Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TriPath: A multimodal deep learning system for MDS diagnosis from joint bone marrow aspirate and biopsy slides

Palak Dave

palak.dave@moffitt.org

H. Lee Moffitt Cancer Center and Research Institute https://orcid.org/0000-0002-7107-3085

Ling Zhang

Moffitt Cancer Center

Mohammad Hussaini

H. Lee Moffitt Cancer Center and Research Institute

Jordan Rai

H. Lee Moffitt Cancer Center and Research Institute https://orcid.org/0009-0009-5066-4650

Sameer Andani

H. Lee Moffitt Cancer Center and Research Institute

Joseph Pierce

H. Lee Moffitt Cancer Center and Research Institute https://orcid.org/0009-0000-2720-5282

Mary Ellen Walker

H. Lee Moffitt Cancer Center and Research Institute

Nancy Gillis

Moffitt Cancer Center and Research Institute https://orcid.org/0000-0002-7744-8490

Lynn Moscinski

H. Lee Moffitt Cancer Center https://orcid.org/0000-0002-2609-4063

Mikkael Sekeres

University of Miami

Amy DeZern

Johns Hopkins University School of Medicine https://orcid.org/0000-0001-5225-4609

Eric Padron

Hematologic Malignancies, H. Lee Moffitt Cancer Center and Research Institute
https://orcid.org/0000-0002-4707-7916

Rami Komrokji

Moffitt Cancer Center

Issam El-Naqa

H. Lee Moffitt Cancer Center & Research Institute https://orcid.org/0000-0001-6023-1132

Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10743890/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

TriPath: A multimodal deep learning system for MDS diagnosis from joint bone marrow aspirate and biopsy slides

Palak Dave[1], Ling Zhang[2], Mohammad Hussaini[2], Jordan Rai[1], Sameer Andani[2], Joseph Pierce[1], Mary Ellen Walker[3], Nancy Gillis[3,6], Lynn C. Moscinski[7], Mikkael A. Sekeres[4], Amy E. DeZern[5], Eric Padron[6], Rami Komrokji[*6], and Issam El Naqa[*1]

[1]Departments of Machine Learning, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida
[2]Pathology and Laboratory Medicine, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida
[3]Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida
[4]Division of Hematology, Sylvester Comprehensive Cancer Center, University of Miami, Miami, Florida
[5]Division of Hematological Malignancies, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, Maryland
[6]Malignant Hematology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida
[7]Pathology (Hematopathology), H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida

October 2, 2026

[*]Co-senior authors
Corresponding authors: Palak Dave (palak.dave@moffitt.org); Issam El Naqa (issam.elnaqa@moffitt.org)

Abstract

Myelodysplastic syndromes (MDS) are clonal hematopoietic disorders characterized by ineffective hematopoiesis and a risk of progression to acute myeloid leukemia. Differentiation of MDS from idiopathic cytopenia of undetermined significance (ICUS) and clonal cytopenia of undetermined significance (CCUS), remains challenging due to the inherent subjectivity of bone marrow morphology assessment. Here, we present TriPath, a novel diagnostic deep learning framework that integrates bone marrow core biopsy and aspirate whole-slide images via a joint foundation and specialized, modality-specific encoders, with a gated fusion mechanism. TriPath was trained on a multi-institutional real-world cohort (n=645) involving rigorously adjudicated expert diagnoses from the MDS Natural History Study (NHS) (NCT02775383) and evaluated on both held-out (n=129) and external validation sets (n=95). The model

achieved robust performance across cross-validation (AUC of 0.81 ± 0.01), held-out (AUC=0.84), and external validation (AUC=0.92). Association of model predictions with genomic alterations indicated that the diagnostic performance remains robust across molecular subgroups while simultaneously assigning higher MDS probabilities to cases harboring genomic alterations in splicing factor genes and TP53 (mutation present vs absent p-value <0.01). We also conducted a pilot blinded multi-reader study to evaluate the effect of AI assistance on expert interpretation, assessing diagnostic accuracy, inter-reader agreement, confidence calibration, and human–AI interaction. The standalone AI model achieved performance comparable to expert readers on morphology-only cases, while AI-assisted review was associated with numerically higher diagnostic accuracy and improved inter-reader agreement (46.2% to 73.1%, Cohen's κ +0.426). Collectively, these findings suggest that the proposed system has the potential to function both as an autonomous diagnostic tool and as an AI-assisted decision-support system, reducing inter-observer variability and improving the reliability and efficiency of morphological assessment in hematopathology workflow.

Keywords: Myelodysplastic syndromes, Pathology, Deep learning, Multimodality.

1 Main

Myelodysplastic syndromes (MDS) are a group of hematologic malignancies characterized by ineffective hematopoiesis, leading to cytopenias, transfusion dependence, and reduced survival. As one of the most common blood cancers, MDS also carries a risk of progression to acute myeloid leukemia (AML), underscoring the importance of accurate diagnosis.

The diagnosis of MDS relies on the morphological assessment of peripheral blood and bone marrow (BM) specimens by specialized hematopathologists, with particular emphasis on identifying dysplastic features - a labor intensive and subjective process. Moreover, differentiation of MDS from idiopathic cytopenia of undetermined significance (ICUS) and clonal cytopenia of undetermined significance (CCUS), in which cytopenia and mild to absent morphologic abnormalities that do not meet MDS diagnostic criteria, remains challenging due to the inherent subjectivity of bone marrow morphological assessment. This subjectivity contributes to substantial inter-observer variability, with discordance rates of up to 40% between academic and community settings, affecting clinical management in approximately 26% of cases [34, 8].

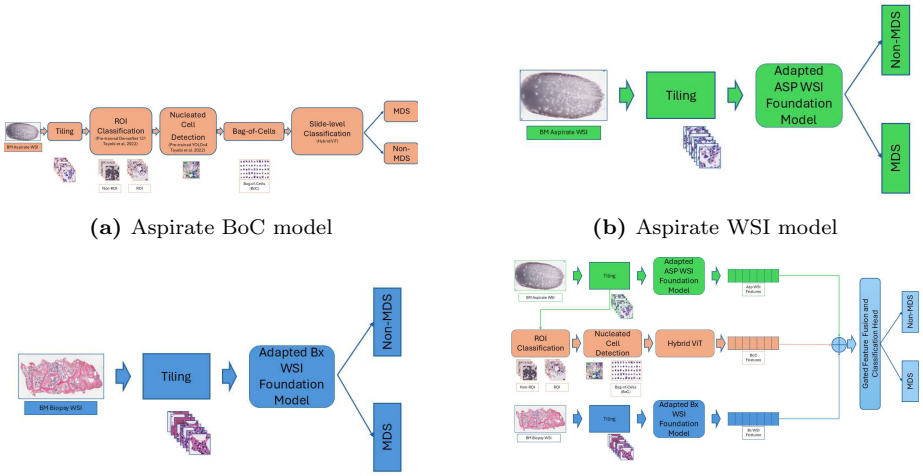
While automated cell counting in peripheral blood is well established in routine clinical practice, cytomorphologic evaluation of blood and BM smears for hematologic neoplasms remains largely manual and dependent on expert interpretation [6, 14]. Artificial intelligence (AI) offers a scalable approach to standardize morphologic assessment by enabling high-throughput analysis of whole slide images (WSIs), with the potential to improve diagnostic consistency and precision.

The evaluation of the bone marrow (BM) involves a biopsy (BMB) for structural and stromal evaluations and an aspirate (BMA) for a detailed examination of trilineage cellular morphology. Prior work in BM aspirate analysis has largely focused on automated detection and classification of nucleated cells [27, 15, 28, 26, 7, 33, 18, 30], often preceded by region-of-interest (ROI) localization within whole slide images (WSIs) [3, 25, 29]. However, applications to MDS remain limited to specific dysplastic features or disease subtypes [20, 13, 11]. These constraints reflect the substantial annotation burden required to capture lineage identity, the continuum of hematopoietic maturation, and the morphological heterogeneity of dysplasia.

More recent efforts have explored end-to-end, weakly supervised deep learning models for direct disease classification from bone marrow whole-slide images (WSIs), which are well suited for clinical tasks focused on overall diagnosis rather than differential cell counting. Early work using bone marrow biopsy (BMB)-derived tissue microarrays demonstrated the feasibility of distinguishing MDS from MDS/MPN overlap neoplasms [2]. Subsequent studies have extended BMB WSI-based approaches to a range of hematologic classification tasks, including differentiation of MDS from clinical mimics [19], discrimination of prefibrotic primary myelofibrosis from essential thrombocythemia [24], and classification of normal marrow, myelodysplastic neoplasms, and immune thrombocytopenic purpura based on megakaryocyte features [10]. Parallel efforts in BMA analysis have demonstrated the utility of deep learning for distinguishing acute promyelocytic leukemia from acute myeloid leukemia [17], as well as separating MDS from AML and donor samples [5]. Despite these advances, comprehensive WSI-based diagnostic frameworks for MDS remain limited [6], highlighting the need for approaches that can integrate complementary bone marrow modalities while accounting for the diffuse and heterogeneous nature of hematologic disease.

