Proceedings of the 7th International Conference on Inventive Research in Computing Applications (ICIRCA-2026)
DVD Part Number: CFP26N67-DVD; ISBN: 979-8-3315-5939-7

MorphX: A WHO-Aligned Ante-Hoc Explainable AI Framework for Acute Lymphoblastic Leukaemia Diagnosis from Peripheral Blood Smear Images

Lourdu Rayappan
Research Scholar
Department of Computer Science and Information Technology
School of Computing Sciences
Vels Institute of Science, Technology & Advanced Studies
Pallavaram, Chennai, India
abyssrayappan@gmail.com

R. Parameshwari
Professor,
Department of Computer Science and Information Technology
School of Computing Sciences
Vels Institute of Science, Technology & Advanced Studies
Pallavaram, Chennai, India
parameswari.scs@vistas.ac.in

***Abstract*— We present MorphX (morphology-informed explainable AI), an ante-hoc deep learning framework for interpretable diagnosis of acute lymphoblastic leukaemia (ALL) from peripheral blood smear images. MorphX combines a morphology-aware feature extraction backbone with a World Health Organization (WHO)-aligned concept bottleneck that predicts 28 morphological descriptors, including chromatin texture, nuclear contour irregularity, and cytoplasmic granularity. These concept activations are mapped through a differentiable WHO layer to a clinically weighted scalar risk score. To align training with diagnostic priorities, we used a risk-weighted objective that placed a greater emphasis on clinically important descriptors. MorphX further integrates concept-level attribution, Grad-CAM++ visual overlays, concept sensitivity scoring in the WHO concept space, and sparse counterfactual WHO perturbations that identify the minimal descriptor changes required to alter the predicted class. On average, only 7.2 of the 28 features required adjustment per case (mean $|\Delta| = 0.361$), indicating compact and semantically constrained counterfactual explanations. Across the ablation settings, the WHO feature supervision preserved a strong classification performance while reducing the mean squared error of the concept prediction by 15.2%. The framework achieved a classification accuracy of 94.1% and an AUC of above 0.97 on the ALL-IDB2 dataset. MorphX also generates a structured clinical report artefact that combines diagnostic output, concept vectors, risk scores, visual evidence, and counterfactual summaries. These results show that MorphX improves the interpretability of ALL image analysis while maintaining a strong predictive performance in the present dataset.**

Keywords*— *Explainable AI, Hematopathology, Acute Lymphoblastic Leukaemia, Concept Bottleneck Models, WHO-Aligned Concepts, Differentiable Risk Mapping, Grad-CAM++, TCAV, Counterfactual Explanations, Peripheral Blood Smear Analysis, Deep Learning.

I. INTRODUCTION

Deep learning has achieved strong performance in medical imaging; however, the limited transparency of many models remains a major barrier to clinical trust in high-stakes settings [1]. In blood smear analysis, explainability is important because clinicians do not trust a diagnosis from a score alone; they want to see the morphological evidence behind it. Therefore, an explainable system should make its reasoning reviewable in clinical terms, link predictions to named WHO-style descriptors, highlight uncertain or contestable cases, and support safer use than a black-box probability alone [3][4][5].

Early research on explainable artificial intelligence (XAI) in the field of medical imaging has predominantly focused on post-hoc methodologies that attempt to explain model predictions after the fact. Saliency maps, generated by techniques such as Grad-CAM, highlight regions that influence model outputs using the gradient flowing into the last convolutional layer, pinpointing which image areas "activate" the network for a given prediction [6]. However, saliency maps mainly show where the model is looking and not the morphological evidence it uses. This limitation is important in hematopathology, where experts reason through named cellular descriptors rather than pixel-level regions. In addition, saliency maps can appear plausible without faithfully reflecting the true decision rule and may remain unstable or noisy in some cases. In our evaluation, this was reflected by a high saliency sparsity (0.96), low structural similarity (approximately 0.04), and only modest saliency-attention correlation (approximately 0.12) [7][8].

Other model-agnostic approaches, including LIME, relevance propagation, and perturbation-based testing, provide useful post-hoc information about local model behaviour [2][5][9]. However, post-hoc explanations remain insufficient for clinical trust because they explain the decision only after the classifier has made a decision. It does not require the model to reason through named clinical concepts, cannot reliably identify which descriptors drive the diagnosis, and may produce convincing heatmaps without a reliable clinical rationale. Previous studies have shown that black-box models may rely on spurious cues or dataset-specific shortcuts, which can reduce their reliability on new data [7][10][11][12]. These observations motivate the need for models in which interpretability is part of the predictive pathway rather than an external add-on.

