)

	Asp WSI	Bx WSI	Asp BoC	Confidence
Asp WSI	1.000	-0.750	0.551	-0.061
Bx WSI	-0.750	1.000	-0.965	0.098
Asp BoC	0.551	-0.965	1.000	-0.100
Confidence	-0.061	0.098	-0.100	1.000

(g) External validation

	Asp WSI	Bx WSI	Asp BoC	Confidence
Asp WSI	1.000	-0.788	0.643	0.004
Bx WSI	-0.788	1.000	-0.978	0.249
Asp BoC	0.643	-0.978	1.000	-0.310
Confidence	0.004	0.249	-0.310	1.000

Values denote Spearman correlation coefficients (ρ) between modality gate weights and model confidence.

Table S4: Summary of fusion gate weights and auxiliary head probabilities across cross-validation folds, held-out test, and external datasets (mean ± SD).

source	Fusion (Asp WSI)	Fusion (Bx WSI)	Fusion (Asp BoC)	Aux Head (Asp WSI)	Aux Head (Bx WSI)	Aux Head (Asp BoC)
fold1	0.200 ± 0.035	0.155 ± 0.030	0.646 ± 0.063	0.550 ± 0.106	0.494 ± 0.165	0.410 ± 0.150
fold2	0.157 ± 0.027	0.182 ± 0.025	0.661 ± 0.025	0.533 ± 0.449	0.479 ± 0.431	0.440 ± 0.309
fold3	0.227 ± 0.016	0.216 ± 0.039	0.557 ± 0.036	0.580 ± 0.423	0.568 ± 0.398	0.460 ± 0.314
fold4	0.331 ± 0.020	0.328 ± 0.030	0.340 ± 0.022	0.390 ± 0.395	0.439 ± 0.363	0.502 ± 0.051
fold5	0.324 ± 0.017	0.312 ± 0.035	0.364 ± 0.032	0.434 ± 0.439	0.542 ± 0.423	0.487 ± 0.177
heldout	0.151 ± 0.008	0.141 ± 0.025	0.708 ± 0.020	0.555 ± 0.192	0.585 ± 0.178	0.474 ± 0.124
external	0.155 ± 0.005	0.140 ± 0.020	0.705 ± 0.016	0.504 ± 0.146	0.622 ± 0.131	0.547 ± 0.084

Fusion weights reflect each modality's contribution within the TriPath gating mechanism; auxiliary heads report class-specific probabilities learned for each modality branch.

Calibration performance under domain shift

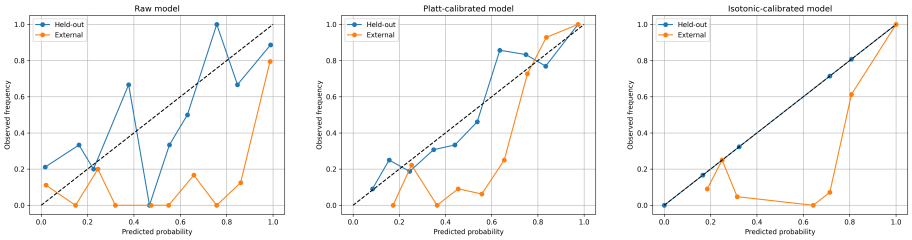
We next evaluated the calibration of predicted probabilities on the held-out and external cohorts (Figure S1(a)). On the held-out set, the model exhibited suboptimal calibration despite reasonable discrimination, with a Brier score of 0.180(95% CI: 0.126–0.238), calibration slope of 0.26, and expected calibration error (ECE) of 0.16, indicating overconfident predictions. Calibration performance further degraded on the external cohort (Brier score 0.241 (95% CI: 0.174–0.310), slope 0.58, ECE 0.28), consistent with systematic probability misestimation under domain shift.

To address this, we applied post-hoc recalibration using logistic (Platt) scaling, fitted on the held-out cohort. This substantially improved calibration on the held-out set (Brier score 0.162(95% CI: 0.128–0.198), slope 0.97, intercept -0.004, ECE 0.07), restoring near-ideal reliability (Figure S1(b)). Importantly, Platt scaling also improved calibration on the external cohort, reducing the Brier score from 0.241 to 0.182 (95% CI: 0.148–0.216), although residual miscalibration persisted (slope 1.89, intercept -1.70, ECE 0.24), indicating incomplete transfer of calibration across cohorts (Figure S1(b)).

