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.

Interpretable Multiple Instance Learning for Hematologic Diagnosis from Peripheral Blood Smears

Siddharth Singi

sss2312@columbia.edu

Memorial Sloan Kettering Cancer Center https://orcid.org/0009-0005-6184-5239

Shenghuan Sun

University of California San Francisco https://orcid.org/0000-0002-4339-2716

Zhanghan Yin

Memorial Sloan Kettering Cancer Center

Riya Gupta

Memorial Sloan Kettering Cancer Center

Dylan Webb

Memorial Sloan Kettering Cancer Center

Khawaja Bilal

Memorial Sloan Kettering Cancer Center

Deepika Dilip

Memorial Sloan Kettering Cancer Center

Linlin Wang

University of California, San Francisco

Neeraj Kumar

University of Alberta, Edmonton, Alberta, Canada.

Swaraj Nanda

Memorial Sloan Kettering Cancer Center

Nicholas Sanchez

University of California Berkeley

Jacob Cleaves

Memorial Sloan Kettering Cancer Center

Brenda Fried

Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0001-5972-6146

Sean Paulsen

Memorial Sloan Kettering Cancer Center

Ethan Yan

Memorial Sloan Kettering Cancer Center https://orcid.org/0009-0005-6332-6803
Ali Kamali
Memorial Sloan Kettering Cancer Center
Argho Sarkar
Memorial Sloan Kettering Cancer Center
Allyne Manzo
Memorial Sloan Kettering Cancer Center https://orcid.org/0009-0004-3056-1180
Jeeyeon Baik
Memorial Sloan Kettering Cancer Center
Irem Isgor
Memorial Sloan Kettering Cancer Center
Cesar Colorado-Jimenez
Memorial Sloan Kettering Cancer Center
Anthony Cardillo
Memorial Sloan Kettering Cancer Center
Leonardo Boiocchi
Memorial Sloan Kettering Cancer Center
Aijazuddin Syed
Memorial Sloan Kettering Cancer Center
David Kim
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0002-7593-6052
Brie Kezlarian-Sachs
Memorial Sloan Kettering Cancer Center
Maly Fenelus
Memorial Sloan Kettering Cancer Center
Alexander Chan
Department of Pathology https://orcid.org/0000-0001-5884-8424
Mariko Yabe
Memorial Sloan Kettering Cancer Center
Samuel McCash
Memorial Sloan Kettering Cancer Center
Menglei Zhu
Memorial Sloan Kettering Cancer Center
Simon Mantha
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0003-4277-5261
Orly Ardon
Memorial Sloan-Kettering Cancer Center https://orcid.org/0000-0001-8147-933X
Lauren McVoy
Memorial Sloan Kettering Cancer Center

Wenbin Xiao
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0001-8586-8500
Mikhail Roshal
Memorial Sloan Kettering Cancer Center
Oscar Lin
Memorial Sloan Kettering Cancer Center
Ahmet Dogan
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0001-6576-5256
Iain Carmichael
University of North Carolina at Chapel Hill
Chad Vanderbilt
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0002-8114-0237
Gregory Goldgof
Memorial Sloan Kettering Cancer Center https://orcid.org/0000-0001-8732-9834

Article

Keywords:

Posted Date: October 31st, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6933141/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.

Interpretable Multiple Instance Learning for Hematologic Diagnosis from Peripheral Blood Smears

Siddharth Singi[1], Shenghuan Sun[2], Zhanghan Yin[1,3], Riya Gupta[1], Dylan C. Webb[1,3], Khawaja H. Bilal[1], Deepika Dilip[4], Linlin Wang[5], Neeraj Kumar[1], Swaraj Nanda[1], Nicolas Sanchez[1,3], Jacob G. Van Cleave[1], Brenda Fried[1], Sean Paulsen[1], Ethan S. Yan[1], Ali Kamali[1], Argho Sarkar[1], Allyne Manzo[1], Jeeyeon Baik[1], Irem S Isgor[1], Cesar Colorado-Jimenez[1], Anthony Cardillo[1], Leonardo Boiocchi[1], Aijazuddin Syed[1], David Kim[1], Brie Kezlarian-Sachs[1], Maly Fenelus[1], Alexander Chan[1], Mariko Yabe[1], Samuel I McCash[1], Menglei Zhu[1], Simon Mantha[6], Orly Ardon[1], Lauren McVoy[1], Wenbin Xiao[1], Mikhail Roshal[1], Oscar Lin[1], Ahmet Dogan[1], Iain Carmichael[7], Chad Vanderbilt[1*], and Gregory M. Goldgof[1,8,9*†]

[1]Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[2]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA

[3]Department of Statistics, University of California Berkeley, USA

[1,4]New York Medical College, Valhalla, NY

[5]Department of Laboratory Medicine, University of California, San Francisco, CA, USA

[6]Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[7]School of Data Science and Society; Department of Pathology and Laboratory Medicine, University of North Carolina-Chapel Hill

[8]Department of Pathology and Laboratory Medicine, Weill Cornell School of Medicine

[9]Halvorsen Center for Computational Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[*]These authors contributed equally to this work

[†]goldgofg@mskcc.org, vanderbc@mskcc.org

ABSTRACT

Accurate diagnosis of hematologic malignancies from peripheral blood smears (PBSs) requires integrating cellular morphology and composition across hundreds of white blood cells. Existing approaches primarily automate single-cell classification and do not provide whole-slide diagnostic predictions. We present a full network that utilizes a highly performative cell-based encoder (DeepHeme) for feature extraction paired with our weakly supervised framework using attention-based multiple instance learning (MIL) that we call CAREMIL (Cell AggRegation, Explainable, Multiple Instance Learning). Upon evaluating various popular image encoders and MIL architectures, the combination of DeepHeme and CAREMIL is the best performing pipeline on our disease classification task. CAREMIL proves to be a robust aggregation function that outperforms the most commonly used slide level aggregation function (gated multiple instance learning) across several encoder types. The greatest improvements in performance gain with CAREMIL is observed when using out-of-domain encoders, including an encoder trained on ImageNet and leading open-source pathology foundational models (UNI2 and Virchow2). CAREMIL plus DeepHeme achieves the highest diagnostic performance across acute leukemia (AML), myelodysplastic syndromes (MDS), and hairy cell leukemia (HCL) (AUROCs 0.999, 0.891, and 0.945, respectively), and identifies AML disease even in cases with minimal or absent circulating blasts. Attention values assigned by CAREMIL highlight diagnostically relevant cells and reveal disease-specific morphometric signatures, enabling biological interpretability and case-level insight. CAREMIL remains robust to misclassified cell types by the cell image encoder and does not require explicit cell-level supervision. These findings position CAREMIL as an effective and interpretable multiple instance learning framework for hematologic slide diagnosis, with potential to extend to bone marrow aspirates, cytology, and other liquid biopsy specimens, and to support a broader shift toward quantitative, morphology-informed diagnostics in hematology.

1 Introduction

Accurate diagnosis of hematologic malignancies from peripheral blood smears (PBSs) requires integrating both the morphology of individual white blood cells and the overall cellular composition (1; 2; 3; 4). Peripheral smears serve as a non-invasive, rapid "liquid biopsy", offering critical information about underlying bone marrow pathology. However, diagnosis often hinges on detecting subtle morphological abnormalities or shifts in cellular proportions, making interpretation highly dependent on expert review(5). These challenges are particularly acute in cases of low circulating disease burden or early-stage dysplasia, where abnormalities may escape detection through routine screening (6; 7). The clinical and computational workflow for PBS analysis is outlined in Figure 1A. Recent

computational efforts have focused primarily on automating aspects of PBS interpretation, including single-cell classification and cytometric quantification (8), (9), (10). While these approaches improve efficiency, they fail to capture the integrative reasoning required for whole-slide diagnosis. In contrast to solid tumor pathology, where tiling approaches can leverage preserved tissue architecture (11), (12), hematologic specimens lack cohesive spatial organization(13), rendering traditional whole slide image (WSI) analysis methods poorly suited for blood smears, where proportions of cell types and single-cell morphology are central to diagnostic interpretation, rather than tissue architecture(14; 15). As a result, new frameworks are needed to model PBSs at the single-cell level while enabling slide-level diagnostic predictions. Multiple instance learning (MIL) offers a promising solution for weakly supervised learning in pathology (16; 17; 18; 19; 20; 21), treating individual cells as instances and the PBS as a labeled bag. Attention-based MIL architectures have shown success in solid tumor WSIs by improving both performance and interpretability, but their adaptation to hematologic specimens remains limited. Prior work has applied MIL to blood smear images using general-purpose encoders without attention-based aggregation, but comprehensive strategies for integrating morphology and composition with domain-specific features have not been explored (22).