In this study, we present TriPath, a first diagnostic deep learning framework for discriminating myelodysplastic syndromes (MDS) from other cytopenia-associated conditions, namely idiopathic cytopenia of undetermined significance (ICUS) and clonal cytopenia of undetermined significance (CCUS). We leverage the whole-slide foundation model Prov-GigaPath [32] to capture structural tissue architecture in BMB and BMA (Figures 1b and 1c), alongside a hybrid vision transformer (ViT) model for cytomorphological analysis of BMA-derived Bag-of-Cells (BoC) (Figure 1a). The outputs of these complementary models are integrated within TriPath via a late-fusion gating strategy to generate the final diagnostic prediction (Figure 1d). TriPath was developed using only bone marrow aspirate and biopsy whole-slide images as model inputs. No genomic, cytogenetic, laboratory, or clinical variables were used during training or inference.

A multi-institutional real-world dataset along with gold-standard diagnoses made by pathologists with expertise in MDS, who themselves had a rigorous process for diagnosing MDS and adjudicating discrepancies in diagnosis [8] from the National Heart, Lung, and Blood Institute (NHLBI) MDS Natural History Study (NHS) (NCT02775383) [22] was used as the discovery cohort for model training and internal validation. The model was further validated using an



(a) Aspirate BoC model

(b) Aspirate WSI model

(c) Biopsy WSI model

(d) TriPath multimodal model

Figure 1: Overview of single-modality and multimodal architectures for bone marrow evaluation. (**a**) The *Aspirate BoC model* aggregates single-cell embeddings from bag-of-cells (BoC) derived from aspirate WSI using a hybrid Vision Transformer to produce WSI-level representations. (**b**) The *Aspirate WSI model* employs a fine-tuned WSI foundation model to extract global contextual information from the aspirate WSI. (**c**) The *Biopsy WSI model* utilizes a fine-tuned WSI foundation model to extract tissue-level contextual features from H&E-stained sections. (**d**) The *TriPath multimodal fusion model* integrates the three encoded representations via a gated fusion module that adaptively learns modality-specific contributions for final classification.

independent institutional cohort of 95 patients as an external validation set.

To enhance interpretability of the deep learning model, we performed a comprehensive explainability analysis of the model's behavior. We analyzed learned fusion gate weights to quantify modality contributions and assessed their stability across training folds, held-out data, and external cohorts. We further examined correlations between fusion weights and model confidence to probe adaptive behavior under domain shift. In addition, we conducted subgroup-level analyses to evaluate performance across clinically relevant diagnostic subcategories and molecular subgroups.

To evaluate the clinical utility beyond discriminative performance, we conducted a pilot, blinded, multi-reader study involving two board-certified hematopathologists, with 13 and 21 years of experience, assessing the impact of AI assistance on expert diagnostic behavior, including diagnostic accuracy, inter-reader agreement, confidence calibration, and human–AI interaction. Each expert independently reviewed bone marrow aspirate and biopsy WSIs from the same set of cases under two conditions: (i) Arm A (unassisted), in which only the WSIs were provided; and (ii) Arm B (AI-assisted), in which the WSIs were accompanied by Integrated Gradients–based visual explanation overlays highlighting regions contributing to the predicted diagnosis (Figure 8).

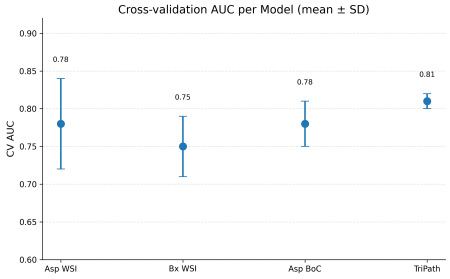
2 Results

2.1 Overall Performance

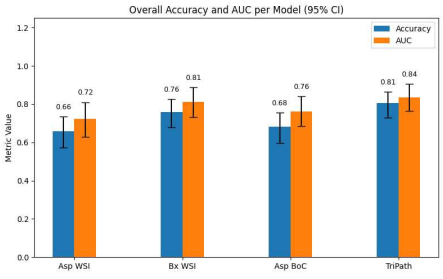
We first evaluated the performance of each single-modality model and the integrated TriPath model across a five-fold cross-validation (CV) folds, held-out test data, and an external validation cohort. Figure 2 (Table S2) summarizes the overall discriminative performance of each single-modality encoder and the integrated TriPath model. AUC values were interpreted as measures of discrimination, with values approaching 1.0 indicating better separation between MDS and non-MDS cases and 0.5 representing no discriminative ability. Single-modality models demonstrated moderate classification performance, with Asp BoC and Asp WSI showing comparable cross-validation (AUC=0.78) and Bx WSI modestly lower (AUC=0.75) on average. Across held-out and external cohorts, unimodal performance varied by modality, with Bx WSI performing strongest on the held-out set (AUC=0.81) and both Bx WSI and Asp BoC achieving high external AUCs (0.89). In all evaluations, TriPath achieved the highest AUC (0.81 CV, 0.84 Held-out, and 0.91 External), supporting the complementary value of multimodal integration.

2.2 Subcategory-wise performance analysis

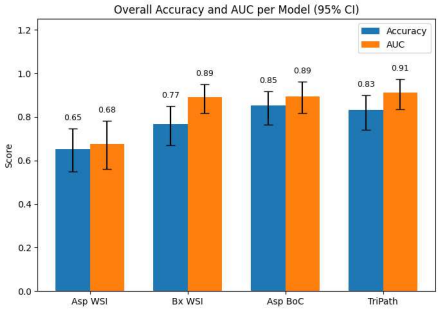
To further assess diagnostic robustness, model performance was evaluated across individual diagnostic subcategories in the held-out test set (Figure 3a, Table S5) and the external validation cohort (Figure 3b, Table S6).



(a) CV



(b) Held-out



(c) External

Figure 2: Model performance across cross-validation (a), held-out test (n=129) (b), and external validation (n=95)(c) cohorts. Cross-validation AUC is shown as mean ± s.d. across five folds. 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. TriPath integrates aspirate WSI (Asp WSI), biopsy WSI (Bx WSI), and aspirate bag-of-cells features (Asp BoC) via gated fusion and demonstrated consistent performance gains across evaluation settings.

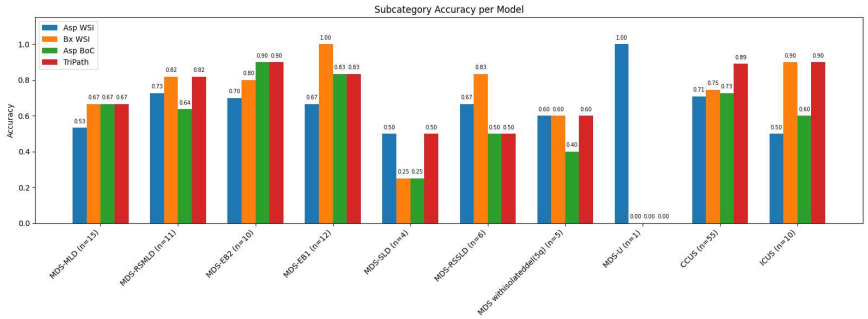
Within the held-out test set, performance was generally higher for higher-grade subtypes, particularly MDS with excess blasts 1 (MDS-EB1) and MDS with excess blasts 2 (MDS-EB2), across all models (Table S5), consistent with their more overt dysplastic features and increased blast burden. In contrast, lower-grade categories such as MDS with single lineage dysplasia (MDS-SLD) and MDS with ring sideroblasts and single lineage dysplasia (MDS-RSSLD) showed more variable and overall modest accuracies, likely reflecting both limited sample sizes and subtler morphologic abnormalities. Although MDS-RSSLD and MDS with ring sideroblasts and multi-lineage dysplasia (MDS-RSMLD) were analyzed separately from their non–ring sideroblast counterparts, the defining ring sideroblast morphology requires iron staining and is not directly visible on Wright-stained aspirates or H&E-stained biopsy sections, which may further limit discriminability in routine histologic inputs.

Notably, TriPath demonstrated improved performance in CCUS (0.89 [0.78–0.95]) compared to the strongest single-modality model (0.75 [0.62–0.84]), suggesting that integrating aspirate, biopsy, and smear-derived representations enhances discrimination in borderline or early-stage disease. For ICUS, both TriPath and the biopsy-based model achieved comparable performance (0.90 [0.60–0.98]), indicating that biopsy morphology alone may already provide a strong discriminative signal in certain low-cellularity contexts. Overall, TriPath matched or exceeded the best-performing unimodal model in most diagnostic subcategories, while performance variability in MDS-EB1, MDS-RSSLD, and MDS unclassifiable (MDS-U) suggests that the benefits of multimodal fusion are not uniform across all disease presentations, although they remain broadly consistent across the majority of clinically relevant groups.