We further evaluated isotonic regression to account for potential nonlinear probability distortions. While isotonic calibration achieved near-perfect calibration on the held-out cohort (Brier score 0.144 (95% CI: 0.110–0.181), slope 0.96, ECE ≈ 0), it did not generalize as effectively to the external dataset (Brier score 0.209 (95% CI: 0.160–0.259), ECE 0.25), suggesting overfitting to the held-out distribution (Figure S1(c)).

Notably, the optimal decision threshold differed substantially between cohorts (0.63 for held-out vs 0.88 for external), reflecting shifts in the predicted probability distribution. This threshold instability further underscores the impact of domain shift on probability calibration, even in the presence of improved discrimination performance on the external cohort.

Consistent with the fusion analysis, these findings suggest that modality-specific confidence amplification—particularly increased reliance on biopsy-derived features—may contribute to systematic distortion of predicted probabilities under domain shift.



(a) Raw model **(b)** Post-hoc Platt calibration **(c)** Post-hoc isotonic calibration

Figure S1: Calibration performance under domain shift before and after post-hoc recalibration. (a) Raw model predictions demonstrate miscalibration on both held-out and external cohorts, with greater deviation in the external dataset, indicating systematic probability distortion under domain shift. (b) Post-hoc logistic (Platt) recalibration improves calibration on the held-out cohort and partially corrects miscalibration on the external cohort, although residual deviations persist, likely reflecting domain shift associated with differences in slide quality and disease severity distribution. (c) Isotonic recalibration achieves near-perfect calibration on the held-out cohort but generalizes less effectively to the external dataset, consistent with overfitting of the non-parametric calibration model to the held-out probability distribution. The dashed line represents perfect calibration.

Subcategory Analysis Tables

Table S5: Accuracy of each model per diagnostic subcategory in the held-out test set. Values are presented as proportion correct with 95% Clopper–Pearson exact confidence intervals.

Model subcat	Asp BoC	Asp WSI	Bx WSI	TriPath
MDS-MLD (n=15)	0.67 [0.42–0.85]	0.53 [0.30–0.75]	0.67 [0.42–0.85]	0.67 [0.42–0.85]
MDS-RSMLD (n=11)	0.64 [0.35–0.85]	0.73 [0.43–0.90]	0.82 [0.52–0.95]	0.82 [0.52–0.95]
MDS-EB2 (n=10)	0.90 [0.60–0.98]	0.70 [0.40–0.89]	0.80 [0.49–0.94]	0.90 [0.60–0.98]
MDS-EB1 (n=12)	0.83 [0.55–0.95]	0.67 [0.39–0.86]	1.00 [0.76–1.00]	0.83 [0.55–0.95]
MDS-SLD (n=4)	0.25 [0.05–0.70]	0.50 [0.15–0.85]	0.25 [0.05–0.70]	0.50 [0.15–0.85]
MDS-RSSLD (n=6)	0.50 [0.19–0.81]	0.67 [0.30–0.90]	0.83 [0.44–0.97]	0.50 [0.19–0.81]
MDS withisolateddel(5q) (n=5)	0.40 [0.12–0.77]	0.60 [0.23–0.88]	0.60 [0.23–0.88]	0.60 [0.23–0.88]
MDS-U (n=1)	0.00 [0.00–0.79]	1.00 [0.21–1.00]	0.00 [0.00–0.79]	0.00 [0.00–0.79]
CCUS (n=55)	0.73 [0.60–0.83]	0.71 [0.58–0.81]	0.75 [0.62–0.84]	0.89 [0.78–0.95]
ICUS (n=10)	0.60 [0.31–0.83]	0.50 [0.24–0.76]	0.90 [0.60–0.98]	0.90 [0.60–0.98]

Note: Subcategories with smaller sample sizes (n) exhibit wider 95% confidence intervals, reflecting greater uncertainty in model accuracy estimates.

Table S6: Accuracy of each model per diagnostic subcategory in the external validation set. Values are presented as proportion correct with 95% Clopper–Pearson exact intervals.