This need has led to growing interest in interpretability by design [7]. Concept bottleneck models address this problem by predicting human-defined concepts as an intermediate step and using these concepts to support the final decision-making [13]. Such models are especially relevant in hematopathology, where diagnosis depends on codified morphological descriptors such as chromatin texture, nuclear shape, nucleoli, and cytoplasmic appearance [14]. Embedding these descriptors directly into the model provides a route to predictions that are both accurate and clinically interpretable.

To address these limitations, we introduced MorphX, a morphology-informed explainable framework for ALL diagnosis using peripheral blood smear images. MorphX uses the ALL-MorphNet backbone [16] as a feature extraction module and augments it with a WHO-aligned concept bottleneck that predicts 28 morphologic descriptors relevant to ALL, including chromatin pattern, nuclear contour irregularity, and nucleolar prominence [14]. These descriptor predictions serve as interpretable intermediate representations between images and final diagnostic outputs. By embedding this WHO-

aligned concept layer into the predictive pathway, MorphX constrains internal representations to remain clinically meaningful while preserving end-to-end learning.

MorphX further incorporates a clinically weighted risk mapping layer and risk-weighted concept objective so that diagnostically important descriptors receive greater emphasis during training. In addition to this ante-hoc concept structure, the framework provides integrated explanation modules, including Grad-CAM++ visual overlays, concept-level attribution, concept-sensitivity analysis in the World Health Organization (WHO) concept space, and sparse counterfactual perturbations that identify the minimal descriptor changes required to alter the predicted diagnosis [15]. These components allow the model output to be examined visually, semantically, and contrastively within the same concept space.

The contribution of this study lies in integrating a World Health Organization (WHO)-aligned concept bottleneck, clinically weighted differentiable risk mapping, concept-space sensitivity analysis, uncertainty-aware interpretation, sparse counterfactual reasoning, and structured report generation within a single diagnostic framework for ALL smear-image analysis [1]. Unlike conventional post-hoc explainability approaches, MorphX embeds clinically named morphological descriptors directly into the predictive pathway, allowing diagnostic outputs to be interpreted in a semantically meaningful conceptual space. Although MorphX uses a previously developed morphology-aware backbone [16], the present study is self-contained and focuses on the explainability architecture, its outputs, and its evaluation on the ALL-IDB2 dataset. Therefore, the framework links classification, concept prediction, visual explanation, and contrastive explanation within a unified model representation. The following sections describe the dataset, model design, explanation modules, and experimental evaluation.

II. Proposed Methodology

A. Dataset and Experimental Setup

The MorphX framework was evaluated on the ALL-IDB2 dataset [17], a widely used benchmark comprising 260 high-resolution peripheral blood smear images, including 130 lymphoblasts and 130 normal lymphocytes, each measuring 257 × 257 pixels in size. The images were cropped from ALL-IDB1 [17] and acquired using a Canon PowerShot G5 camera mounted on a microscope at 300×-500× magnification. This resolution preserves diagnostically relevant morphological details, including chromatin texture, cytoplasmic granularity, and vacuolisation, making the dataset suitable for concept-level evaluation in a controlled single-cell setting. In this study, each image was treated as an individual sample for classification, concept prediction, saliency analysis and counterfactual evaluation. Accordingly, the reported results should be interpreted as proof of method within the ALL-IDB2 setting, with broader validation still required across multi-institutional data.

All experiments were conducted on a Windows 10 system with WSL 2 running Ubuntu 22.04, using an Intel 6-core 2.59 GHz CPU, 32 GB RAM, 400 GB storage, and an NVIDIA RTX 2070A Max-Q GPU with 6 GB of memory. This setup was sufficient for training, ablation analysis, and explanation generation, although large-scale studies will require higher-throughput hardware.

B. System Architecture

MorphX is an ante-hoc explainable framework in which clinically named concepts are embedded within the predictive pathway rather than appended after classification. Fig. 1a shows the overall MorphX pipeline from the peripheral blood smear image input to the structured diagnostic output. Fig. 1b shows the corresponding inference and explanation workflow, including concept prediction, risk mapping, visual explanation, counterfactual analysis, and uncertainty estimation. Table I compares MorphX with common XAI approaches and highlights the roles of concept-guided prediction, clinically weighted risk estimation, and counterfactual reasoning in the proposed framework.