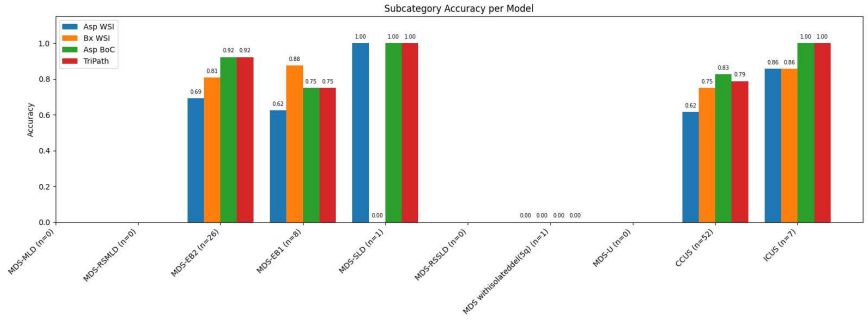
A broadly similar pattern was observed in the external validation cohort. High-grade subtypes (MDS-EB2 and MDS-EB1) were classified with accuracies of 0.92 [95% CI: 0.75–0.99] and 0.75 [95% CI: 0.35–0.97], respectively. Interpretation of performance in lower-prevalence subcategories was limited by small sample sizes (e.g., MDS-SLD and isolated del(5q), n=1), resulting in wide confidence intervals. Several subcategories were not represented in the external cohort (MDS-MLD, MDS-RSMLD, MDS-RSSLD, and MDS-U), reflecting cohort-specific sampling distributions.

2.3 Association of morphology-based model predictions with genomic alterations

Because TriPath was trained and evaluated exclusively using bone marrow morphology, without access to genomic information, we next investigated whether its morphology-based predictions nevertheless reflected underlying molecular alterations. To this end, we analyzed targeted sequencing data available for the held-out cohort using a clinically curated 96-gene myeloid neoplasm panel. We first quantified overall mutation burden and then evaluated expert-defined molecular groups representing recurrent genomic programs in myeloid neoplasia. These groups included splicing factor genes (*SF3B1, SRSF2, U2AF1, ZRSR2*), *TP53*, tumor suppressor/cohesin-related genes (*RUNX1,*



(a) Held-out test set



(b) External validation set

Figure 3: Subcategory-level diagnostic performance across held-out and external cohorts. Bars represent subgroup accuracy without confidence intervals for visual clarity. Exact estimates with 95% confidence intervals are provided in the Supplementary Tables S5 and S6.

BCOR, BCORL1, STAG2, RAD21), epigenetic regulators (*TET2, IDH1, IDH2*), *ASXL1*, *DNMT3A*, and signaling pathway genes (*JAK2, KRAS, NRAS, FLT3, CBL, CBLB, CBLC, PTPN11*).

A modest positive association was observed between mutation burden and morphology-based model-predicted probability of MDS (Spearman $\rho = 0.22$, $p = 0.012$), indicating that cases harboring a greater number of genomic alterations tended to receive higher MDS probabilities (Figure S2). Although statistically significant, the magnitude of the correlation was relatively small, suggesting that mutation burden explains only a portion of the variability in model predictions.

To determine whether genomic alterations influenced morphology-based model performance, classification accuracy was evaluated separately within mutation-positive and mutation-negative subsets using the predefined operating threshold (P(MDS)$>$ 0.63) (Figure 4, Table S7). Performance remained broadly consistent across molecular subgroups. For example, overall accuracy was 81.5% in splicing-negative cases and 78.4% in splicing-positive cases, while epigenetic-positive and epigenetic-negative cases achieved similar accuracies (81.3% and 80.4%, respectively). Higher accuracies were observed in several mutation-positive groups, including *DNMT3A*-mutated (87.9%), tumor suppressor/cohesin-mutated (90.9%), and *TP53*-mutated (100%) cases; however, these subsets contained relatively small numbers of patients. Collectively, these findings suggest that model discrimination is maintained across diverse genomic backgrounds rather than being restricted to a specific mutational subtype.

We next investigated whether specific molecular alterations were associated with higher morphology-based model-assigned risk/MDS probability. Presence of splicing-factor mutations was associated with substantially higher predicted probabilities of MDS compared with mutation-negative cases (median probability 0.96 vs. 0.15; FDR-adjusted $p = 0.0046$). Similarly, *TP53*-mutated cases demonstrated markedly elevated model probabilities (median probability 1.00 vs. 0.24; FDR-adjusted $p = 0.024$). In contrast, tumor suppressor/cohesin-related genes, *ASXL1*, *DNMT3A*, epigenetic regulators, and signaling pathway genes were not significantly associated with model probability following correction for multiple testing (Figure S3; Table S8).

Taken together, these results indicate that TriPath, despite using only morphologic whole-slide images as input, maintains robust diagnostic performance across diverse molecular subgroups while assigning higher MDS probabilities to cases harboring specific genomic alterations strongly associated with myelodysplastic neoplasia. The concordance between morphology-derived predictions and established molecular risk features suggests that the model captures clinically meaningful morphologic correlates of underlying disease biology, even without direct access to genomic information.

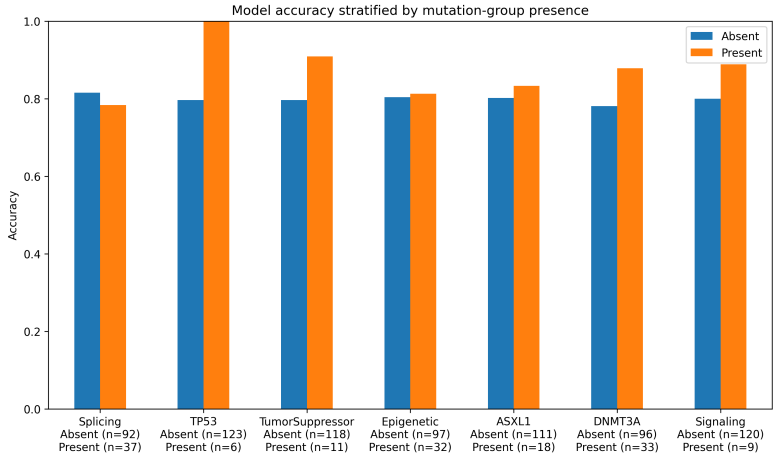


Figure 4: Morphology-based model performance stratified by mutation-group presence. Accuracy was calculated using the predefined classification threshold and compared between mutation-positive and mutation-negative cases for each expert-defined molecular category. Across all genomic subgroups, classification accuracy remained stable. Although modest differences were observed for *TP53*-mutated, tumor suppressor–mutated, *DNMT3A*-mutated, and signaling pathway–mutated cases, these groups contained relatively small numbers of patients. Overall, the results suggest that model discrimination is preserved across diverse molecular backgrounds rather than being driven by a single genomic subtype.

2.4 Modality Contribution and Synergistic Discrimination

In the TriPath model, the gated fusion module learns sample-specific gate weights that determine how much each modality—aspirate WSI, biopsy WSI, and aspirate BoC features—contributes to the final prediction. Analyzing the gate-weight distribution provides insight into how the model integrates multimodal information and whether its decision strategy remains stable across the evaluation settings. Stable gating patterns indicate consistent multimodal reasoning, while shifts in modality weighting may reflect cohort- or domain-specific dependencies.

2.4.1 Fusion gate weight analysis

We analyzed the fusion gate to quantify modality contributions. For each sample, we recorded the three per-modality weights ($\alpha_{\text{Asp WSI}}$, $\alpha_{\text{Bx WSI}}$, $\alpha_{\text{Asp BoC}}$; summing to 1). The fusion-weight distributions (Figure 6a) show moderate fold-to-fold variability, yet the aspirate-derived BoC modality (Asp BoC) consistently received the highest fusion weight across all folds, including held-out and external validations. Across the five cross-validation folds, the mean fusion weights were 0.25 for Asp WSI, 0.24 for Bx WSI, and 0.51 for Asp BoC.

The held-out and external models maintained this trend, assigning the highest weight to Asp BoC (0.71 ± 0.02 and 0.71 ± 0.02, respectively) (Table S4). This indicates that the fused model systematically relies more on discriminative cues embedded in cytological features from aspirate smears to inform its final predictions—an observation aligned with manual diagnostic reasoning, where cellular morphology in aspirate smears provides the most critical morphological diagnostic evidence, while histomorphologic features from biopsy and whole-slide aspirate imaging contribute complementary secondary information.

Interestingly, although the Asp BoC modality exhibits limited class separation when considered alone (Figure 5), the TriPath model assigns it the highest fusion weight across folds. This suggests that, while its standalone diagnostic signal is weak, Asp BoC provides complementary morphological cues that become highly informative when integrated with WSI features—highlighting the model's ability to exploit synergistic, cross-modality interactions rather than simple additive effects.