Model subcat	Asp BoC	Asp WSI	Bx WSI	TriPath
MDS-MLD (n=0)	-	-	-	-
MDS-RSMLD (n=0)	-	-	-	-
MDS-EB2 (n=26)	0.92 [0.75–0.99]	0.69 [0.48–0.86]	0.81 [0.61–0.93]	0.92 [0.75–0.99]
MDS-EB1 (n=8)	0.75 [0.35–0.97]	0.62 [0.24–0.91]	0.88 [0.47–1.00]	0.75 [0.35–0.97]
MDS-SLD (n=1)	1.00 [0.03–1.00]	1.00 [0.03–1.00]	0.00 [0.00–0.97]	1.00 [0.03–1.00]
MDS-RSSLD (n=0)	-	-	-	-
MDS withisolateddel(5q) (n=1)	0.00 [0.00–0.97]	0.00 [0.00–0.97]	0.00 [0.00–0.97]	0.00 [0.00–0.97]
MDS-U (n=0)	-	-	-	-
CCUS (n=52)	0.83 [0.70–0.92]	0.62 [0.47–0.75]	0.75 [0.61–0.86]	0.79 [0.65–0.89]
ICUS (n=7)	1.00 [0.59–1.00]	0.86 [0.42–1.00]	0.86 [0.42–1.00]	1.00 [0.59–1.00]

Note: Subcategories with smaller sample sizes (n) exhibit wider 95% confidence intervals, reflecting greater uncertainty in model accuracy estimates.

Association of model predictions with genomic alterations

Table S7: Model performance stratified by mutation-group presence using the optimal classification threshold of 0.63.

Genomic group	Status	N	MDS	Non-MDS	Accuracy	Sensitivity	Specificity
Splicing	Present	37	27	10	0.784	0.778	0.800
Splicing	Absent	92	37	55	0.815	0.676	0.909
TP53	Present	6	6	0	1.000	1.000	–
TP53	Absent	123	58	65	0.797	0.690	0.892
Tumor Suppressor	Present	11	9	2	0.909	0.889	1.000
Tumor Suppressor	Absent	118	55	63	0.797	0.691	0.889
Epigenetic	Present	32	15	17	0.813	0.733	0.882
Epigenetic	Absent	97	49	48	0.804	0.714	0.896
ASXL1	Present	18	9	9	0.833	0.889	0.778
ASXL1	Absent	111	55	56	0.802	0.691	0.911
DNMT3A	Present	33	15	18	0.879	0.867	0.889
DNMT3A	Absent	96	49	47	0.781	0.673	0.894
Signaling	Present	9	6	3	0.889	0.833	1.000
Signaling	Absent	120	58	62	0.800	0.707	0.887

Sensitivity and specificity estimates for subgroups with small sample sizes (e.g., TP53 and signaling pathway alterations) should be interpreted cautiously.

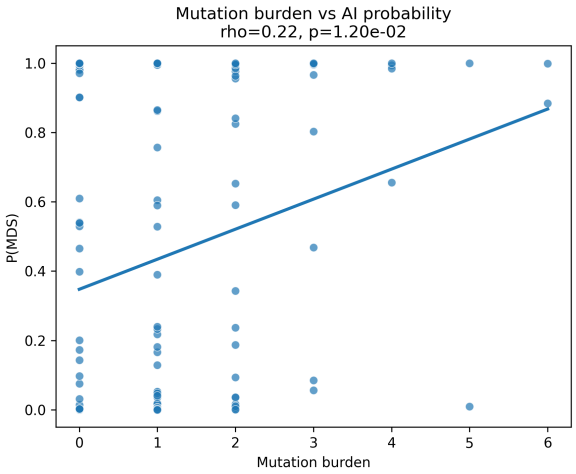


Figure S2: Association between mutation burden and AI-predicted probability of myelodysplastic syndrome (MDS). Each point represents an individual patient. Mutation burden was defined as the total number of detected variants across the targeted sequencing panel. The fitted regression line illustrates the overall positive association between mutation burden and model-predicted probability of MDS. A weak positive correlation was observed (Spearman $\rho = 0.22$, $p = 0.012$), indicating that cases harboring a greater number of genomic alterations tended to receive higher model-predicted MDS probabilities.

Table S8: Association between AI-predicted MDS probability and mutation-group presence. P-values were calculated using two-sided Mann–Whitney U tests comparing predicted MDS probabilities between mutation-present and mutation-absent cases. False discovery rate (FDR) correction was performed across all tested groups using the Benjamini–Hochberg procedure.