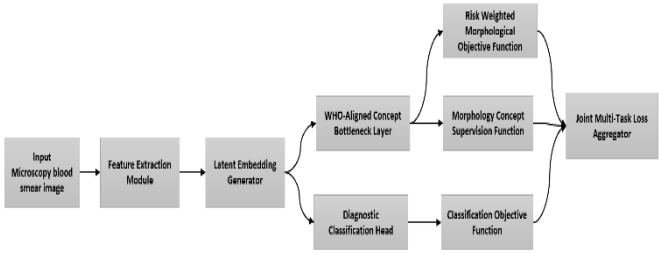

Figure-1a: Pipeline Diagram of MorphX

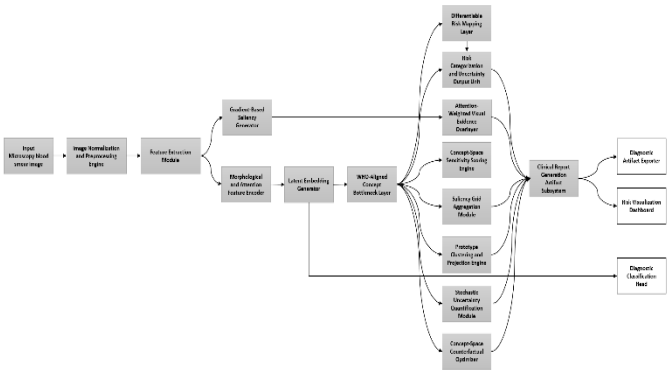

Figure-1b: Schematic Block Diagram of the Inference and Explanation Pipeline

Backbone and WHO Concept Bottleneck: As shown in Fig. 1a, MorphX uses ALL-MorphNet [16] as its morphology-aware backbone to extract the local texture, multi-scale cellular structure, and broader visual context from peripheral blood smear images. These morphology-rich features are then mapped to a 28-dimensional WHO morphology concept vector,

$$\mathbf{c} = [c1, c2, \ldots, c28]$$

where each $c_j \in [0,1]$ denotes the predicted activation of a named WHO descriptor. This concept vector serves as an interpretable intermediate representation that links the image input to a clinically readable concept space and forms the basis of the downstream diagnostic and explanation modules.

Differentiable WHO Layer: Fig. 1b shows that the predicted WHO concept vector is passed to a differentiable WHO layer that maps the descriptor activations to a scalar risk score using clinically defined relevance weights. Let c_{ij} denote the predicted activation of descriptor j for sample i, and let w_j denote the corresponding clinical importance weight. The scalar risk score is computed as

$$r_i = \frac{\sum_{j=1}^{28} w_j c_{ij}}{\sum_{j=1}^{28} w_j}.$$

This formulation produces a transparent score that remains comparable across samples and directly reflects the descriptor-level clinical relevance.

Risk-Weighted WHO Loss: MorphX uses a risk-weighted WHO loss to prioritise diagnostically important descriptors during training, Let y_{ij} denote the target value of descriptor j for sample i, and let $l(.,.)$ denotes the descriptor-level loss, and B denotes the batch size. The loss is defined as

$$\mathcal{L}_{\text{WHO}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{28} w_j \, \ell(c_{ij}, y_{ij}).$$

This loss increases the contribution of clinically important descriptors to the optimisation objective and aligns the concept learning with the final classification task.

TCAV-Based Concept Sensitivity: To quantify the directional influence of each WHO descriptor on the diagnostic output, MorphX applies TCAV-style sensitivity analysis directly in the concept space. As summarised in Fig. 1b, this analysis perturbs each descriptor and measures the resulting changes in the model response. For descriptor j, the sensitivity score is computed as

$$s_j(\mathbf{c}) = f(\mathbf{c} + \delta \mathbf{e}_j) - f(\mathbf{c}),$$

Here, $f(\cdot)$ denotes the diagnostic output, δ is a small perturbation, and e_j is the unit vector for descriptor j. This yields descriptor-level influence scores in a clinically named concept space, rather than in opaque latent features.

Visual and Descriptor-Level Explanation: Fig. 1b shows that MorphX integrates visual, semantic, contrastive, and uncertainty-aware explanations within the same inference pipeline. Grad-CAM++ was used to localise image regions that contributed to the diagnostic output, and anatomical masking was applied to constrain attention to biologically meaningful regions, such as the nuclei and cytoplasm. The Grad-CAM++ map is computed as

$$L^c_{\text{GradCAM++}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right),$$

Where A^k denotes the k-th feature map and α_k^c denotes its class-specific weight. In parallel, the differentiable WHO layer decomposes the scalar risk score into descriptor-level contributions, enabling the final prediction to be inspected in terms of the clinically named features.

Semantically Constrained Counterfactuals: As indicated in Fig. 1b, MorphX generates counterfactual explanations in the WHO concept space. Let c denote the original concept vector and c' the counterfactual vector. The corresponding counterfactual risk score is

$$r(\mathbf{c}') = \frac{\sum_{j=1}^{28} w_j c'_j}{\sum_{j=1}^{28} w_j}.$$

The counterfactual objective is to find the smallest concept-level perturbation that changes the predicted diagnosis.

$$\mathbf{c}^* = \arg\min_{\mathbf{c}'} \|\mathbf{c}' - \mathbf{c}\|_1 \quad \text{subject to} \quad \hat{y}(\mathbf{c}') \neq \hat{y}(\mathbf{c}).$$

This yields sparse and semantically grounded what-if explanations that show how limited changes in WHO descriptors can alter predicted outcomes.