Here, we present CAREMIL (Cell AggRegation, Explainable, Multiple Instance Learning), an attention-based multiple instance learning framework optimized for hematologic liquid biopsy interpretation. CAREMIL integrates both cell morphology and cytometric composition at the instance level to generate accurate, interpretable slide-level predictions. We evaluate CAREMIL across three diagnostically distinct hematologic malignancies: acute leukemia, characterized by increased circulating blasts; hairy cell leukemia, marked by morphologically abnormal lymphocytes; and myelodysplastic syndromes, distinguished by subtle dysplastic (morphologic) changes across multiple hematologic lineages. To support cell-level representation learning, we leverage DeepHeme (14), a domain-specific encoder trained on hematologic cell classification, applying its feature embeddings for the first time to weakly supervised, slide-level diagnosis.

CAREMIL consistently outperforms gated MIL architectures across multiple encoder types, with the greatest performance gains observed when using in-domain foundational models trained on non-hematologic images. The model accurately identifies diagnostically relevant cell populations without requiring explicit cell-level supervision and remains robust even when upstream classifiers mislabel abnormal morphologic variants. These findings demonstrate the utility of attention-based weakly supervised learning for morphology-informed, quantitative diagnostics in hematology and highlight the potential for broader application across liquid biopsy specimens.

2 Results

2.1 Dataset Curation

We curated a dataset of 1,017 whole slide images (WSIs) of peripheral blood smears (PBSs) scanned at 400x magnification (40x objective), representing a diverse cohort of 573 patients evaluated by the hematopathology service. Each PBS was paired with a bone marrow biopsy and clinical pathology report, with diagnoses established by board-certified hematopathologists based on the full clinical context in the bone marrow report. The dataset includes 307 cases of acute leukemia (AML), 95 cases of myelodysplastic syndromes (MDS), 41 cases of hairy cell leukemia (HCL), and 130 normal (NL) controls. Notably, control cases represent patients with normal overall reports, rather than healthy individuals, reflecting the real-world diagnostic setting. Each WSI contains hundreds of white blood cells, spanning a wide range of disease burden and morphologic representations, including blast-negative AL and subtle dysplasia in MDS. To ensure independence across datasets, train, validation, and test splits were performed at the patient level.

2.2 Improving the Attention Mechanism for Slide-Level Predictions

Multiple instance learning (MIL) provides a weakly supervised framework for training models when only bag-level (i.e. slide-level) labels are available, a setting well-suited to pathology applications where instance-level annotations (e.g., per-cell labels) are often impractical. In MIL, each bag consists of a collection of instances, and the model must learn to aggregate information across instances to make a bag-level prediction. We compare our CAREMIL model with the Gated attention MIL.

Gated attention MIL is common approach used in histology WSIs where attention is is applied by using a gating mechanism (17) [Ilse et al, 2018]. Instance features undergo two separate nonlinear transformations, one with a tanh activation and one with a σ activation. The outputs of these transformations are combined via element-wise multiplication before attention scoring. This gating mechanism allows the model to better modulate instance contributions by enhancing or suppressing different aspects of the feature representation prior to computing attention. However, aggregation remains a weighted sum across instances, and the model primarily focuses on selecting a few highly relevant instances for diagnosis.

CAREMIL builds upon the MIL paradigm by explicitly modeling both the individual morphology of instances and broader distributional properties of the instance population. Instead of aggregating raw cell embeddings through a weighted sum, CAREMIL first computes global statistical summaries (e.g. mean and variance) across the set of instance embeddings. These summary features are combined with the original cell embeddings and jointly processed

through a transformer-style self-attention mechanism. This architecture enables CAREMIL to reason both about rare high-attention instances and about subtle global shifts in cell population characteristics, improving robustness and clinical interpretability.

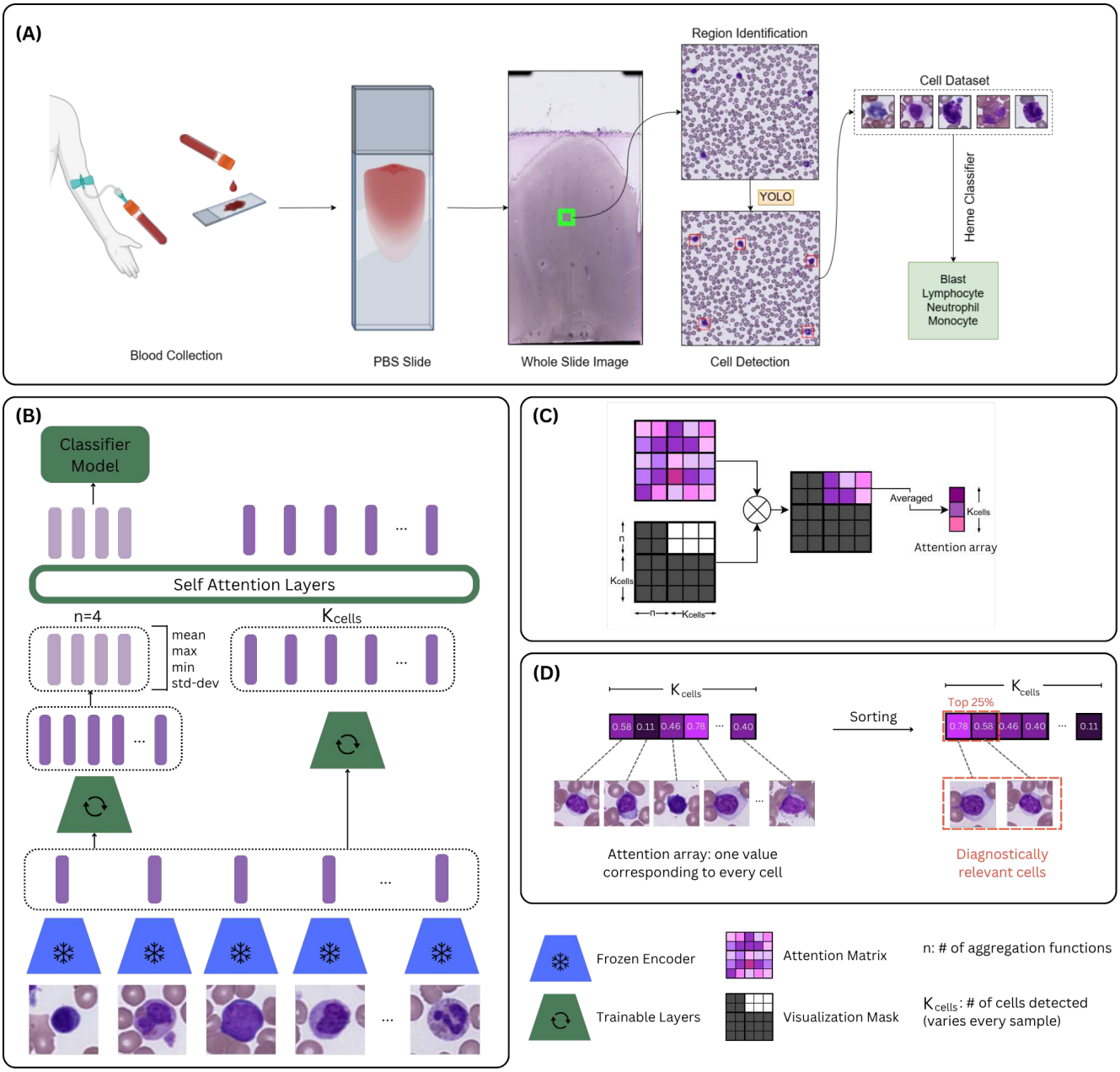

Figure 1. Overview of CAREMIL. CAREMIL is a deep learning pipeline designed for pathologists to diagnose and analyze liquid biosamples. **A)** Data collection process. **B)** Model architecture: The model is designed to take in a varying number of cells for every sample and make an explainable classification prediction. **C)** Masked Attention Visualization. We isolate the part of the attention matrix that relates the aggregation functions to every cell in the input of the model. **D)** Cell Ranking: Cell ranking by prediction importance identifies the cells with highest and lowest diagnostic relevance to verify a decision.