Mutation Group	n_{present}	n_{absent}	Median P(MDS) Present	Median P(MDS) Absent	p value	FDR
Splicing	37	92	0.956	0.154	6.52×10^{-4}	0.0046
TP53	6	123	0.999	0.237	0.0069	0.0240
Tumor Suppressor	11	118	0.963	0.238	0.0575	0.1342
ASXL1	18	111	0.728	0.233	0.1926	0.3371
DNMT3A	33	96	0.187	0.432	0.4365	0.5455
Signaling	9	120	0.984	0.367	0.4676	0.5455
Epigenetic	32	97	0.499	0.240	0.5777	0.5777

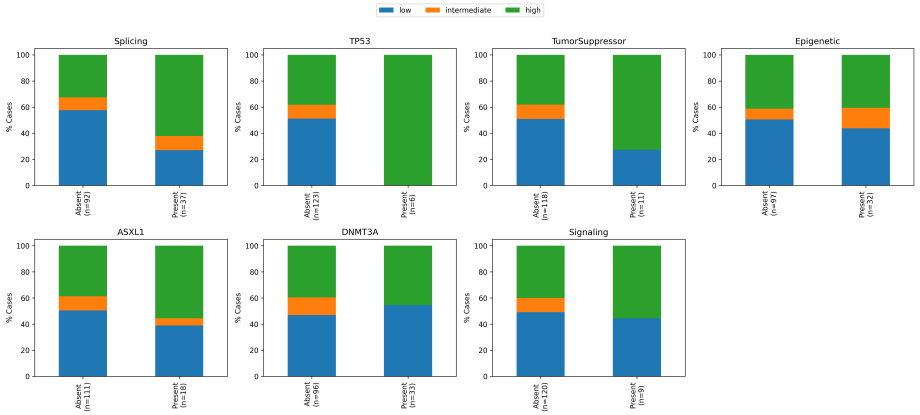


Figure S3: Distribution of AI-predicted MDS probability categories stratified by genomic alteration status. For each genomic group, cases were categorized as mutation absent or present and subdivided into low ($P_{\mathrm{MDS}} < 0.33$), intermediate ($0.33 \leq P_{\mathrm{MDS}} < 0.63$), and high ($P_{\mathrm{MDS}} \geq 0.63$) probability categories. The high-probability category corresponds to predictions exceeding the model operating threshold used for binary MDS classification ($P_{\mathrm{MDS}} = 0.63$). Splicing-factor and TP53-mutated cases demonstrated enrichment within the high-probability category, whereas other genomic groups showed more heterogeneous distributions. Sample sizes for mutation-negative and mutation-positive subsets are indicated beneath each bar group.

Reader Study

Integrated Gradients-based model explanation

To generate interpretable explanations for the TriPath model used in the reader study Arm B, we computed feature attributions using Integrated Gradients (IG) across all three model branches: aspirate whole-slide images (Asp WSI), biopsy whole-slide images (Bx WSI), and aspirate bag-of-cells (Asp BoC) representations. IG was implemented using the Captum interpretability library for PyTorch [2] with 80–120 interpolation steps depending on modality.

To enable modality-specific attribution while maintaining consistency with the full TriPath model, separate forward-pass wrappers were defined for each modality branch. Each wrapper fixed all non-target modalities while computing gradients only with respect to the modality of interest. A zero baseline was used for the tile features (Asp WSI and Bx WSI), while for cell image-based inputs (Asp BoC), a mean-image baseline was constructed by averaging cell images across spatial dimensions to preserve distributional consistency. IG attributions were computed per tile embedding and aggregated by summing across feature dimensions to obtain a tile attribution score for the WSI encoders. For Asp BoC encoder, attributions were computed over spatial cell-image tokens and aggregated across spatial dimensions to yield a single signed importance score per cell.

To ensure numerical stability of IG estimates, completeness (summation) checks were performed for each modality, verifying that the sum of attributions approximated the difference between model outputs at the input and baseline conditions. Convergence diagnostics were recorded via the IG internal convergence delta provided by Captum.

For visualization in the reader study, attributions were ranked per modality, and the top 25% most positively and negatively contributing tiles or cells were selected for display. Positive attributions correspond to features supporting the predicted class, whereas negative attributions indicate features opposing the prediction. For each slide, attributions were computed with respect to the predicted MDS class to ensure a consistent positive-class interpretation across both MDS and non-MDS cases.

Attribution maps were overlaid onto the original spatial coordinates of the slide. For biopsy slides, only Bx WSI-derived tiles were visualized. For aspirate slides, both Asp WSI-derived patches and Asp BoC-derived cellular regions were visualized jointly. This allowed simultaneous interpretation of tissue architecture (WSI-level) and cellular composition (BoC-level) within the same explanatory framework.