Uncertainty Estimation: Finally, Fig. 1b shows that uncertainty is estimated using the Monte Carlo dropout and predictive entropy over the WHO concept vector. Samples with high uncertainty were flagged for manual review to ensure that low-confidence predictions were not treated as definitive.

Overall, MorphX links prediction, concept reasoning, risk estimation, and explanation within a single, concept-centred architecture. Fig. 1a emphasizes the end-to-end diagnostic pipeline, whereas Fig. 1b details the internal inference and explanation flow that supports interpretability and auditability.

III Results and Discussions

Table II summarizes the main MorphX results across classification, explanation, and robustness. Overall, MorphX achieved strong performance on ALL-IDB2 while preserving a clinically structured explanation pathway through the WHO concept prediction, weighted risk mapping, and integrated explanation modules. These results suggest that WHO-aligned concept supervision can improve interpretability without materially reducing the diagnostic performance.

Table II. Summary of predictive, explanatory, and robustness-related results for MorphX

Model or setting	Main Result	Interpretation
Morphx Test Accuracy	94.1%	Strong final classification performance
Morphx ROC-AUC	**>0.97>0.97**	Strong class discrimination
Full_Explainable Validation Accuracy	96.5%	Best-performing ablation result
Baseline Validation Accuracy	95.8%	Slightly below the full explainable model
Who_Loss_Weighted Validation Accuracy	94.7%	Competitive, but below the full model
Who_Loss_Only Validation Accuracy	90.2%	Lowest classification performance among the ablations
Explanation Overhead	about 0.8 s per image	Added computational cost of integrated explainability
Counterfactual Sparsity	7.2 of 28 descriptors changed on average	Decision boundary is relatively compact
Mean Counterfactual Change	0.361, SD 0.243	Moderate perturbation in concept space
Saliency Sparsity	0.96	Visual attention is highly concentrated
Saliency Map Structural Similarity	about 0.04	Visual explanations are unstable across cases
Saliency-Attention Correlation	about 0.12	Only modest agreement between saliency and attention

MorphX achieved a classification accuracy of 94.1 percent on ALL-IDB2. Fig. 2 shows the classification accuracy curve, and Fig. 3 shows the corresponding loss curve. Both plots indicate stable optimisation and smooth convergence, supporting the regularising role of the concept-guided supervision. Fig. 4 shows the confusion matrix, which contains relatively few false positives and false negatives, indicating a clear separation between the lymphoblast and normal cell classes. Fig. 5 shows the ROC curve, with an AUC greater than 0.97, confirming a strong discriminative performance in the present dataset setting.

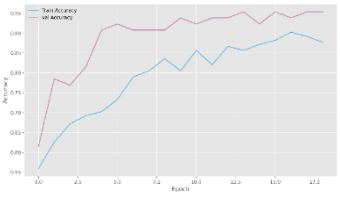

Figure-2: Classification Accuracy Curve

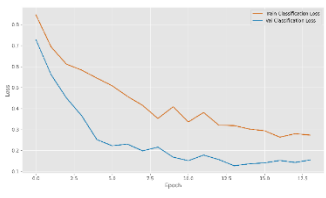

Figure-3: Classification Loss Curve

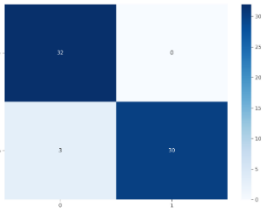

Figure-4: Classification Confusion Matrix

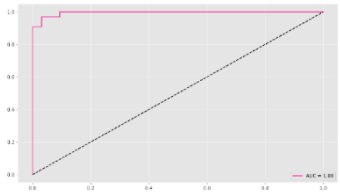

Figure-5: Classification ROC Curve

MorphX also maintains a light computational profile. The model required 3.71 GMac FLOPs and 630.32 thousand parameters with an average inference latency of 4.89 ms. The main additional cost arose from the integrated explanation stack, which required approximately 0.8 s per image on an available GPU. This indicates that the classification core remained efficient, whereas explanation generation accounted for most of the added latency.

The WHO concept module produced meaningful intermediate predictions for 28 descriptors. Fig. 6 shows the relationship between the predicted and reference descriptor values. Although the predictions clustered rather than following a perfect linear trend, they captured the relevant morphological structure under a limited-data setting. Fig. 7 shows that the WHO feature loss converged stably, indicating that the concept head remained trainable despite the descriptor-level variability. Together, these findings support the feasibility of concept-guided learning in the present experimental regime.