2.3 Model evaluation and performance

CAREMIL's performance was compared to two different approaches. The first is a set of optimized machine learning classifiers operating on the cell count alone (cytometry). To set a strong baseline, we performed hyperparameter optimization using a grid search over 6 different standard models and 227 different hyperparameter groups. Performance was compared with gated MIL attention (17). We also tested the performance of both the CAREMIL and the Gated MIL architecture on four different image encoders (see Figure 2. The first is DeepHeme, a domain-specific RexNeXt-50-based classifier, trained to identify 23 different cell hematopoietic cell classes found in the normal bone marrow. To create strong baselines for our work we also make comparisons on UNI2-h (23) and Virchow2, self-supervised computational pathology foundation models (24). Lastly, we also use a ResNeXt-50 model trained on ImageNet, to understand the performance of general purpose computer vision models not trained on histopathology. These models and their training hyperparameters are summarized in Supplementary Table 2.

Each approach was evaluated in four distinct clinical contexts, as summarized in Supplementary Table 3. The first task involved identifying acute leukemia (AML) from normal (NL) samples. AL encompasses both acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) and is defined by an increased percentage of blast cells in the peripheral blood or bone marrow.

The second task focused on distinguishing myelodysplastic syndromes (MDS) from normal samples. MDS is characterized by abnormal morphology (dysplasia) in hematopoietic cells. While MDS is typically diagnosed in patients presenting with cytopenias, the ability to detect myelodysplasia morphologically in patients with cytometrically normal blood counts may provide a low-cost method for identifying individuals at risk of developing MDS, or who already have cytogenetically defined MDS. Earlier identification of these patients could enable interventions aimed at preventing disease progression.

The third task aimed to separate hairy cell leukemia (HCL) from normal samples. HCL is defined by the presence of morphologically abnormal lymphocytes with "hairy" cytoplasmic projections, although these projections were rarely observed in this dataset due to slide preparation artifacts. HCL does not necessarily cause abnormal cytometry.

The fourth task involves distinguishing among AL, MDS, HCL, and NL, presenting a significantly more challenging quaternary classification problem compared to binary classification. The increased complexity stems from the need to separate four distinct classes with overlapping morphologic and cytometric features. For instance, AL and MDS may share dysplastic characteristics and overlapping blast counts, while all four classes can exhibit normal cytometric profiles under certain conditions. This task is further complicated by the requirement for the classifier to generalize across all classes while minimizing errors specific to each, highlighting the difficulty of

achieving robust and accurate multi-class classification.

Slides were labeled with diagnoses derived from pathology reports at the case level. Model performance was evaluated based on its ability to correctly predict these case-level diagnoses. For example, a diagnosis of acute leukemia (AML) did not require detectable disease in the peripheral blood, as the diagnosis may have been based on findings from the bone marrow. This approach ensured the inclusion of cases with low-level circulating disease that might be identified by the AI system, even when undetectable by automated hematology analyzers or human pathologists.

CAREMIL outperforms Gated MIL on all tasks (see Supplementary Table 5) and performs better irrespective of the image encoder being used (see Figure 2A). Our results also show that the hematology-specific image encoder, DeepHeme, demonstrated substantially better performance compared to large general purpose pathology foundation models like UNI2-h or Virchow2. Both the encoder and the aggregator model architecture play an important role in the overall performance. The CAREMIL model architecture with the DeepHeme image encoder demonstrates superior performance compared to other models and encoders for the hematology domain.

Supplementary Table 4 and 5 summarizes the average AUROC-scores for each model and experiment. In the MDS vs NL and multiclass task, UNI2-h achieved slightly higher performance, though the difference was not significant. Across all four experiments, models using the CAREMIL architecture performed significantly better than their Gated MIL equivalent. Additionally, cytometry-based (e.g. based on cell counts alone) machine learning classifiers consistently outperformed MIL models that relied on ImageNet encoders.

2.4 Explainability and patient safety

AI models deployed in medical environments must include a layer of explainability to ensure patient safety and clinical utility. A key aspect of explainability is enabling physicians to rapidly determine whether the model's performance is based on sound biology and not spurious correlations. Additionally, adding explainability outputs can facilitate broader analyses on how abnormal cells influence diagnostic decisions. For this reason, we developed a workflow designed to provide pathologists with interpretable outputs for each sample. The output consists of a figure showcasing three key cell subsets (see Supplementary Figure 7) : the cells with the highest attention values (top row), a random selection of cells to represent the whole-slide context (middle row), and the cells with the lowest attention values (bottom row). This structure allows pathologists to confirm whether the model is focusing on diagnostically relevant cells while avoiding irrelevant or morphologically normal cells.

The top row specifically highlights the morphology of disease cells, offering a clear visual confirmation of abnormal cells. In cases of very low disease burden, where abnormal cells may not appear in the random selection,

the top row ensures the clinician can still validate the diagnosis. The middle row provides a representative sampling of randomly selected cells, enabling clinicians to quickly estimate the disease burden in the patient by assessing the proportion of abnormal cells. The bottom row allows clinicians to verify that the model correctly identifies and de-emphasizes cell types that are least diagnostically relevant, ensuring it focuses on cells that contribute meaningfully to the predicted diagnosis. While the figures presented in this study are limited to 10 cells per row for visual clarity, a clinical deployment would include a larger number of cells. This would make it possible to accurately estimate disease burden in cases with a low percentage of abnormal cells.

In the context of model explainability, specific patterns of cell morphology were observed for each disease state, corresponding to the high attention cells. Figure 2C is composed of disease specific examples of high attention-based outputs that demonstrate the model's capacity to align its predictions with clinically relevant morphological features while providing a broader context for disease burden and cell-type distribution.

In Figure 2C the top panel shows that in a sample with acute myeloid leukemia (AML), the high-attention cells consistently included leukemic blasts characterized by fine chromatin and high nuclear-to-cytoplasmic ratios. The next panel shows a case of MDS, with high-attention cells including examples of both dysplastic neutrophils and immature granulocytes which, when circulating in peripheral blood, are characteristic of MDS. The bottom panel shows an HCL sample, where the high-attention cells highlight the characteristic abnormal "hairy cell" lymphocytes associated with the disease.

Supplementary Figure 7 highlights an additional case of MDS. High-attention cells in this case include dysplastic hypogranular neutrophils, monolobated neutrophils, dysplastic erythroblasts exhibiting nuclear-cytoplasmic dyssynchrony, and abnormal immature granulocytes circulating in the peripheral blood. The inclusion of two examples for MDS demonstrates the model's ability to identify different types of dysplasia across specimens, reflecting the diversity of morphological abnormalities seen in MDS, which represents a family of syndromes, rather than one specific entity. These findings underscore the model's adaptability in detecting patient-specific manifestations of dysplasia, providing critical information for diagnosis and model explainability.