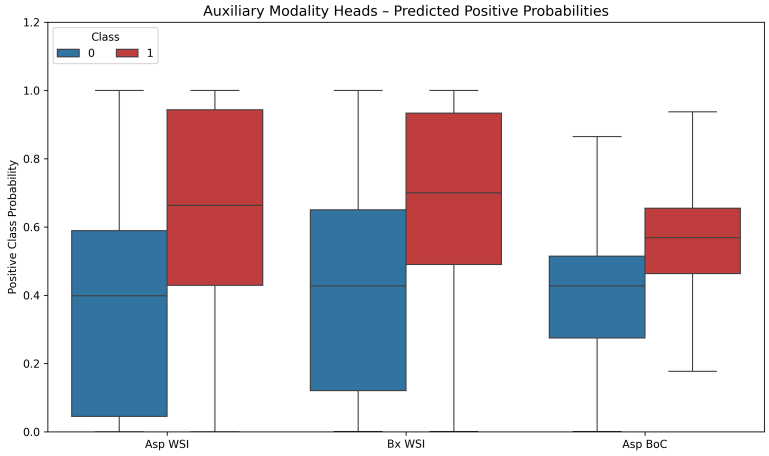


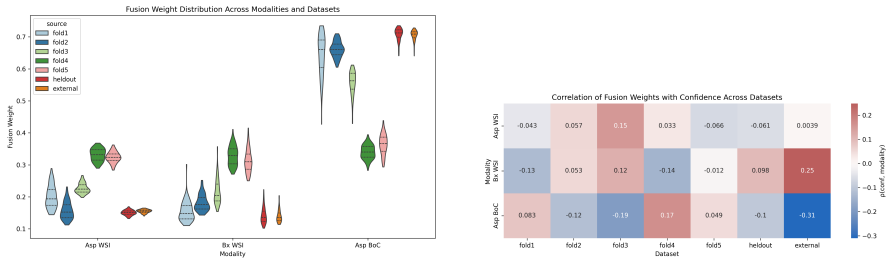
Figure 5: Auxiliary modality heads – predicted positive class probabilities stratified by true label. Boxplots show the distribution of auxiliary predictions for each modality (Asp WSI, Bx WSI, Asp BoC) across Non-MDS (blue) and MDS (red) samples. Outliers are omitted for clarity.

2.4.2 Relationship between fusion weights and model confidence

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). Fold-wise thresholds are provided in the Supplementary Materials. Correlation analysis between modality weights and model confidence (Figure 6b; Table S3) revealed weak relationships across cross-validation folds and held-out set ($|\rho| < 0.2$), suggesting that no single modality disproportionately influenced confidence. In contrast, the external

cohort displayed a moderate positive correlation for biopsy weights ($\rho = +0.25$) and a moderate negative correlation for BoC weights ($\rho = -0.31$), suggesting adaptive re-weighting under domain shift. These findings imply that the fusion mechanism is stable under internal validation but flexible enough to adjust modality reliance when faced with possible distributional changes.

Together, these analyses provide mechanistic interpretability, revealing that TriPath's decision logic is governed by stable yet context-adaptive cross-modal weighting, reflecting biologically grounded diagnostic reasoning.



(a) Distribution of modality-specific gate weights across datasets.

(b) Correlation (ρ) between fusion weights and prediction confidence across datasets.

Figure 6: Gate weight explainability analysis. (**A**) Violin plots showing the distribution of learned fusion (gate) weights for each modality (*Asp WSI*, *Bx WSI*, and *Asp BoC*) across cross-validation folds, held-out, and external cohorts. While the relative dominance of modalities varies across cross-validation folds, the overall distribution of gate weights remains within a comparable range, reflecting balanced multimodal integration. Held-out and external cohorts reveal consistent Asp BoC dominance, suggesting that BoC-derived features provide the most stable and transferable signal across domains. (**B**) Summary heatmap depicting correlation coefficients (ρ) between each modality's gate weight and prediction confidence. Weak correlations ($|\rho| \approx 0$) suggest that gating behavior is largely independent of prediction certainty, indicating that the model integrates information consistently rather than over-relying on any modality only when confident—supporting robust and context-adaptive fusion. Marked correlations in the external cohort indicate adaptive re-weighting under domain shift.

2.4.3 Modality synergy

To assess the synergistic effect of multi-modal integration, we performed ablation experiments in which the TriPath model was trained with subsets of the three available modalities, while maintaining identical cross-validation splits and training settings as the main model (Table 1). This included both single-modality and dual-modality configurations to isolate individual and joint contributions. Specifically, the evaluated settings were: (i) aspirate BoC only (A1), (ii) aspirate WSI only (A2), (iii) biopsy WSI only (B), (iv) aspirate BoC + biopsy WSI (A1B), (v) aspirate WSI + biopsy WSI (A2B), and (vi) aspirate BoC + aspirate WSI (A1A2). The full TriPath configuration (A1A2B) served

as a reference.

Table 1: Cross-validation AUCs for single- and dual-modality configurations compared to the full TriPath model.

Configuration	Mean AUC ± SD	ΔAUC vs. Full
A1 (Asp BoC only)	0.78 ± 0.03	–0.03
A2 (Asp WSI only)	0.78 ± 0.06	–0.03
B (Bx WSI only)	0.75 ± 0.04	–0.06
A1B (Asp BoC + Bx WSI)	0.77 ± 0.04	–0.04
A2B (Asp WSI + Bx WSI)	0.78 ± 0.03	–0.03
A1A2 (Asp BoC + Asp WSI)	0.77 ± 0.02	–0.04
A1A2B (Full TriPath)	**0.81 ± 0.01**	–

Among individual modalities (reported previously in Section 2), both aspirate-based views (BoC and WSI) perform comparably (AUC $\approx$ 0.78), while the biopsy WSI alone lags slightly behind (AUC = 0.75), suggesting that cytological morphology provides a stronger independent signal. However, combining all three modalities (A1A2B) yields the highest and most stable performance (AUC=0.81 ± 0.01), outperforming all dual- and single-modality variants by 0.03–0.06. These results complement the fusion-weight findings (Section 2.4.1), showing that while the BoC-derived modality provides the strongest single-modality performance, maximal diagnostic utility is achieved only when its features are integrated with histological context—underscoring the cooperative nature of cytological and tissue-level information within the TriPath framework.

2.5 Reader Study: Clinical Impact of AI Assistance

Diagnostic performance AI assistance resulted in a consistent trend toward improved diagnostic accuracy across modalities (Figure 7(A), Table S9), with the largest gains observed for Expert 2 (combined diagnosis: 65.4% to 84.6%). Expert 1 showed smaller but consistent improvements (73.1% to 80.8%). However, these differences did not reach statistical significance (McNemar's test, $p > 0.1$), likely reflecting the limited sample size.

Inter-expert agreement AI assistance substantially improved inter-expert agreement (Figure 7(B), Table S10). Cohen's κ increased from 0.071 (slight agreement) under Arm A to 0.497 (moderate agreement) under Arm B for the combined diagnosis ($\Delta = +0.426$). Similar trends were observed across aspirate and biopsy modalities. These findings indicate that AI assistance enhances diagnostic consistency between experts and reduces inter-observer variability.

Confidence calibration AI assistance induced modality- and reader-dependent shifts in diagnostic confidence (Figure 7(C), Table S11). Expert 1 demonstrated a significant increase in confidence (mean across cases) under AI assistance (aspirate: 2.00 to 2.54, $p = 0.001$; biopsy: 1.65 to 2.08, $p = 0.012$), with a similar but non-significant increase in combined assessment (2.00 to 2.23, $p = 0.109$). In contrast, Expert 2 exhibited a consistent decrease in confidence across modalities (aspirate: 2.31 to 2.08, $p = 0.29$; biopsy: 2.65 to 2.38, $p = 0.052$; combined: 2.58 to 2.31, $p = 0.052$), none of which reached statistical significance.

Despite these directional differences, all comparisons included 11–18 non-zero paired differences, suggesting a systematic shift rather than random fluctuation. Collectively, these findings are consistent with a calibration effect, whereby AI assistance increases confidence in less certain readers while tempering confidence in more confident readers, thereby reducing inter-reader variability.

Error correction dynamics AI assistance resulted in a net improvement in diagnostic correctness across readers (Figure 7(D), Table S12). For the combined diagnosis, Expert 1 corrected 4 previously incorrect cases while introducing 2 new errors, whereas Expert 2 corrected 7 cases and introduced 2 errors. This consistent imbalance in favor of error correction suggests that AI assistance primarily acts to resolve ambiguous cases rather than introduce systematic misclassification.

Expert-AI agreement We further evaluated the relationship between expert agreement with model overlay and diagnostic performance in Arm B. While higher agreement was associated with improved accuracy in certain biopsy evaluations, this relationship was not consistent across modalities or readers. These findings likely reflect limited sample size and potential differences in feature utilization between human experts and the model, and should be interpreted as exploratory.

Together, these results demonstrate that the proposed multimodal TriPath framework not only improves diagnostic performance but also enhances human decision consistency through improved agreement, calibrated confidence, and asymmetric error correction.

2.5.1 Human vs AI vs Human+AI Evaluation

Although the reader-study cohort was intentionally enriched for both model-correct and model-incorrect cases and therefore should not be interpreted as a prevalence-representative diagnostic cohort, the AI model demonstrated performance comparable to expert readers, with more than 13 years of experience, on the same set of morphology-only cases. For the combined diagnosis task, standalone model accuracy was 73.1%, compared with 73.1% and 65.4% for Expert 1 and Expert 2, respectively (Figure 9a).