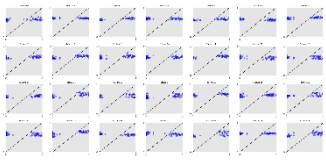

Figure-6: WHO Feature Prediction VS Ground Truth (After Sigmoid)

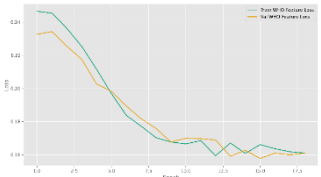

Figure-7: Who Feature Module Loss Curve

The concept-level interpretability was further supported by a sensitivity analysis. Fig. 8 shows the TCAV ranking of the most influential WHO descriptors. High scores for features such as polychromasia and azurophilic granules indicate that MorphX relied on readable morphological cues rather than saliency patterns alone. This is important because it links the model output to named descriptors that are interpretable in hematopathological terms.

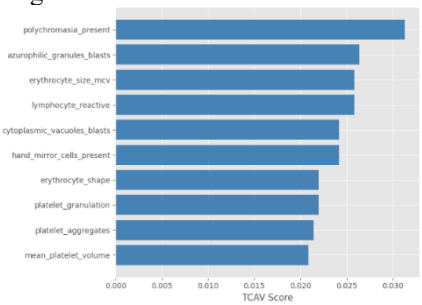

Figure-8: TCAV ranking of top WHO descriptors

Saliency and conceptual explanation quality: Saliency analysis showed high sparsity (mean = 0.96), indicating concentrated attention on the selected regions. The low structural similarity index measure (SSIM; mean ≈ 0.04) suggests structural variability in the saliency maps, which remains a limitation of gradient-based methods. A moderate correlation between saliency and attention (≈0.12) supported the use of an anatomically constrained overlay to reduce noise.

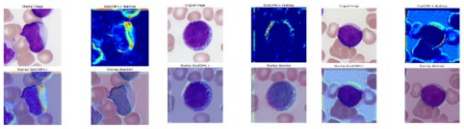

Figure-9: GradCAM++ and attention-based visual explanation

The differentiable WHO layer produced a transparent scalar risk score by aggregating the predicted descriptors using clinically defined weights. Fig. 10 shows the descriptor-level contributions to this risk score, making the final output inspectable in clinically meaningful terms. In practice, this risk mapping was integrated with local visual overlays, counterfactual summaries, and uncertainty indicators to produce a structured, clinician-facing report.

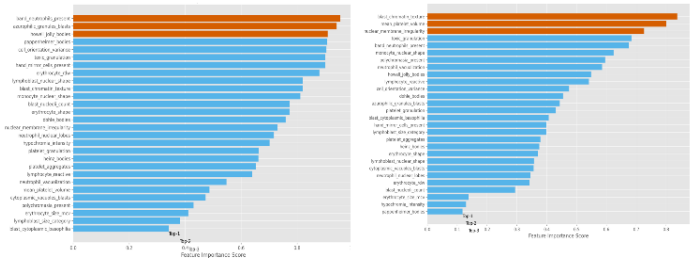

Figure-10: Feature Importance Ranking Based on Model Interpretability

Counterfactual morphological perturbation analysis further supported the interpretability of the framework. Fig. 11 shows the comparison between the original and counterfactual WHO descriptor values. On average, only 7.2 of the 28 descriptors required perturbation to reverse the predicted diagnosis, with a mean absolute change of 0.361 and a standard deviation of 0.243, indicating sparse and quantitatively compact counterfactuals, respectively. The largest changes were observed for descriptors such as hand_mirror_cells_present, cell_orientation_variance, and mean_platelet_volume, whereas erythrocyte_rdw and toxic_granulation changed minimally. These findings support the plausibility of counterfactuals and suggest that class reversal depends on a semantically interpretable subset of morphological factors.