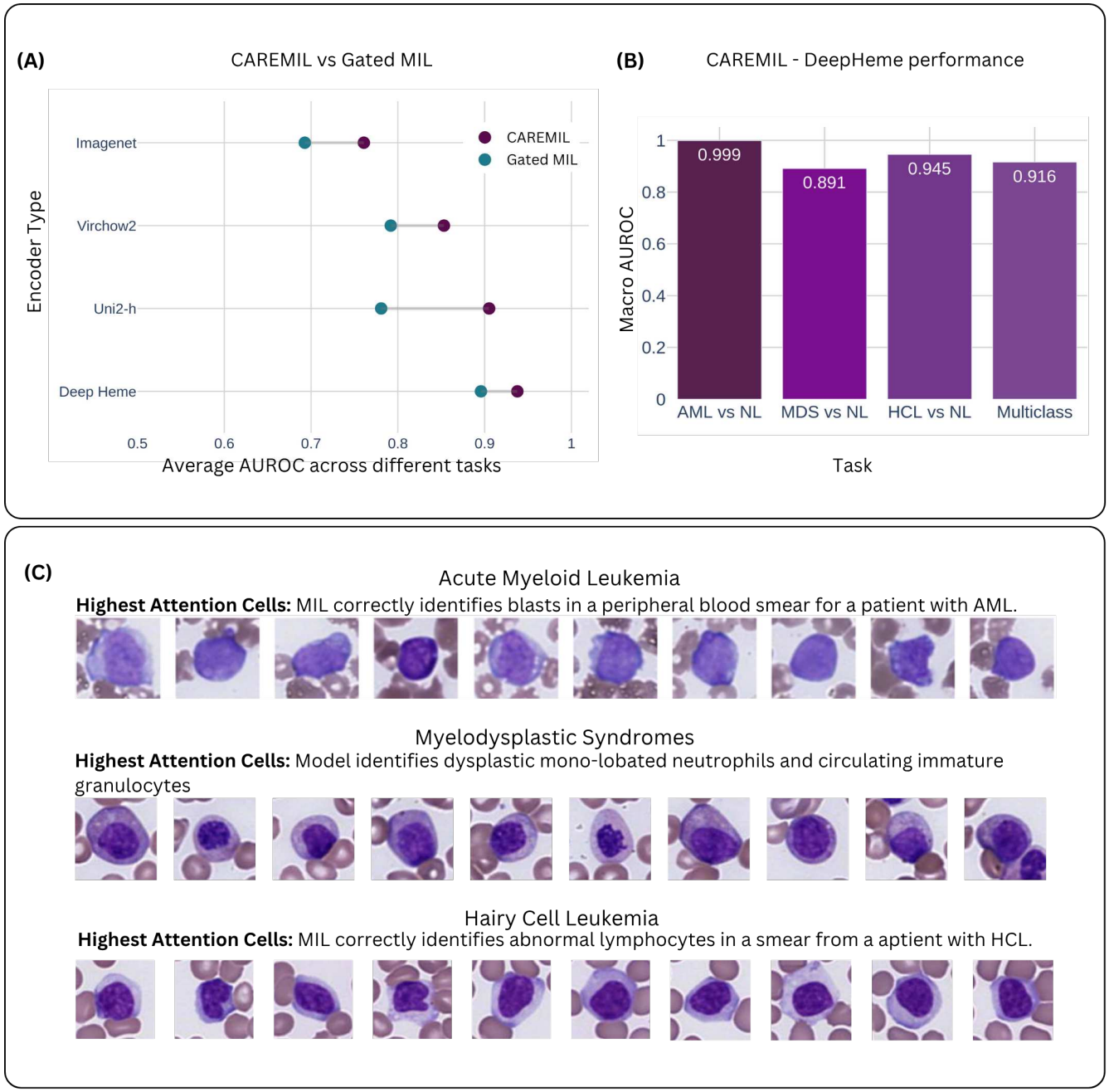

Figure 2. (A): Comparison of CAREMIL and Gated MIL performance across different encoder types. The plot shows the average AUROC across multiple diagnostic tasks using four different pretrained encoder types: ImageNet, Virchow2, Uni2-h, and DeepHeme. Each horizontal pair of dots represents the performance of CAREMIL (purple) and Gated MIL (teal) using the same encoder. CAREMIL consistently outperforms or matches Gated MIL across all encoder types, with the largest performance margin observed for the Virchow2 and Uni2-h encoders. Horizontal lines connect paired values for visual comparison. **(B): CAREMIL + DeepHeme performance:** The best performing combination of architecture and encoder is evaluated on 3 binary classifications and 1 multiclass classification. **(C) Explainability Analysis:** The top 10 highest attention cells generated by CAREMIL + DeepHeme are displayed from selected examples from AML, MDS, and HCL, respectively.

2.5 Limits of detection in synthetic patients

To evaluate the sensitivity of CAREMIL across varying blast percentages, we generated a dataset of synthetic patients by combining fixed proportions of blasts from leukemia patients with randomly sampled non-blast cells from normal individuals. Leukemia is typically diagnosed when blast percentages exceed 20%. However, lower blast levels may be observed in post-treatment patients, cases where the disease is predominantly confined to the marrow, or for certain disease genotypes. As illustrated in Figure 3A, the probability of correctly diagnosing acute leukemia increases with rising blast percentages. Using a cutoff of 0.5, at 0% blast proportion, the algorithm identifies all but one sample as normal and at 20% the algorithm identifies all samples as acute leukemia.

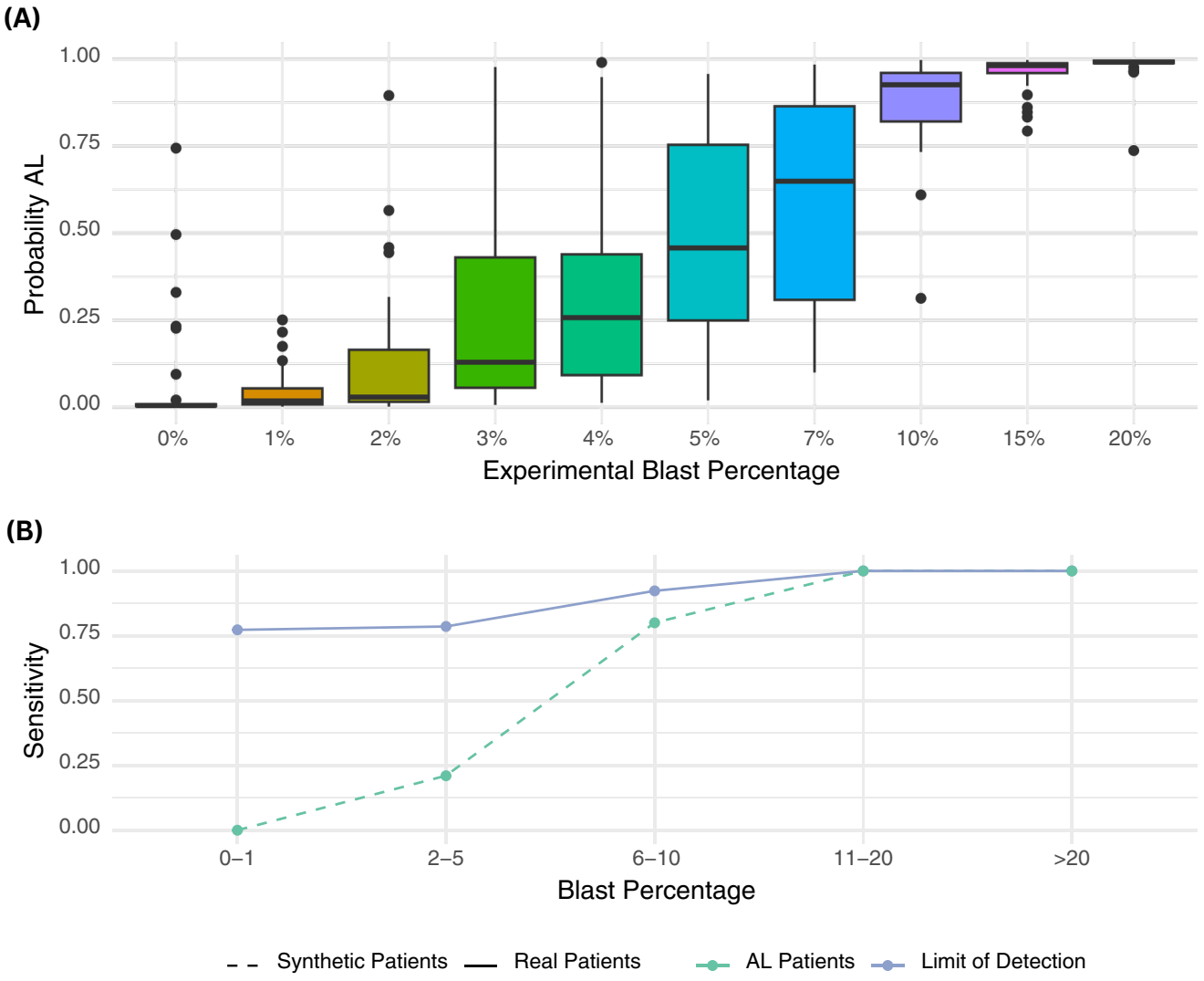

Figure 3. CAREMIL Ablation Studies. (A) Limit of detection study. This figure illustrates the CAREMIL scores across a range of blast percentages in synthetic patients. Synthetic patients were generated by combining a specific percentage of randomly selected blast cells from acute leukemia patient samples into random cells taken from normal patient samples. **(B) CAREMIL performance in minimal leukemia detection.** This graph compares the sensitivity of CAREMIL for diagnosing leukemia in real patients (solid line) versus synthetic patients (dashed line) across various blast percentage categories. Sensitivity increases with higher blast percentages in both patient groups. However, CAREMIL performs better on real patients, suggesting the model is incorporating data from non-blast cells into its prediction. Surprisingly, it achieves a sensitivity of 75% for samples with blast count <=1%, which might be indistinguishable from normal samples by traditional methods.