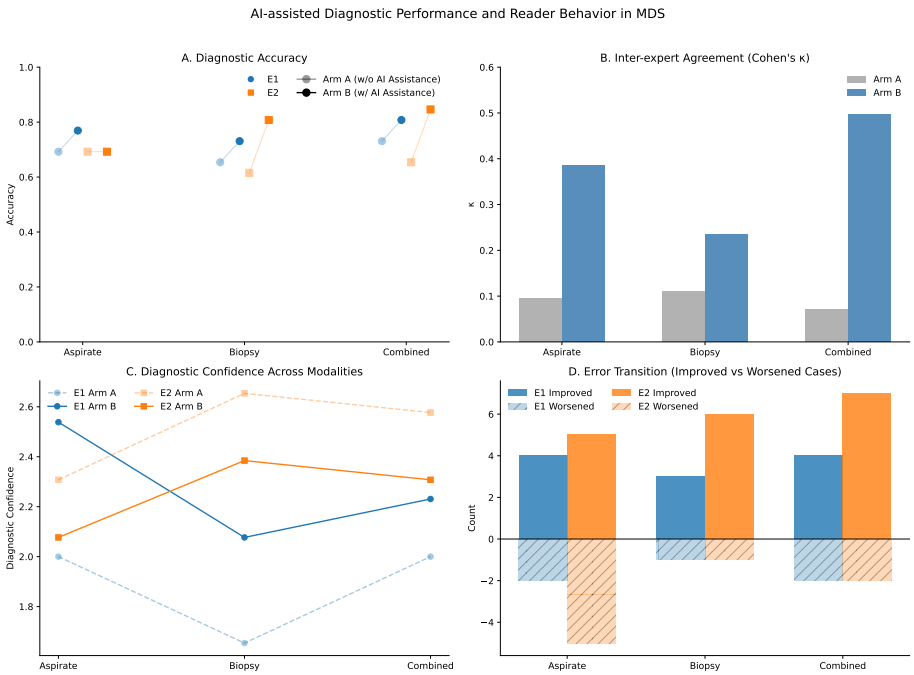


Figure 7: AI-assisted diagnostic performance, inter-reader agreement, and behavioral shifts in MDS evaluation. **(A)** Diagnostic accuracy of two expert hematopathologists (E1, E2) across aspirate, biopsy, and combined modalities under Arm A (w/o AI Assistance) and Arm B (w/ AI Assistance). Points represent paired performance on identical cases, enabling within-reader comparison across study arms. **(B)** Inter-expert agreement quantified using Cohen's κ for each modality under both arms, demonstrating improved diagnostic consistency with AI assistance. **(C)** Diagnostic confidence trajectories across modalities for each expert. Lines depict mean confidence scores (1-3 ordinal scale) across cases, with lighter lines indicating Arm A and darker lines indicating Arm B; colors denote expert identity (E1: blue; E2: orange). Convergence of expert trajectories under Arm B indicates reduced inter-reader variability in confidence. **(D)** Error transition analysis showing the number of cases that improved (incorrect to correct) or worsened (correct to incorrect) with AI assistance for each expert and modality. All analyses are based on 26 paired cases.

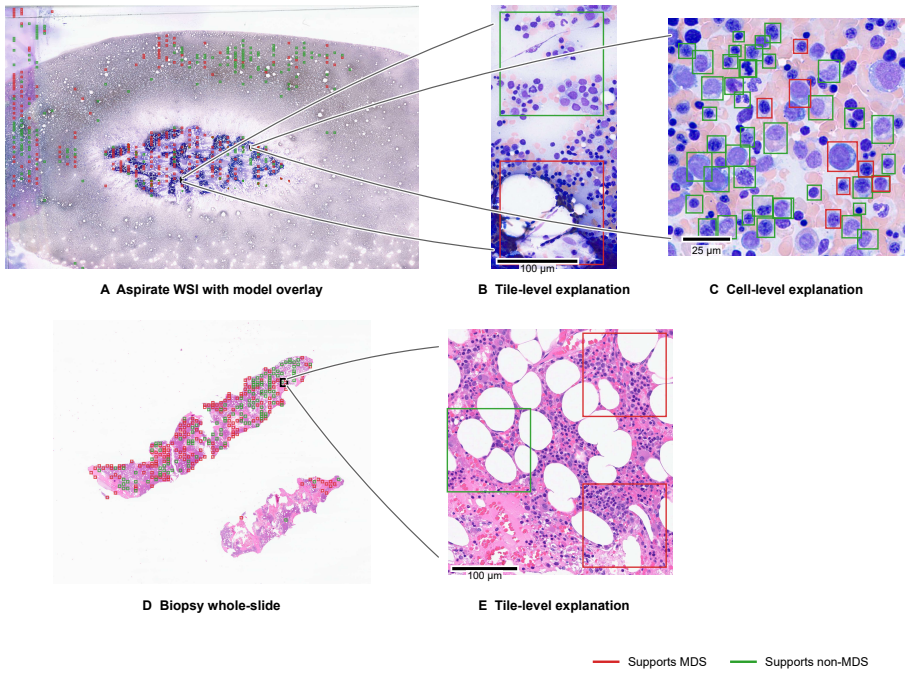


Figure 8: AI-generated explanation overlays used in the reader study (Arm B). Representative examples of model-derived explanation maps presented to expert readers during the AI-assisted evaluation are shown for aspirate and biopsy specimens from a case of MDS. (**A**) Aspirate whole-slide image (WSI) with overlaid regions identified by the model as contributing to the predicted diagnosis. (**B**) Zoomed tile-level region highlighting contributions from the aspirate WSI branch of TriPath. (**C**) Higher magnification view illustrating cell-level contributions from the aspirate bag-of-cells branch. (**D**) Biopsy WSI overview with corresponding model-highlighted regions. (**E**) Zoomed biopsy region showing tile-level contributions from the biopsy WSI branch. Red overlays indicate regions supporting MDS prediction, whereas green overlays indicate regions supporting non-MDS classification. Attributions were computed using Integrated Gradients, ranked separately within each modality, and the top 25% of tiles or cells with the highest positive and negative contributions were selected for visualization. Scale bars: 100 μm (tile-level), 25 μm (cell-level).

Importantly, AI assistance improved reader performance beyond either the model or the unaided experts alone. For the combined diagnosis task, Expert 1 improved from 73.1% to 80.8% accuracy, while Expert 2 improved from 65.4% to 84.6%. Similar trends were observed for aspirate-only and biopsy-only assessments.

To better understand the interaction between readers and AI recommendations, we stratified cases according to whether the AI prediction was correct or incorrect (Figure 9b). These analyses were performed to evaluate reader behavior and should not be interpreted as estimates of population-level diagnostic performance.

When the AI prediction was correct, AI assistance frequently improved reader accuracy. Importantly, AI assistance did not result in systematic propagation of incorrect model predictions. Among model-incorrect cases, readers often maintained their original performance or improved after reviewing the AI output. Across combined-diagnosis assessments, Expert 1 and Expert 2 correctly overrode 57.1% and 85.7% of incorrect AI predictions, respectively. These findings indicate that readers retained independent diagnostic judgment and critically evaluated AI recommendations rather than adopting them indiscriminately. Exploratory analyses stratified by diagnostic subcategory is presented in the Supplementary material (Figure S4).

These findings indicate that the proposed system possesses dual clinical utility. First, the model can function as an autonomous morphology-based diagnostic classifier with performance comparable to expert review. Second, when integrated into the diagnostic workflow, AI assistance augments expert interpretation and yields higher accuracy than either the AI model or the expert alone.

3 Discussion

In this study, we present a first multimodal deep learning system for discriminating myelodysplastic syndromes (MDS) from other cytopenia-related conditions using bone marrow aspirate and biopsy specimens. The TriPath model integrates aspirate-derived bag-of-cells (BoC), aspirate WSI, and biopsy WSIs, and adaptively combines these complementary modalities through a gated fusion mechanism.