Feature	Original	Counterfactual	Δ Change
erythrocyte_size_mcv	0.624	0.418	-0.206
erythrocyte_rdw	0.652	0.656	+0.004
erythrocyte_shape	0.705	0.288	-0.417
hypochromia_intensity	0.574	0.410	-0.164
polychromasia_present	0.628	0.390	-0.238
howell_jolly_bodies	0.166	0.587	+0.421
pappenheimer_bodies	0.683	0.448	-0.235
heinz_bodies	0.727	0.229	-0.497
neutrophil_nuclear_lobes	0.811	0.376	-0.435
band_neutrophils_present	0.811	0.341	-0.471
toxic_granulation	0.605	0.650	+0.045
dohle_bodies	0.859	0.587	-0.271
neutrophil_vacuolization	0.168	0.663	+0.495
monocyte_nuclear_shape	0.302	0.357	+0.055
lymphocyte_reactive	0.487	0.224	-0.263
mean_platelet_volume	0.972	0.283	-0.689
platelet_granulation	0.803	0.570	-0.233
platelet_aggregates	0.556	0.124	-0.432
lymphoblast_size_category	0.906	0.360	-0.546
lymphoblast_nuclear_shape	0.563	0.648	+0.085
blast_chromatin_texture	0.960	0.538	-0.422
blast_nucleoli_count	0.187	0.394	+0.207
blast_cytoplasmic_basophilia	0.989	0.154	-0.835
azurophilic_granules_blasts	0.184	0.769	+0.586
cytoplasmic_vacuoles_blasts	0.380	0.556	+0.175
hand_mirror_cells_present	0.980	0.217	-0.763
nuclear_membrane_irregularity	0.559	0.617	+0.058
cell_orientation_variance	0.858	0.099	-0.760

Figure 11: Original and counterfactual WHO descriptor scores

The uncertainty was estimated using the Monte Carlo dropout over the WHO concept vector. Low-confidence cases, including borderline morphological patterns, were flagged for manual review rather than being treated as fully reliable. Fig. 12 shows the resulting clinical interpretation report, which

combines the concept predictions, diagnostic classification, weighted risk score, visual overlays, counterfactual information, and confidence indicators in a single structured artefact. This report format improves traceability and aligns the model output more closely with clinical review requirements.

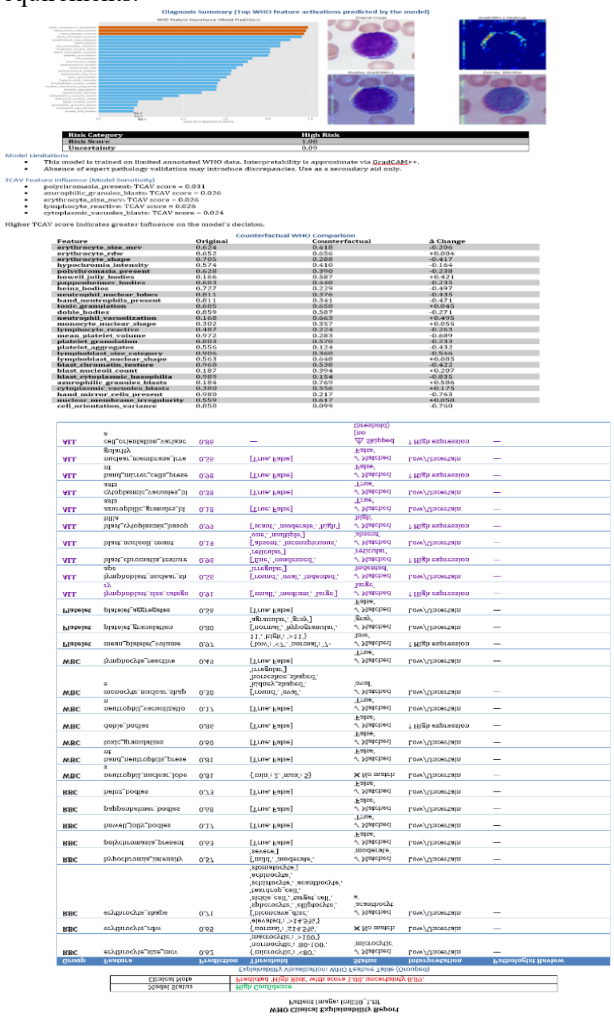

Figure-12: Clinical Interpretation Report Using WHO Feature Attribution

Ablation Studies: The ablation study provides quantitative evidence that concept bottleneck supervision improved classification performance when integrated with the diagnostic objective. The full_explainable configuration achieved the highest validation accuracy at 96.5 percent, compared with 95.8 percent for the baseline, 94.7 percent for who_loss_weighted, and 90.2 percent for who_loss_only. These results show that concept supervision alone is insufficient, but that joint optimization of class prediction and WHO descriptor learning yields the best overall tradeoff between predictive accuracy and interpretability. The gain over the baseline, although modest, indicates that clinically structured supervision can regularize learning without sacrificing diagnostic utility.