2.6 False Positive Review Reveals True Positive Relapse

Interestingly, CAREMIL's performance at low levels of circulating blasts in synthetic datasets appeared substantially lower than in the WSI dataset. The blast percentages in our WSI dataset spans from 0% to 100% in cases of acute leukemia. Slides were included in the study based on pathology reports diagnosing acute leukemia, even if circulating blasts were not detected by hematology analyzers or visual inspection noted in the pathology report. In cases where circulating blast percentages were below 20%, patients typically exhibited higher blast counts in the bone marrow, consistent with the diagnostic criteria for acute leukemia.

In Figure 3B, the continuous line delineates the performance of CAREMIL across varying levels of circulating disease. For clarity, we categorized our test data into different bins: high (>=20%), low (<20%), very low (<=10%), extremely low (<=5%), and invisible leukemia (<=1%). CAREMIL demonstrates high performance at low levels of blasts on real patients compared to synthetic patients, which are represented by the dashed line. Surprisingly, CAREMIL achieves 75% sensitivity, even in cases of with 0% blast count. This suggests that CAREMIL may be recognizing changes in the composition or morphology of the non-blast cell population. However, these changes are not necessarily specific to acute leukemia and could also occur in other hematological disorders. For instance, they may result from marrow damage caused by leukemia cells, leading to alterations in peripheral blood morphology or cytometry. Alternatively, acute leukemia might induce specific but subtle, multifactorial changes in non-blast cells that are not easily recognizable by pathologists, but identifiable by CAREMIL.

Although missed diagnoses in cases with low-level diseases are anticipated, encountering false positives where normal slides are misclassified as acute leukemia is less expected. In our test set, no such instances occurred. However, in the validation set, four cases were identified. Upon review by our pathologist colleagues, two were found in patients with a history of leukemia but no current detectable disease. Another involved a patient who had recently received granulocyte colony stimulating factor (G-CSF) treatment prior to a bone marrow transplant for lymphoma. For patients whose bone marrow stem cells are being harvested for transplant, G-CSF treatment is performed to massively increase circulating hematopoietic stem cells (blasts) prior to harvesting.

The most notable "false positive" involved a patient with a history of acute myeloid leukemia (AML) in remission who presented to the clinic with an enlarged lymph node. Although imaging raised concerns for relapse, initial diagnostic evaluations, including a peripheral blood smear, bone marrow biopsy, bone marrow flow cytometry, and bone marrow biopsy genetic testing for a known *FLT3* mutation did not confirm disease recurrence. Based on this workup, the bone marrow biopsy report was classified as morphologically normal, resulting in its inclusion in our test set as a morphologically normal sample. However, a lymph node biopsy confirmed AML, prompting a repeat bone

marrow biopsy three weeks later, which also revealed AML with an 80% blast count. This suggests that CAREMIL may have identified the presence of AML approximately three weeks earlier than the standard diagnostic workup.

2.7 Model-Derived Morphometric Signatures Reflect Underlying Disease Biology

The attention values created from the model provide valuable interpreteability beyond the individual patient level. For each individual patient we create an attention panel as shown in Figure 3, through which a pathologist can verify the cells being looked into for each prediction. To evaluate the overall effectiveness of our interpretability mechanism beyond individual attention maps, we compare the cell type distribution within the top 10% most-attended cells in each disease category to the distribution of all cells in the same disease category. This highlights the cell types that are relevant to the decision making for each disease category.

As shown in 4A the relative composition of blast cells in top 10% highest attention cells is 2.5 times higher than in the total population, which indicates that the attention mechanism is correctly identifying blasts as the important cell class for diagnosing AML. Supplemental Figure 4C extends this analysis using a log-odds ratio to highlight the contribution of rare cell classes to the disease signature. Notably, we observe increases in mitotic bodies, orthochromic/polychromatophilic erythroblasts, plasma cells, and basophils. While mitotic bodies are rarely seen in circulation, their presence is elevated in acute leukemia, and the model appears to have captured this pattern. The increases in other rare circulating cell types likely reflect underlying marrow stress, further supporting the model's ability to detect subtle morphological indicators of disease.

In Figure 4B, the MDS disease signature is illustrated. For MDS patients, the model assigns high attention to lymphocytes, blasts, and basophils. We hypothesize that the increased attention to lymphocytes is due to the relative lymphocytosis observed in patients with myeloid cytopenias, a hallmark feature of MDS. On further examination some patients in the cohort had lymphocyte percentages in excess of 50% of all circulating white blood cells. Likewise, abnormal blasts and basophilia are also commonly seen in MDS.

The comparative attention histogram for HCL shows attention less focused on lymphocytes. Thus, the DeepHeme model does not classify the hairy cells as normal lymphocytes. Figure 4C shows that the high attention category contains overrepresentation of monocytes and blasts. Pathologist review revealed that the cells interpreted as blasts and monocytes by DeepHeme in these samples were hairy cells. DeepHeme is trained on patients with no known neoplastic disease. Thus, hairy cells were absent from the training corpora. Despite this limitation, the CAREMIL model determined the features associated with hairy cells was important to the underlying diagnosis despite the encoder model (DeepHeme) not being trained to specifically classify hairy cells. This highlights the strength of the CAREMIL model, which does not rely on the cell-level classification provided by the encoder. Even when the

cell labels are incorrect, the model can learn to identify the important cell morphologies provided by the feature embedding from the encoder.

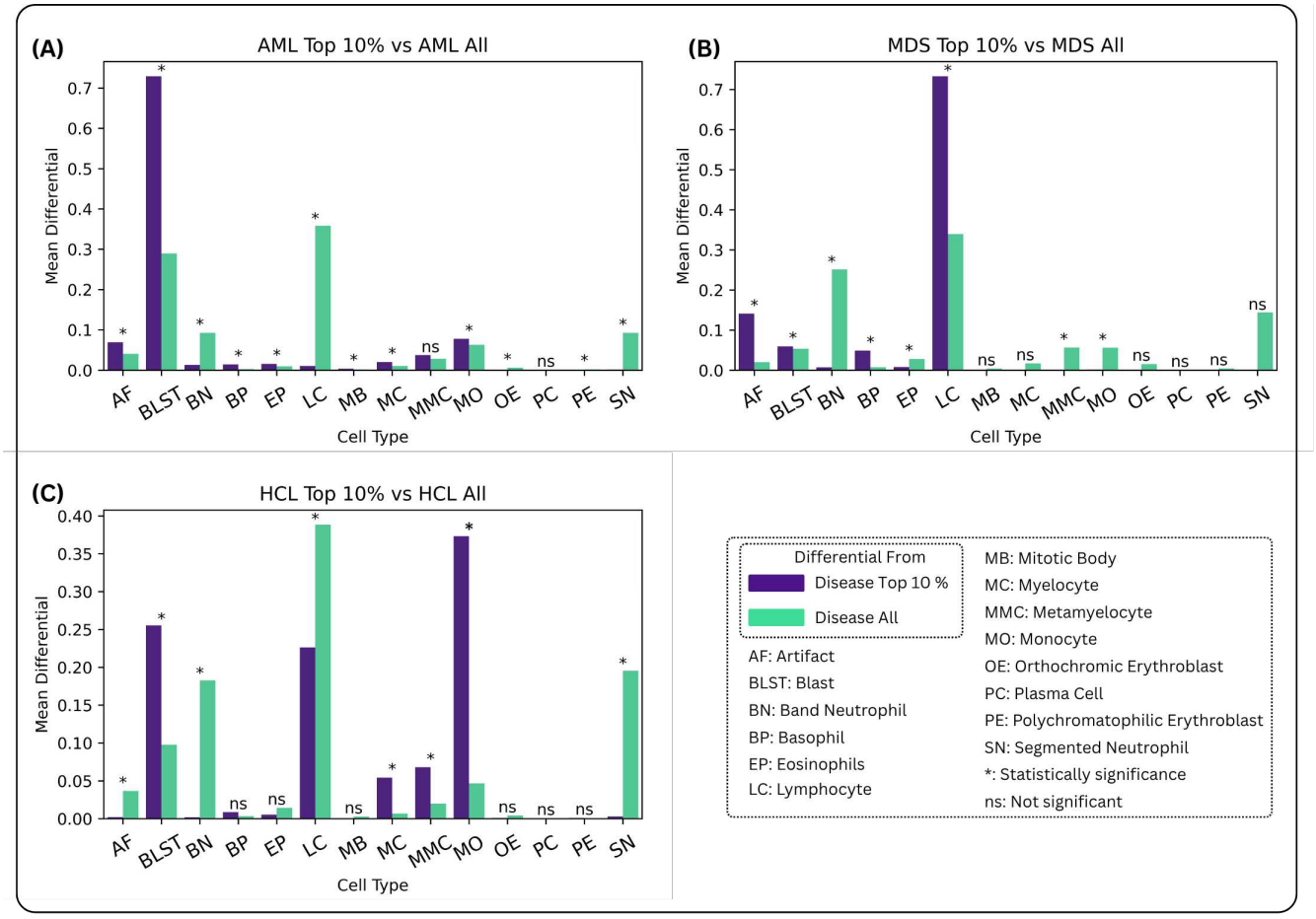

Figure 4. Comparative Attention Histograms. Each graph shows the distribution of cell types among the top 10% highest-attention cells, identified by the CAREMIL+DeepHeme architecture, in correctly-classified cases, compared to the overall distribution of cell types in the same cases. A chi-squared test is used to determine the statistical significance of the enrichment for each cell type in the top 10% highest-attention cells vs. all cells. A Bonferroni correction is used to account for the multiple hypotheses being tested. Statistically significant differences are highlighted with *. Bar chart for A) AML, B) MDS, C) HCL