TriPath was trained on a large, real-world, population-based dataset with rigorously adjudicated expert diagnoses as gold-standard diagnoses from the MDS NHS cohort leading to robust performance in the held-out set and an external validation set, emphasizing generalizability to routine digital slides. Interestingly, the TriPath model exhibits higher performance on the external cohort compared to the held-out internal test set, which is counterintuitive given that external validation typically presents additional variability. One contributing factor is that the internal held-out set is multi-institutional, encompassing a broader range of staining, slide preparation, and imaging variability, as well as a higher proportion of diagnostically challenging cases such as low-grade

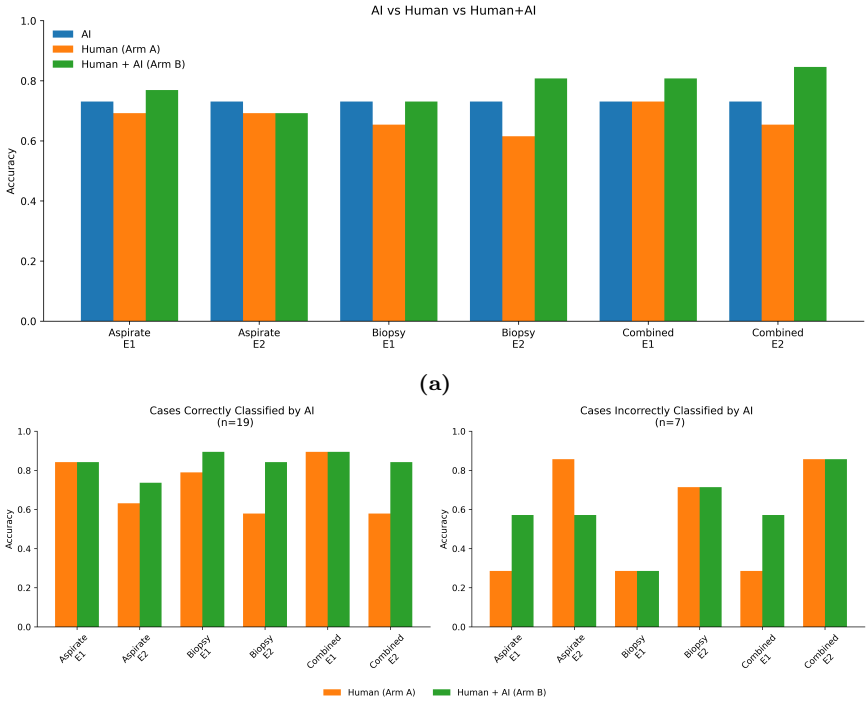


Figure 9: (a) Comparison of standalone AI performance, unaided expert performance (Arm A), and AI-assisted expert performance (Arm B) across aspirate, biopsy, and combined diagnostic assessments for two experts (E1 and E2) in the reader study. Although the reader-study cohort was intentionally enriched for both model-correct and model-incorrect cases, the AI model achieved performance comparable to expert readers on the same morphology-only cases, while AI assistance often improved reader performance beyond either the model or the unaided experts alone. (b) Reader performance stratified according to whether the AI prediction was correct or incorrect. Correct AI recommendations frequently improved reader performance, whereas readers often maintained performance or overrode erroneous AI recommendations when the model was incorrect, supporting the role of the system as a decision-support tool while preserving independent clinical judgment.

MDS. In contrast, the external set originates from a single institution with more standardized slide preparation and a case mix skewed toward higher-grade or morphologically distinct samples, making classification relatively easier. This suggests that cohort composition and pre-analytical variability can influence observed performance, while also confirming that TriPath generalizes robustly to unseen institutional data without retraining.

Fusion weight analysis suggests that cytological features contribute the strongest individual signal; however, optimal diagnostic performance is achieved only when these features are integrated with histological and architectural context, underscoring the complementary roles of cellular- and tissue-level information. This synergistic effect is further corroborated by ablation experiments, in which the TriPath model consistently outperformed unimodal and dual-modal configurations across all evaluation settings. Collectively, these findings demonstrate that the model leverages biologically meaningful feature integration to enhance diagnostic accuracy, aligning closely with established expert diagnostic reasoning in MDS.

Fusion weight–confidence analysis further revealed increased reliance on biopsy-derived features in the external cohort, which predominantly consisted of higher-grade MDS cases and originated from a single institution with more uniform slide quality, potentially yielding stronger and more homogeneous morphologic signals. Despite improved discrimination performance (AUC), the substantial shift in the optimal operating threshold (0.63 in the held-out cohort versus 0.88 in the external cohort) indicates that raw predicted probabilities were not directly transferable across institutions. This observation highlights the distinction between discrimination and calibration: while the model accurately ranked cases, the absolute probabilities required for clinical decision-making changed under external domain shift. Consistent with this, the external cohort demonstrated substantial probability miscalibration (Figure S1), reflected by worsening Brier score (0.241 vs 0.180 in the held-out cohort) and increased expected calibration error. Post-hoc Platt recalibration partially improved probability calibration, whereas isotonic regression demonstrated limited external generalization, consistent with overfitting to the held-out probability distribution. As is well recognized in computational pathology, probability calibration often generalizes less robustly than discrimination across institutions. Together, these findings emphasize the importance of evaluating calibration independently of discrimination and suggest that institution-specific probability recalibration may be beneficial prior to clinical deployment. Detailed calibration analyses are provided in the Supplementary Materials.

An important finding from our genomic analyses was that model accuracy remained robust across diverse molecular backgrounds while the model-assigned MDS probability showed biologically meaningful associations with specific genomic alterations. Despite being developed using morphologic whole-slide images without genomic input, the model maintained comparable classification accuracy across mutation-positive and mutation-negative subsets, suggesting that its performance was not dependent on a single molecular subtype. At the same time, cases harboring splicing-factor mutations and TP53 alterations, two

molecular groups strongly associated with myeloid neoplasia, received significantly higher model-assigned MDS probabilities, indicating that the learned morphologic representations may capture disease features associated with underlying molecular states. The modest association between overall mutation burden and predicted MDS probability further supports the same indication. Those findings are consistent with known clinical data, both CCUS and MDS are reported to have somatic mutations but not ICUS by definition and certain mutations are more likely to progress from CCUS to MDS clinically. Notably, prior work predicting genetic and cytogenetic aberrations in MDS patients from BMB-derived morphologic features also reported high performance in predicting spliceosome mutations (AUC=0.89) and TP53 (AUC=0.82) [2]. Together, these findings provide additional evidence that AI-based morphologic assessment can achieve stable diagnostic performance while capturing clinically relevant molecular heterogeneity.

Prior computational approaches in bone marrow pathology have largely focused on either cell-level analysis or single-modality whole-slide image (WSI) classification. Earlier studies demonstrated the feasibility of applying artificial intelligence to bone marrow biopsy or aspirate images for specific diagnostic tasks, including discrimination of MDS from normal marrow, differentiation of MDS from MDS/MPN overlap neoplasms, distinction of acute promyelocytic leukemia from acute myeloid leukemia, classification among MDS, AML, and donor bone marrow samples, and recognition of other hematologic conditions such as immune thrombocytopenic purpura based on marrow morphology [2, 19, 24, 10, 17, 5, 27, 15, 28, 26, 7, 33, 18, 30]. However, these approaches typically rely on a single source of morphologic information and do not reflect the multimodal nature of routine hematopathology workflows, where diagnosis is informed by the complementary evaluation of bone marrow aspirate cytology and biopsy architecture. Moreover, most prior studies have focused on diagnostically overt hematologic diseases or established hematologic entities with comparatively distinct morphologic phenotypes, rather than the diagnostically challenging distinction between MDS and precursor cytopenic conditions such as ICUS and CCUS, where abnormalities can be subtle, heterogeneous, and diagnostically ambiguous.

Beyond diagnostic performance, a central challenge for the clinical adoption of AI systems is understanding how they influence physician decision-making in real-world practice. In our pilot reader study, AI assistance demonstrated encouraging trends toward higher diagnostic accuracy, increased inter-reader agreement, and reduced variability in diagnostic confidence between experts. One potential explanation for the observed increase in diagnostic accuracy is that the AI overlay guided pathologists to examine a broader and more informative portion of the bone marrow slide than might be routinely assessed during conventional review. By highlighting diagnostically relevant regions, the system may have enabled more comprehensive slide evaluation within a comparable review time. Furthermore, the standardized visualization provided by the AI overlay may have encouraged readers to evaluate a more consistent set of morphologic regions, thereby contributing to the observed improvement in

inter-reader agreement. For example, the entire aspirate slide may not always be reviewed in its entirety. Blast distribution can be uneven. The AI overlay may help reduce selection bias in the areas reviewed, potentially leading to greater agreement. Importantly, readers frequently benefited from correct AI recommendations while retaining the ability to reject incorrect model predictions, suggesting that AI assistance complemented rather than replaced expert judgment. Collectively, these findings indicate that AI may serve not only as a diagnostic support tool but also as a mechanism for standardizing morphologic interpretation, addressing a major source of variability in MDS diagnosis across institutions and observers making it valuable for clinical practice and future clinical trials, and potentially extending to a broader range of hematopathologic disorders.

Notably, the proposed morphology-based AI system demonstrated both autonomous diagnostic capability and decision-support utility. In the reader-study cohort composed exclusively of morphology-only cases, the standalone model achieved performance comparable to experienced hematopathologists, while AI-assisted review further improved diagnostic accuracy beyond either the model or the unaided experts alone. These findings suggest a dual-use paradigm in which the system can function independently as an automated screening or triage tool, while also augmenting expert performance when incorporated into routine diagnostic workflows. Furthermore, the observation that experts frequently overrode incorrect AI predictions provides preliminary evidence against excessive automation bias and supports the feasibility of human–AI collaboration in bone marrow pathology.

Limitations include small sample sizes in rare subtypes (e.g., MDS-U, MDS-RSSLD, and isolated del(5q)), resulting in wide confidence intervals and uncertainty in performance estimates. In the Asp BoC branch, ROI selection, cell detection, and cell feature extraction and aggregation are performed as separate steps; integrating these components into an end-to-end pipeline could improve both efficiency and reproducibility. Moreover, the impact of slide quality was not explicitly quantified in this study, although low-cellularity aspirates and poorly preserved biopsy specimens may reduce model reliability and effect reader interpretability.