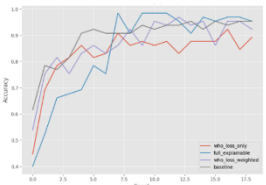

Figure-13: Ablation Study: Validation Accuracy

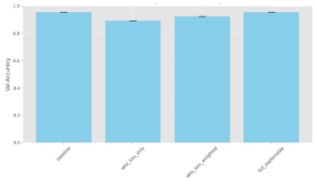

Figure-14: Ablation Study: Validation Accuracy Comparison

Figs. 14-18 summarize the corresponding loss trends. The baseline and full_explainable models converged more smoothly, whereas the who_loss_only model showed higher variance and slower convergence. The who_loss_only model achieved the lowest WHO-specific validation loss, which was consistent with its exclusive optimisation target. However, full_explainable and who_loss_weighted maintained comparably low WHO losses while preserving a stronger classification performance. Overall, the ablation results show that balanced concept supervision improves interpretability without undermining the diagnostic utility.

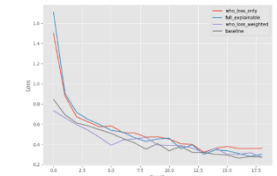

Figure-15: Ablation Study - Training Loss Comparison

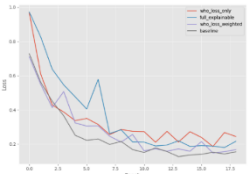

Figure-16: Ablation Study – Validation Loss Comparison

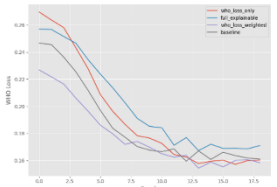

Figure-17: Ablation Study - WHO Training Loss Comparison

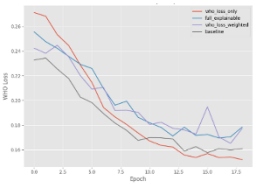

Figure-18: Ablation Study - WHO Validation Loss Comparison

Compared with recent attribute-aware and concept-based methods, MorphX remains competitive while offering a more integrated ante-hoc explanation pathway. For related haematologic imaging tasks, [18] reported 91.28±2.17%91.28±2.17% accuracy, [22] reported 97.85±0.19%97.85±0.19% accuracy on C_NMC_2019, and [21] achieved an AUROC of 0.97 on real-world peripheral smear data. Although these are not strictly like-for-like comparisons, they show that MorphX preserves strong predictive performance while adding structured interpretability to the model.

IV Discussion

MorphX demonstrates how ante-hoc concept modelling can improve the transparency of deep learning in haematological image analysis. The framework does not treat explanations as post hoc add-ons. Instead, it builds a diagnostic prediction around a WHO-aligned concept bottleneck, a clinically weighted risk mapping layer, and an integrated explanation pipeline. This design is important because it expresses model reasoning in terms that are closer to hematopathological assessment than those of a conventional black-box classifier.

The results show that this added structure did not compromise practical predictive performance in the present setting. MorphX maintained 94.1 percent classification

accuracy and an AUC above 0.97 on ALL-IDB2, while also producing concept predictions, visual overlays, risk scores, uncertainty indicators, and counterfactual summaries. Therefore, the value of the framework lies not only in classification accuracy but also in making the full diagnostic pathway more inspectable and auditable.

The interpretability evidence was consistent across several analysis levels. TCAV identified pathology-relevant descriptors as important drivers of the final decision-making process. Saliency overlays localise attention to morphologically meaningful regions. Counterfactual analysis showed that diagnosis reversal usually requires changes in only a limited subset of WHO descriptors. Taken together, these findings suggest that MorphX operates within a compact and semantically meaningful concept space rather than relying solely on opaque internal features.

However, the present evidence should not be overstated. First, the study did not include a fully matched numeric benchmark against a separate post hoc explainer; therefore, claims of superiority are stronger at the conceptual and methodological levels than at strict like-for-like performance comparisons. Second, gradient-based saliency remained structurally noisy, as reflected by the low SSIM values, particularly in the rare morphological settings. Third, concept prediction was less stable for underrepresented phenotypes, likely because such patterns were sparse in the training data. Fourth, the explanation pipeline introduced a measurable per-image latency, which may limit the throughput unless further optimised. Finally, the results were obtained from ALL-IDB2 and should therefore be interpreted as proof of the method rather than evidence of deployment readiness.

Overall, MorphX provides a useful framework for interpretable AI in hematopathology by linking predictions, explanations, and clinical traceability within a single architecture. However, robustness remains weaker for rare or underrepresented morphological patterns, where concept prediction is less reliable because of limited training coverage. Accordingly, broader clinical use will require stronger expert annotation, targeted sampling of rare morphologies, larger multi-institutional cohorts, and prospective and external validation.