2.8 Model attention aligns with diagnostic cell identity

The disease signature helps pathologists review the cells the model takes into consideration for diagnosis along with the weight given to each cell. However, this pictorial and manual verification is not enough to prove the model's capabilities. We want to understand the overlap between the cells that are given a high attention value (top 25% by attention value) by the model vs the diagnostically relevant cells identified for that patient. We choose a sample that has been diagnosed with AML for this analysis, as AML has a simple biomarker of the presence of blast cells. In Fig 5 almost all high attention cells are blast cells showcasing the practical utility of the model to identify important diagnostic markers without the need to explicitly teach the model specific biomarkers for every disease.

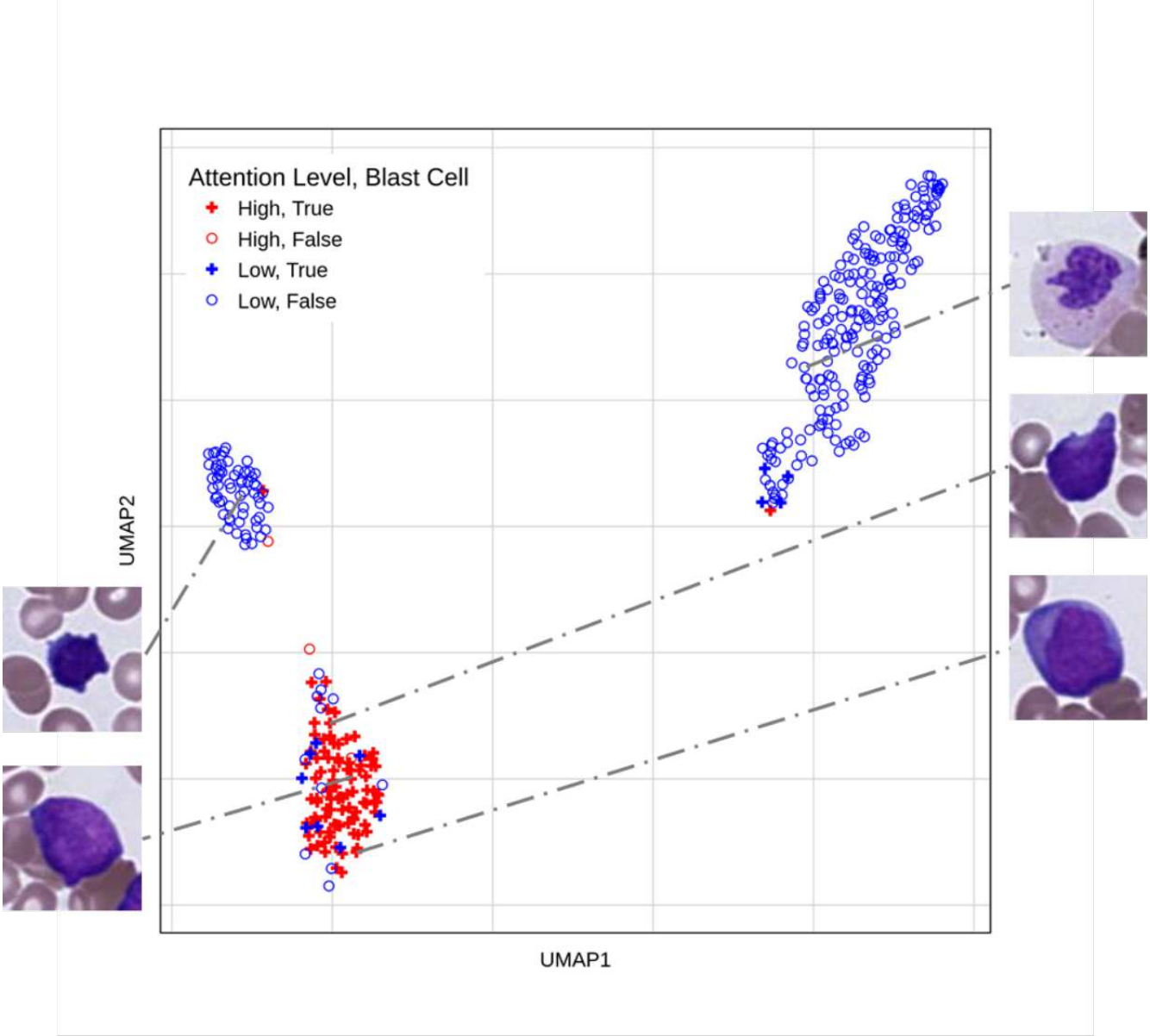

Figure 5. Verifying high attention cells with diagnostically relevant cells. This is a UMAP of the cell embeddings of a patient diagnosed with AML. High attention cells are shown in red, with low attention cells in blue. Blasts, a biomarker of acute leukemia, are depicted as filled diamonds, whereas other cells are empty circles. Most high attention cells (top 25% cells by attention value from the model) are blast cells which are a biomarker for AML diagnoses (filled red diamonds).

3 Discussion

Overall we proposed a novel framework CAREMIL with DeepHeme which is the best performative model for hematological diagnostic tasks. Instead of traditional diagnosis of hematologic cancers which rely on enumerating different blood cell types (the blood differential) and applying heuristic rules alongside morphological analysis of individual cells, our pipeline leverages a hematologic domain-specific encoder to integrate cellular morphology with cytometric features for case-level diagnosis of hematologic malignancies from blood smears.

Unlike prior AI efforts that classify cells individually and base diagnosis on cell differentials alone, our model also integrates cell-level morphologies across the entire case to inform diagnostic predictions. Moreover, conventional cell classifiers often offer limited interpretability, making it difficult to meaningfully involve pathologists in the diagnostic process (25), a limitation CAREMIL explicitly addresses by providing cell-level attention scores.

In computational pathology for tissue sections, whole slide images (WSIs) are analyzed through a tiling approach for slide-level diagnosis. However, these methods are unsuitable for hematology, where diagnostically relevant cells, such as white blood cells (WBCs), are sparsely and randomly distributed. Our proposed methods overcomes these limitations by implementing a cell-aggregation-based multiple instance learning (MIL) method with attention mechanisms that prioritize diagnostically important cells (26) for pathologist review, facilitating accurate slide-level predictions.

Our findings demonstrate that CAREMIL paired with DeepHeme cell image encoding achieves substantial improvements in diagnostic accuracy by combining individual cell-level morphological features with cytometric composition, which is pivotal for diagnosing and staging hematologic malignancies. Notably, our pipeline achieved AUROC scores of 0.999, 0.891 and 0.945 for AML, MDS, and HCL, respectively, underscoring its clinical relevance across diverse disease contexts. We showed that use of DeepHeme outperforms other image encoders, including pathology foundation models and ImageNet-trained encoders. This result highlights the importance of domain-specific feature extraction for arriving at the most accurate diagnosis.

An important aspect of the MIL strategy for hematologic diagnosis is the explainability framework we highlight. By employing an attention-based MIL aggregator, the model identifies cells critical to the diagnostic prediction, allowing pathologists to validate and interpret the model's rationale in the context of each case. This explainability, alongside performance, is essential for clinical adoption, as it enables medical professionals to ensure model predictions align with established diagnostic criteria, thereby promoting clinician trust (27) and patient safety.