Future work will explore high-grade versus low- and intermediate-risk CCUS cases to understand modality-specific contributions. While the current model was designed to isolate the diagnostic contribution of bone marrow morphology, future studies will evaluate whether integrating peripheral blood, clinical, and genomic information further improves diagnostic performance and risk stratification. We also aim to investigate correlations between histopathologic features and molecular subtypes in CCUS and MDS. Finally, comparisons of model explanations with hematopathologist reports will assess generalizability as a diagnostic aid.

4 Methods

Dataset

A multi-institutional real-world dataset along with gold-standard diagnoses made by pathologists with expertise in MDS, who themselves had a rigorous process for diagnosing MDS and adjudicating discrepancies in diagnosis [8] from the National Heart, Lung, and Blood Institute (NHLBI) MDS Natural History Study (NHS) (NCT02775383) [22] was used as the discovery cohort for model training and internal validation. For each patient, a baseline whole-slide image (WSI) pair was included, consisting of a hematoxylin and eosin (H&E)-stained bone marrow biopsy (BMB) and a Wright-stained bone marrow aspirate (BMA), both scanned at 40× magnification with a base resolution of 0.252 μm per pixel. Only patients (n=645) with both images available were included in the discovery cohort. An independent institutional cohort of 95 patients served as the external validation set, containing corresponding baseline BMB and BMA slides scanned under similar conditions (40× magnification, 0.263 μm per pixel).

CCUS and ICUS were combined into a "non-MDS" category, based on the absence of definitive morphologic dysplasia ($< 10\%$ dysplasia in any cell lineage and $< 5\%$ blasts). The number of samples in each diagnostic category, as well as the subcategory distribution within the MDS cases, is summarized in Table 2.

Modality-specific representation learning

Tissue-level representation learning from biopsy

A WSI foundation model Prov-GigaPath [32] was utilized for the examination of H&E-stained BMB WSI (Figure 1c). A post hoc comparison with several other histopathology foundation models to evaluate the relative performance of Prov-GigaPath is provided in the Supplementary Table S1. Each WSI was tiled into 512×512 tiles to extract the tile-level visual features from the tile-encoder of the two-stage architecture of Prov-GigaPath. The tile features (D=1536) were aggregated to obtain slide-level features (D=768) using the LongNet based slide-encoder of the same foundation model.

The slide-encoder was fine-tuned on the discovery dataset using a 5-fold cross-validation. The training parameters were set to 20 epochs, with gradient accumulation over 32 steps. Optimization employed a base learning rate of 0.002 with a weight decay of 0.05, and layer-wise learning rate decay was set to 0.95. A dropout rate of 0.1 was applied to mitigate overfitting. Feature embeddings extracted from layer 11 were subsequently fed into a fully connected classification layer to generate final predictions.

Cell-level representation learning from aspirates

The BMA cell morphology examination pipeline (Figure 1a) consisted of a DenseNet-121 model for region-of-interest (ROI) tile selection and a YOLOv4

Table 2: Distribution of MDS and non-MDS diagnostic subcategories across the NHLBI discovery and institutional validation cohorts. Counts are shown separately for each cohort and in total. Group totals and the overall total are highlighted in bold. Dashes (–) indicate that no cases were available.

Category	Subcategory	# Discovery	# Validation
MDS	Single Lineage Dysplasia (SLD)	18	1
	Ring Sideroblasts Single Lineage Dysplasia (RSSLD)	25	–
	Multi Lineage Dysplasia (MLD)	84	–
	Ring Sideroblasts Multi Lineage Dysplasia (RSMLD)	49	–
	Excess Blasts 1 (EB1)	56	8
	Excess Blasts 2 (EB2)	50	26
	Isolated del(5q)	16	1
	Unclassifiable (MDS-U)	21	–
Total MDS cases		**319**	**36**
Non-MDS	Clonal Cytopenia of Undetermined Significance (CCUS)	276	52
	Idiopathic Cytopenia of Undetermined Significance (ICUS)	50	7
Total Non-MDS cases		**326**	**59**
Overall Total		**645**	**95**

detector for nucleated cell localization within the selected ROI tiles. To overcome the requirement for cell-level annotations across the wide spectrum of dysplastic morphologies, we employed a semi-supervised learning approach that aggregates information across a bag-of-cells comprising N (=500, consistent with international standards for bone marrow aspirate review [1]) detected nucleated cells detected incrementally within tiles ranked by their likelihood of containing diagnostically relevant ROIs for each BMA WSI. Vision features for the detected cells were subsequently extracted using ResNet50 and aggregated using a hybrid Vision Transformer (ViT) model [4].

Pre-trained weights from Tayebi et al. [27] were utilized for both ROI selection and cell detection. The pretrained YOLOv4 model detects 19 categories of cellular and non-cellular objects, namely "neutrophil", "metamyelocyte", "myelocyte", "promyelocyte", "blast", "erythroblast", "lymphocyte", "monocyte", "plasma cell", "eosinophil", "basophil", "megakaryocyte", "megakaryocyte nuclei", "platelets", "platelet clumps", "histiocytes", "debris", "mast cells", and "other cell". Objects detected as "debris", "mast cells", "histiocyte", "platelet", and "platelet clump" were excluded from the bag-of-cells in accordance with hematopathological guidelines.

Global context modeling from aspirates

To capture global contextual information from the entire BMA whole-slide image—such as the spatial distribution, relative organization, and scale of cellular and non-cellular components within the smear—we incorporated an additional global context modeling branch. This branch (Figure 1b) employed the same foundation model, Prov-GigaPath [32], as used for the BMB evaluation, fine-tuned under identical settings described in Section 4. By doing so, the proposed TriPath model leveraged both detailed cellular representations from the bag-of-cells branch and large-scale morphological organization from the whole-slide view, enabling a complementary understanding of aspirate smears.

TriPathModel: Multimodal integration and training strategy

We implemented a gated fusion module to integrate the three feature embeddings extracted from the modality-specific encoders described above—BMB WSI encoder, BMA BoC encoder, and BMA WSI encoder—into a shared representation (Figure 1d). Each modality embedding was projected into a common latent space ($d = 512$) using independent linear layers. The projected embeddings were concatenated and passed through a lightweight gating layer that produced modality-specific weights, normalized via a softmax function with temperature scaling (factor of 2.0), allowing the model to regulate the relative contribution of each modality in a data-driven manner. The final fused representation was obtained as a weighted sum of modality-specific embeddings, followed by layer normalization and dropout (rate of 0.3) for regularization.

All encoders were initialized from domain-specific pretraining and kept frozen except for their final two layers, which were fine-tuned to adapt to the downstream classification task. To preserve modality-specific discriminative capacity, auxiliary classification heads were attached to each encoder and optimized jointly with the fusion module in a multi-branch supervision scheme. The fused representation from the gated fusion layer served as the primary prediction head, while auxiliary predictions from each encoder provided additional gradients for stabilization.

Optimization was performed using stochastic gradient descent (momentum 0.9, weight decay 0.05). Distinct learning rates were applied across parameter groups—10^{-3} for the fusion module and classification heads, and 10^{-4} for the unfrozen encoder layers—to enable gradual fine-tuning while preserving pretrained representations. Auxiliary losses were weighted by $\lambda_{\text{aux}} = 0.2$ (empirically tuned in the range 0.1–0.3). Training was conducted for 20 epochs with gradient accumulation over 32 steps, a layer-wise learning rate decay of 0.95, and a batch-level dropout rate of 0.3 for regularization. Further details of the training objective and loss formulation are provided in the Supplementary Material.

Evaluation Protocol

Experimental Setup

The discovery cohort (n = 645) was randomly split into training and held-out test sets using a stratified 80:20 partition to preserve the distribution of diagnostic classes. All models were trained on the training set, with hyperparameters selected via five-fold cross-validation within the training data. Model performance was evaluated on a held-out test set (n = 129) and an independent external validation set (n = 95), both comprising paired bone marrow aspirate (BMA) and biopsy (BMB) whole-slide images (WSIs).

Performance was quantified using standard classification metrics, including area under the receiver operating characteristic curve (AUROC), accuracy, sensitivity, and specificity. Where appropriate, 95% confidence intervals were estimated using bootstrapping for AUROC and the Clopper–Pearson exact interval for accuracy.

Subcategory analyses were performed on both the held-out and external validation sets to assess robustness across diagnostic subgroups. Fusion behavior was examined via learned gating weights to quantify modality contributions across cases. The proposed TriPath model was additionally compared against unimodal and dual-modal baselines to quantify the incremental benefit of multimodal integration.

Reader Study Design

To evaluate the clinical utility of the proposed model, we conducted a pilot, blinded, paired reader study involving two board-certified hematopathologists.

Each expert independently reviewed bone marrow aspirate and biopsy WSIs of the same set of cases under two conditions: (i) Arm A (unassisted), where only the WSIs were provided, and (ii) Arm B (AI-assisted), where WSIs were accompanied by Integrated Gradients–based visual explanation overlays highlighting regions contributing to the predicted diagnosis (Figure 8). Details of the Integrated Gradients computation are provided in the Supplementary Material.