Table S9: Diagnostic accuracy comparison between study arms. Accuracy of each expert across modalities under Arm A (without AI assistance) and Arm B (with AI assistance). Δ indicates the change in accuracy (Arm B - Arm A). Statistical significance was assessed using McNemar's test.

Modality	**Expert**	**Arm A**	**Arm B**	Δ	p **(McNemar)**
Aspirate	E1	0.692	0.769	+0.077	0.688
Aspirate	E2	0.692	0.692	0.000	1.000
Biopsy	E1	0.654	0.731	+0.077	0.625
Biopsy	E2	0.615	0.808	+0.192	0.125
Combined	E1	0.731	0.808	+0.077	0.688
Combined	E2	0.654	0.846	+0.192	0.180

Table S10: Inter-expert agreement under AI assistance. Cohen's κ values for each modality under Arm A and Arm B. AI assistance improves agreement across all modalities, with the largest gain observed for combined diagnosis.

Modality	κ **(Arm A)**	κ **(Arm B)**	$\Delta\kappa$
Aspirate	0.095	0.385	+0.290
Biopsy	0.110	0.235	+0.126
Combined	0.071	0.497	+0.426

Table S11: Confidence shift with AI assistance. Mean diagnostic confidence (3-point ordinal scale) for each expert and modality under Arm A and Arm B. Δ indicates change in mean confidence. Statistical significance was assessed using the Wilcoxon signed-rank test.

Modality	**Expert**	**Mean A**	**Mean B**	Δ	p **(Wilcoxon)**	**Non-zero pairs**
Aspirate	E1	2.00	2.54	+0.54	0.001	18
Aspirate	E2	2.31	2.08	-0.23	0.293	11
Biopsy	E1	1.65	2.08	+0.42	0.012	13
Biopsy	E2	2.65	2.38	-0.27	0.052	13
Combined	E1	2.00	2.23	+0.23	0.109	14
Combined	E2	2.58	2.31	-0.27	0.052	13

Table S12: Error transition analysis with AI assistance. Number of cases corrected (incorrect to correct) and worsened (correct to incorrect) when transitioning from Arm A to Arm B. Net improvement reflects the balance between corrected and newly introduced errors.

Modality	Expert	Improved	Worsened	Net
Aspirate	E1	4	2	+2
Aspirate	E2	5	5	0
Biopsy	E1	3	1	+2
Biopsy	E2	6	1	+5
Combined	E1	4	2	+2
Combined	E2	7	2	+5

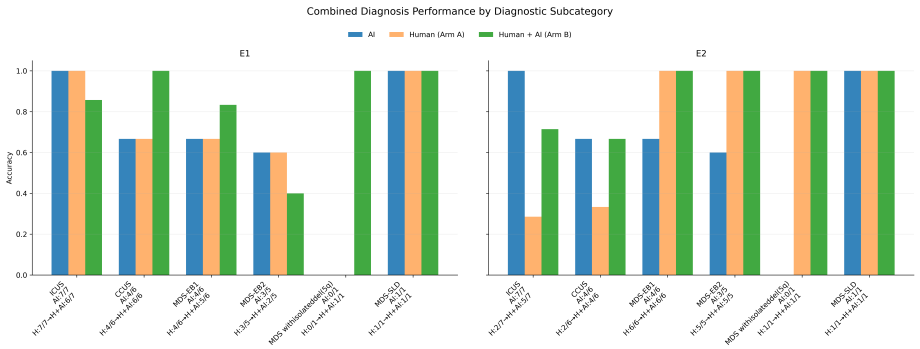


Figure S4: Comparison of standalone AI performance, unaided expert performance (Arm A), and AI-assisted expert performance (Arm B) across diagnostic subcategories for each reader. Numbers beneath each subcategory indicate the number of correctly classified cases out of the total number of cases in that subgroup. AI denotes standalone model performance, whereas H denotes expert performance without AI assistance followed by expert performance with AI assistance (H→H+AI). AI assistance demonstrated heterogeneous effects across disease categories. Notably, both readers showed improved performance in CCUS cases following AI-assisted review, whereas effects in other diagnostic subgroups were more variable. Given the limited number of cases within several subcategories, these findings should be considered exploratory and hypothesis-generating.