The attention-based weakly supervised approach employed in CAREMIL offers unique insights into underlying disease processes by identifying diagnostically relevant cells, rather than relying solely on quantitative metrics

like blast counts. Specific genotypes of AML associated with aggressive phenotypes do not require a specific blast percentage for diagnosis, according to the 5th edition of the World Health Organization Classification of Haematolymphoid Tumours(28). This pipeline can be replicated with different disease labels to build an automated biomarker identification system that will emerge as a valuable tool in the future to stratify high-risk patients, enabling earlier and more precise interventions. This approach can also be used to develop outcome prediction models that are morphology informed, using outcome labels rather than diagnosis labels.

The ability to assign attention values to individual cells based on their relevance to the diagnostic outcome allows for a nuanced interpretation of WSI data that captures both explicit and latent features across the entire sample. This model of disease characterization holds promise for future applications, where attention-based diagnostics may surpass traditional metrics in robustness and specificity, especially for diseases with complex presentations. By focusing on pathologically significant cells rather than fixed thresholds, CAREMIL could enable more personalized and accurate diagnostics that reflect the continuum of disease rather than categorical criteria.

While our techniques demonstrate strong diagnostic potential, several limitations underscore areas for future development. The current approach does not support end-to-end training or fine-tuning, which may be necessary for applications lacking pre-existing in-domain feature encoders (non-hematopoietic cells in cytology applications, for example). Additionally, DeepHeme generates feature embeddings from a relatively modest-sized ResNeXt-50 model, and future implementations may benefit from larger, self-supervised architectures to achieve even better performance, especially as diagnostic complexity increases. (29)

Another consideration is the diagnostic spectrum of the cell-based MIL strategy. This study addresses a four-way classification, yet the full range of hematologic diagnoses is far more extensive. Expanding the study to encompass additional conditions will require large, diverse datasets and further testing to ensure reliability across a broader range of hematologic pathologies. Moreover, while this study focuses solely on peripheral blood smears—a practical choice for screening due to their low cost and minimally invasive nature—a more accurate diagnosis would be achieved by integrating data from bone marrow aspirates, core biopsies, immunohistochemistry, flow cytometry, clinical text histories, and genomic data. Conveniently, the current pipeline provides an encoding strategy that could allow for incorporation in a multimodal study that includes these diverse data types.

The MIL architecture, CAREMIL along with the DeepHeme cell encoder, represents a significant advancement in computational hematology, providing an accurate, interpretable framework for diagnosing hematologic diseases using WSIs of peripheral blood smears. Leveraging a cell-aggregation-based multiple instance learning approach not only preserves essential morphologic information of white blood cells but also achieves superior diagnostic

accuracy compared to cytometry-based models. Future work should explore expanding this pipeline's diagnostic scope and integrating multimodal data to enhance its clinical utility in hematology.

4 Acknowledgements

This research was supported by funding from the Cancer Center Support Grant from the NIH/NCI (P30CA008748), the Warren Alpert Foundation through the Warren Alpert Center for Digital and Computational Pathology at Memorial Sloan Kettering Cancer Center, and the MSK Technology Development Fund. We extend our gratitude to the MSK Hematopathology and Digital Pathology Services for their contributions.

Declarations

Data Availability

Clinical datasets used for this study are not made available in accordance with institutional policies.

Code Availability

Code used in this work is available on GitHub: https://github.com/goldgoflab/Slide-Level-Classificati
tree/nature_comms.

References

1. Kimura, K. *et al.* Automated diagnostic support system with deep learning algorithms for distinction of Philadelphia chromosome-negative myeloproliferative neoplasms using peripheral blood specimen. *Sci. Reports* **11**, 3367 (2021). DOI 10.1038/s41598-021-82826-9.

2. Foucar, K. *et al.* Concordance among hematopathologists in classifying blasts plus promonocytes: a bone marrow pathology group study. *Int. J. Lab. Hematol.* **42**, 418–422 (2020).

3. Döhner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute myeloid leukemia. *New Engl. J. Medicine* **373**, 1136–1152 (2015).

4. Pelcovits, A. & Niroula, R. Acute myeloid leukemia: a review. *Rhode Island medical journal* **103**, 38–40 (2020).

5. Wintrobe, M. M. Clinical hematology. *Acad. Medicine* **37**, 78 (1962).

6. Chase, M. L. *et al.* Consensus recommendations on peripheral blood smear review: defining curricular standards and fellow competency. *Blood Adv.* **7**, 3244–3252 (2023). DOI 10.1182/bloodadvances.2023009843.

7. Campo, E. *et al.* The international consensus classification of mature lymphoid neoplasms: a report from the clinical advisory committee. *Blood, The J. Am. Soc. Hematol.* **140**, 1229–1253 (2022).

8. Sidhom, J.-W. *et al.* Deep Learning for Distinguishing Morphological Features of Acute Promyelocytic Leukemia. *Blood* **136**, 10–12 (2020). URL https://doi.org/10.1182/blood-2020-135836. DOI 10.1182/blood-2020-135836.

9. Matek, C., Schwarz, S., Spiekermann, K. & Marr, C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nat. Mach. Intell.* **1**, 538–544 (2019). URL https://www.nature.com/articles/s42256-019-0101-9. DOI 10.1038/s42256-019-0101-9. Publisher: Nature Publishing Group.

10. Wang, J. Deep Learning in Hematology: From Molecules to Patients. *Clin. Hematol. Int.* **6** (2024). URL https://chi.scholasticahq.com/article/124131-deep-learning-in-hematology-from-molecules-to-patients. DOI 10.46989/001c.124131.

11. Campanella, G. *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Medicine* **25**, 1301–1309 (2019). URL https://www.nature.com/articles/s41591-019-0508-1. DOI 10.1038/s41591-019-0508-1. Epub 2019 Jul 15, PMC7418463.

12. Lu, M. Y. *et al.* Data efficient and weakly supervised computational pathology on whole slide images. URL http://arxiv.org/abs/2004.09666. DOI 10.48550/arXiv.2004.09666. 2004.09666[cs,eess,q-bio].

13. Landau, M. S. & Pantanowitz, L. Artificial intelligence in cytopathology: a review of the literature and overview of commercial landscape. *J. Am. Soc. Cytopathol.* **8**, 230–241 (2019).

14. Sun, S. *et al.* Deepheme, a high-performance, generalizable deep ensemble for bone marrow morphometry and hematologic diagnosis. *Sci. Transl. Medicine* **17**, eadq2162 (2025). URL https://doi.org/10.1126/scitranslmed.adq2162. DOI 10.1126/scitranslmed.adq2162.

15. Matek, C., Krappe, S., Münzenmayer, C., Haferlach, T. & Marr, C. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image dataset. *Blood* (2021). URL https://www.sciencedirect.com/science/article/pii/S0006497121013975. DOI 10.1182/blood.2020010568.

16. Song, A. H. *et al.* Morphological Prototyping for Unsupervised Slide Representation Learning in Computational Pathology (2024). DOI 10.48550/arXiv.2405.11643. 2405.11643.

17. Ilse, M., Tomczak, J. & Welling, M. Attention-based Deep Multiple Instance Learning. In *Proceedings of the 35th International Conference on Machine Learning*, 2127–2136 (PMLR, 2018). URL https://proceedings.mlr.press/v80/ilse18a.html. ISSN: 2640-3498.

18. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021). DOI 10.1038/s41551-020-00682-w.

19. Gadermayr, M. & Tschuchnig, M. Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Comput. Med. Imaging Graph.* **112**, 102337 (2024).

20. Javed, S. A. *et al.* Additive mil: Intrinsically interpretable multiple instance learning for pathology. *Adv. Neural Inf. Process. Syst.* **35**, 20689–20702 (2022).

21. Deng, R. *et al.* Cross-scale multi-instance learning for pathological image diagnosis. *Med. image analysis* **94**, 103124 (2024).

22. Manescu, P. *et al.* Detection of acute promyelocytic leukemia in peripheral blood and bone marrow with annotation-free deep learning. *Sci. Reports* **13**, 2562 (2023). DOI 10.1038/s41598-023-29160-4.

23. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat. Medicine* (2024).

24. Zimmermann, E. *et al.* Virchow2: Scaling self-supervised mixed magnification models in pathology (2024). URL https://arxiv.org/abs/2408.00738. 2408.00738.

25. Srisuwananukorn, A., Salama, M. E. & Pearson, A. T. Deep learning applications in visual data for benign and malignant hematologic conditions: a systematic review and visual glossary. *Haematologica* **108**, 1993–2010 (2023). URL https://haematologica.org/article/view/haematol.2021.280209. DOI 10.3324/haematol.2021.280209. Number: 8.