A washout period of at least 3 months was implemented to minimize recall bias. To facilitate the evaluation of AI-assisted decision making, the reader-study cohort was intentionally enriched for both model-correct and model-incorrect cases from the external validation cohort. Thirty cases (15 MDS and 15 non-MDS) were initially selected while maintaining representation across diagnostic subcategories. Following expert review, four cases were excluded because of poor slide quality or inadequate tissue sampling, resulting in a final cohort of 26 paired cases included in the reader study. For each case, experts provided a binary diagnostic classification (MDS vs non-MDS) and a 3-point ordinal diagnostic confidence score across aspirate, biopsy, and combined modalities. Expert's agreement (3-point ordinal) with the model explanation was also recorded under Arm B.

Reader performance was evaluated in terms of diagnostic accuracy. Differences in diagnostic accuracy between arms were assessed using McNemar's test. Inter-expert agreement was quantified using Cohen's κ statistic. Changes in diagnostic confidence between arms were analyzed using the Wilcoxon signed-rank test to assess confidence calibration. Additionally, we quantified error correction dynamics by identifying cases in which incorrect diagnoses in Arm A were corrected in Arm B. Agreement between expert assessments and model explanations was also analyzed to evaluate alignment between human and AI decision-making.

Data availability

The MDS NHS cohort data can be requested through the NHLBI MDS STUDY Data Coordinating Center (mdscontact@emmes.com). Data from the external validation cohort will be made available upon reasonable request to the authors.

Code availability

The source code used for this study along with the weights of the AI model will be made publicly available through a GitHub repository upon acceptance of this manuscript.

Acknowledgements

The Study is funded by grant EPEF-N2023408-202337 from the Edward P. Evans Foundation. The National MDS Natural History study has been

supported by US Federal Government Contracts HHSN268201400003I and HHSN268201400002I from the National Heart Lung and Blood Institute and additional funding by the National Cancer Institute to clinical centers. This manuscript has been reviewed by the NHLBI MDS STUDY for scientific content and consistency of data interpretation with previous NHLBI MDS STUDY publications.

Author Contributions

P.D., I.E.N., R.K., E.P., L.Z., and M.H. contributed to the conception and design of the work. L.Z., S.A., M.E.W., N.G., L.C.M., M.A.S., A.E.D., R.K., and I.E.N. contributed to the data acquisition and curation. P.D., J.R., and J.P. contributed to the technical implementation. R.K., E.P., L.Z., M.H., N.G., M.A.S., and A.E.D. provided clinical inputs to the study. P.D. and I.E.N. wrote the paper. R.K. and I.E.N. supervised the research. All authors reviewed and approved the manuscript.

References

[1] Ahmed A Abdulrahman et al. "Is a 500-cell count necessary for bone marrow differentials? A proposed analytical method for validating a lower cutoff". In: *American journal of clinical pathology* 150.1 (2018), pp. 84–91.

[2] Oscar E Brück et al. "Machine learning of bone marrow histopathology identifies genetic and clinical determinants in patients with MDS". In: *Blood cancer discovery* 2.3 (2021), pp. 238–249.

[3] Jian Chen et al. "Texture-Unet: A Texture-Aware Network for Bone Marrow Smear Whole-Slide Image Region of Interest Segmentation". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2024, pp. 2006–2010.

[4] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[5] Jan-Niklas Eckardt et al. "Image-based explainable artificial intelligence accurately identifies myelodysplastic neoplasms beyond conventional signs of dysplasia". In: *NPJ Precision Oncology* (2025).

[6] Leander van Eekelen, Geert Litjens, and Konnie M Hebeda. "Artificial intelligence in bone marrow histological diagnostics: potential applications and challenges". In: *Pathobiology* 91.1 (2024), pp. 8–17.

[7] Xinyan Fu et al. "Morphogo: an automatic bone marrow cell classification system on digital images analyzed by artificial intelligence". In: *Acta Cytologica* 64.6 (2020), pp. 588–596.

[8] Edward J Gorak et al. "Discordant pathologic diagnoses of myelodysplastic neoplasms and their implications for registries and therapies". In: *Blood advances* 7.20 (2023), pp. 6120–6129.

[9] Maximilian Ilse, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning". In: *International conference on machine learning*. PMLR. 2018, pp. 2127–2136.

[10] Sosuke Ishijima et al. "Extraction of Novel Features and Diagnosis Prediction in Myelodysplastic Neoplasm Using a Weakly Supervised Artificial Intelligence Model Based on Normal Megakaryocytes". In: *Pathology International* 75.11 (2025), pp. 562–571.

[11] Shizuo Kaji et al. "Deep learning-based morphological assessment of myelodysplastic syndrome on bone marrow smears". In: *Leukemia Research* (2025), p. 107923.

[12] Narine Kokhlikyan et al. "Captum: A unified and generic model interpretability library for pytorch". In: *arXiv preprint arXiv:2009.07896* (2020).

[13] Nuri Lee et al. "Deep learning application of the discrimination of bone marrow aspiration cells in patients with myelodysplastic syndromes". In: *Scientific Reports* 12.1 (2022), p. 18677.

[14] Joshua E Lewis and Olga Pozdnyakova. "Advances in Bone Marrow Evaluation." In: *Clinics in Laboratory Medicine* 44.3 (2024), pp. 431–440.

[15] Joshua E Lewis et al. "An automated pipeline for differential cell counts on whole-slide bone marrow aspirate smears". In: *Modern Pathology* 36.2 (2023), p. 100003.

[16] Bin Li, Yin Li, and Kevin W Eliceiri. "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 14318–14328.

[17] Petru Manescu et al. "Detection of acute promyelocytic leukemia in peripheral blood and bone marrow with annotation-free deep learning". In: *Scientific Reports* 13.1 (2023), p. 2562.

[18] Christian Matek et al. "Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks". In: *Nature Machine Intelligence* 1.11 (2019), pp. 538–544.

[19] Vahid Mehrtash et al. "Self-Supervised Learning Can Distinguish Myelodysplastic Neoplasms from Clinical Mimics Using Bone Marrow Biopsies". In: *medRxiv* (2025), pp. 2025–02.

[20] Jinichi Mori et al. "Assessment of dysplasia in bone marrow smear with convolutional neural network". In: *Scientific reports* 10.1 (2020), p. 14734.

[21] Charlie Saillard et al. *H-optimus-0*. 2024. URL: `https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0`.

[22] Mikkael A Sekeres et al. "The National MDS Natural History Study: design of an integrated data and sample biorepository to promote research studies in myelodysplastic syndromes". In: *Leukemia & lymphoma* 60.13 (2019), pp. 3161–3171.

[23] Zhuchen Shao et al. "Transmil: Transformer based correlated multiple instance learning for whole slide image classification". In: *Advances in neural information processing systems* 34 (2021), pp. 2136–2147.

[24] Andrew Srisuwananukorn et al. "Artificial intelligence differentiates prefibrotic primary myelofibrosis with thrombocytosis from essential thrombocythemia using digitized bone marrow biopsy images". In: *Leukemia* (2026), pp. 1–9.

[25] Jie Su et al. "Roi-bmc-dnnet: An efficient automatic analysis model of whole-slide scanned bone marrow aspirate images for the diagnosis of hematological disorders". In: *Biomedical Signal Processing and Control* 86 (2023), p. 105243.

[26] Jonathan Tarquino et al. "A One-Class Variational Autoencoder (OC-VAE) Cascade for Classifying Atypical Bone Marrow Cell Sub-types". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 725–734.

[27] Rohollah Moosavi Tayebi et al. “Automated bone marrow cytology using deep learning to generate a histogram of cell types”. In: *Communications medicine* 2.1 (2022), p. 45.

[28] Satvik Tripathi et al. “HematoNet: Expert level classification of bone marrow cytology morphology in hematological malignancy with deep learning”. In: *Artificial Intelligence in the Life Sciences* 2 (2022), p. 100043.

[29] Ching-Wei Wang et al. “Deep learning for bone marrow cell detection and classification on whole-slide images”. In: *Medical Image Analysis* 75 (2022), p. 102270.

[30] Weining Wang et al. “Artificial intelligence-assisted diagnosis of hematologic diseases based on bone marrow smears using deep neural networks”. In: *Computer Methods and Programs in Biomedicine* 231 (2023), p. 107343.

[31] Jinxi Xiang et al. “A vision–language foundation model for precision oncology”. In: *Nature* 638.8051 (2025), pp. 769–778.

[32] Hanwen Xu et al. “A whole-slide foundation model for digital pathology from real-world data”. In: *Nature* (2024).

[33] Ta-Chuan Yu et al. “Automatic bone marrow cell identification and classification by deep neural network”. In: *Blood* 134 (2019), p. 2084.

[34] Ling Zhang et al. “Diagnosis of myelodysplastic syndromes and related conditions: rates of discordance between local and central review in the NHLBI MDS Natural History Study”. In: *Blood* 132 (2018), p. 4370.

[35] Eric Zimmermann et al. “Virchow2: Scaling Self-Supervised Mixed Magnification Models in Pathology”. In: *arXiv preprint arXiv:2408.00738* (2024).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement.pdf