26. Araújo, D. J. *et al.* Key Patches Are All You Need: A Multiple Instance Learning Framework For Robust Medical Diagnosis (2024). URL http://arxiv.org/abs/2405.01654. DOI 10.48550/arXiv.2405.01654. ArXiv:2405.01654 [cs].

27. Gaube, S. *et al.* Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Sci. Reports* **13**, 1383 (2023). URL https://www.nature.com/articles/s41598-023-28633-w. DOI 10.1038/s41598-023-28633-w. Publisher: Nature Publishing Group.

28. Khoury, J. D. *et al.* The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. *Leukemia* **36**, 1703–1719 (2022). URL https://www.nature.com/articles/s41375-022-01613-1. DOI 10.1038/s41375-022-01613-1. Number: 7 Publisher: Nature Publishing Group.

29. Wong, B., Hong, S. & Yi, M. Y. Rethinking Pre-Trained Feature Extractor Selection in Multiple Instance Learning for Whole Slide Image Classification (2025). URL http://arxiv.org/abs/2408.01167. DOI 10.48550/arXiv.2408.01167. ArXiv:2408.01167 [cs].

1 Methods

1.1 Dataset Collection

We curated a dataset of 1,017 WSIs of blood smears scanned at 400× optical resolution (i.e. using a 40x objective) using a Hamamatsu S360 WSI scanner. These include 307 acute leakemia, 95 myelodysplastic syndromes, 41 hairy cell leukemias, and 130 morphologically normal control cases (see Supplementary Figure 8A) Scanned images range from 2GB to 20GB and were scanned using a single z-plane with uniform refocusing points across each slide. Conventional computer vision methods that rely on manually defined image features for classification often have difficulty with WSIs, as they include complex biological structures (cell nuclei, cytoplasms, granules, etc.) with high variability in appearance across multiple patients and disease states (such as cancer grades or stages). To reduce the computation required for processing these gigapixel WSIs, we first extract regions of interest containing WBCs at 50× optical resolution (downsampled from the highest 400× resolution by a factor of 8) using a combination of deep learning and classical computer vision metrics. From these regions, we extract cell images of size 96 × 96 at 400× magnification, cropped around the cell centers detected by a YOLOv8 model[1] (Figure 1A)). We examined a cohort of 573 patients (Supplementary Table 8A). For each WSI, we extract ∼ 2000 WBC images, and subsequently extract 1000-dimensional feature embeddings for each of them using the last-layer features of a ResNeXt-50 model trained specifically on WBC classification task[2].

1.2 Incorporating cell morphology into the predictions

Current computational approaches for the diagnosis of hematologic neoplasms involve predicting the blood differential through summation of individual cell classifications. Trained ML classifiers use the cytometric vector as input to predict diagnosis. This approach loses morphological information of individual cells, which pathologists use when making diagnostic decisions. Moreover, it lacks cell-level interpretability, without which the final predictions are difficult for pathologists to interpret and audit, making it difficult to use in a clinical setting.

To incorporate individual cell morphology, we propose a trained encoder to first generate vector-embeddings of all the extracted cell images (Figure 1B). Our MIL model then aggregates these cell embeddings and makes a slide-level prediction. Explainability is built into the MIL aggregator to facilitate pathologists' interpretation of the model's predictions.

1.3 CAREMIL model architecture

The CAREMIL architecture[1] is built to ensure that we can take a varying number of cells as input and get relevant attention values for each cell embedding, while making a diagnosis prediction (see Figure 1C . Once all the cell images are extracted, they are passed to a frozen image encoder, that creates embedding values that represent the morphological information in each cell image.

The cell embeddings are passed to different trainable networks, namely the aggregator network and the parallel network, both will reduce each cell's embedding size to the same dimension. The embeddings outputs from the aggregator network are then passed to aggregator functions. We explore several aggregation functions (Supplementary Figure 8C). For example, Kraus et al.[3] showed that generalized mean can lead to better predictions in some computational pathology applications, whereas, Carmichael et al.[4] showed that variance-based aggregation function can lead to better predictions by capturing intratumoral heterogeneity. These function outputs represent compositional information of the bloodstream, similar to the cytometric vector produced in the cytometric based ML classifiers. The vectors are concatenated with the outputs of the parallel network before being fed into the self-attention layer. Here the aggregation function outputs act as class tokens[5], similar to BERT[6] and Vision Transformers[7]. The class tokens from the attention layer are then fed into a multilayer perceptron (MLP) classifier to predict logits and train the network using a cross-entropy loss.

1.4 Case-Level Attention Visualizations

To validate CAREMIL's interpretability in diagnosing hematologic malignancies, attention panel images were generated for cases of acute myeloid leukemia (AML), hairy cell leukemia (HCL), and myelodysplastic syndromes

[1] https://github.com/goldgoflab/Slide-Level-Classification/tree/nature_comms

(MDS). For each case, we selected the top 10% of cells with the highest attention scores, a random sample of cells, and the lowest attention-scored cells. These panels allowed visual verification of the model's focus, with high-attention cells expected to show disease-relevant morphologies (e.g., blasts in AML, dysplastic myeloid cells in MDS, and abnormal lymphocytes in HCL). Randomly selected cells provide a sampling of the sample's overall composition, while low-attention cells should include morphologically irrelevant or non-diagnostic cells. This approach enables assessment of CAREMIL's attention mechanism in identifying cells most relevant to the diagnostic outcome.

1.5 Limit of Detection Experiment

To assess CAREMIL's sensitivity for detecting acute leukemia (AL) in cases with low blast counts, synthetic patient cases were generated by combining cells from normal patients with varying proportions of blast cells. For each synthetic case, cells from normal peripheral blood smear images were incrementally mixed with an increasing percentage of blast cells (ranging from 0% to 20%), simulating conditions of low to moderate blast presence. CAREMIL was then run on each synthetic patient case to evaluate the model's probability of predicting AL at different experimental blast levels (ref Figure 3) . Real patient cases with similar blast count distributions were also evaluated for comparison, allowing analysis of CAREMIL's performance across synthetic and real datasets.

1.6 Attention Mechanism Evaluation

To evaluate how CAREMIL's attention mechanism selectively prioritizes specific cell types in different hematologic malignancies, we compared the distribution of high-attention cells across acute leukemia (AL), myelodysplastic syndrome (MDS), and hairy cell leukemia (HCL) cases. For each disease type, we identified the cell types based on DeepHeme cell classification top 10% highest attention cells as outputted by the MIL model. This allows for calculation of the mean differential distribution across various cell types with the highest attention. These distributions were then compared to the overall cell population in normal samples to identify disease-specific enrichment patterns. This analysis provides insight into the model's diagnostic process, enabling validation of the attention mechanism's interpretability and effectiveness in distinguishing distinct hematologic conditions.

References

1. Reis, D., Kupec, J., Hong, J. & Daoudi, A. Real-time flying object detection with yolov8. *arXiv* (2023). DOI 10.48550/arxiv.2305.09972.

2. Goldgof, G. M. *et al.* DeepHeme: A generalizable, bone marrow classifier with hematopathologist-level performance (2023). URL https://www.biorxiv.org/content/10.1101/2023.02.20.528987v1. DOI 10.1101/2023.02.20.528987. Pages: 2023.02.20.528987 Section: New Results.

3. Kraus, O. Z., Ba, L. J. & Frey, B. Classifying and Segmenting Microscopy Images Using Convolutional Multiple Instance Learning. *Bioinformatics* **32**, i52–i59 (2016). URL http://arxiv.org/abs/1511.05286. DOI 10.1093/bioinformatics/btw252. ArXiv:1511.05286 [cs, q-bio, stat].

4. Carmichael, I. *et al.* Incorporating intratumoral heterogeneity into weakly-supervised deep learning models via variance pooling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 387–397 (Springer, 2022).

5. Mingote, V., Miguel, A., Ortega, A. & Lleida, E. Class Token and Knowledge Distillation for Multi-head Self-Attention Speaker Verification Systems (2023). URL http://arxiv.org/abs/2111.03842. DOI 10.48550/arXiv.2111.03842. ArXiv:2111.03842 [eess].

6. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019). URL http://arxiv.org/abs/1810.04805. DOI 10.48550/arXiv.1810.04805. ArXiv:1810.04805 [cs].

7. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2020). URL https://arxiv.org/abs/2010.11929v2.

Supplementary Files

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

- naturecommunicationschecklistsigned.pdf
- SupplementalPaper.pdf