IV Conclusion

MorphX presents an ante-hoc explainable framework for acute lymphoblastic leukaemia diagnosis from peripheral blood smear images, in which diagnostic predictions are mediated through a WHO-aligned morphology concept vector. In the reported experiments, the framework achieved 94.1 percent classification accuracy while generating clinically interpretable outputs, including weighted risk scores, concept sensitivity rankings, localised visual overlays, counterfactual explanations, uncertainty estimates, and a structured report artefact. These results show that diagnostic transparency can be improved without sacrificing strong predictive performance in the present dataset. Therefore, this study should be interpreted as proof of method on ALL-IDB2, with future work directed toward larger cohorts, stronger concept annotation, and prospective clinical validation.

References

[1]. L. Stoicu-Tivadar, “Explaining Deep Learning Models Applied in Histopathology: Current Developments and the Path to Sustainability,” Studies in Health Technology and Informatics, pp. 1003–1007, 2024.

[2]. S. N. Saw, Y. Y. Yan, and K. H. Ng, “Current Status and Future Directions of Explainable Artificial Intelligence in Medical Imaging,” European Journal of Radiology, p. 111884, 2024.

[3]. H. Kondylakis et al., “A Review of Methods for Trustworthy AI in Medical Imaging: The FUTURE-AI Guidelines,” IEEE Journal of Biomedical and Health Informatics, pp. 2299–2315, 2025.

[4]. S. Kinger and V. Kulkarni, “A Review of Explainable AI in Medical Imaging: Implications and Applications,” International Journal of Computers and Applications, pp. 983–997, 2024.

[5]. A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, “Causability and Explainability of Artificial Intelligence in Medicine,” WIREs Data Mining and Knowledge Discovery, vol. 9, no. 4, 2019.

[6]. R. R. Selvaraju, D. Parikh, A. Das, D. Batra, R. Vedantam, and M. Cogswell, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” in Proc. IEEE Int. Conf. Computer Vision, 2017, pp. 618–626.

[7]. C. Rudin, “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead,” Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.

[8]. K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, and F. Nensa, “Explainable AI in Medical Imaging: An Overview for Clinical Practitioners – Beyond Saliency-Based XAI Approaches,” European Journal of Radiology, p. 110786, 2023.

[9]. R. Caruana, P. Koch, J. Gehrke, M. Sturm, N. Elhadad, and Y. Lou, “Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission,” in Proc. 21th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2015.

[10]. W. Pang et al., “Integrating Clinical Knowledge into Concept Bottleneck Models,” in Proc. Int. Conf. Medical Image Computing and Computer-Assisted Intervention, 2024, pp. 243–253.

[11]. A. De Santis et al., “Visual-TCAV: Concept-Based Attribution and Saliency Maps for Post-hoc Explainability in Image Classification,” arXiv preprint arXiv, 2024.

[12]. A. Yan et al., “Robust and Interpretable Medical Image Classifiers via Concept Bottleneck Models,” arXiv preprint arXiv, 2023.

[13]. P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, “Concept Bottleneck Models,” in Proc. 37th Int. Conf. Machine Learning, vol. 119, pp. 5338–5348, 2020.

[14]. M. Raj, M. D. Sekar, and P. Manivannan, “Role of Morphology in the Diagnosis of Acute Leukemias: Systematic Review,” Indian Journal of Medical and Paediatric Oncology, vol. 44, no. 5, pp. 464–473, 2023.

[15]. B. Kim, C. Cai, F. Viégas, J. Gilmer, J. Wexler, M. Wattenberg, and R. Sayres, “Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors,” arXiv preprint arXiv:1711.11279, 2018.

[16]. L. Rayappan and R. Parameswari, “ALL-MorphNet: A Novel Hybrid Deep Learning Architecture for Automated Acute Lymphoblastic Leukemia Diagnosis in Microscopic Images,” in Proc. Int. Conf. Intelligent Computing and Communication Technologies, pp. 915–921, 2025.

[17]. R. D. Labati, V. Piuri, and F. Scotti, “ALL-IDB: The Acute Lymphoblastic Leukemia Image Database for Image Processing,” in Proc. IEEE Int. Conf. Image Processing, Sep. 2011.

[18]. S. Tsutsui, W. Pang, and B. Wen, “WBCAtt: A White Blood Cell Dataset Annotated with Detailed Morphological Attributes,” arXiv preprint arXiv, 2023.

[19]. A. S. Pal et al., “Pathologist-Like Explanations Unveiled: An Explainable Deep Learning System for White Blood Cell Classification,” in Proc. IEEE Int. Symp. Biomedical Imaging, 2024, pp. 1–5.

[20]. A. Rehman et al., “A Large-Scale Multi-Domain Leukemia Dataset for White Blood Cells Detection With Morphological Attributes for Explainability,” arXiv preprint arXiv, 2024.

[21]. M. F. Dasdelen et al., “Transformer-Based Hematological Malignancy Prediction From Peripheral Blood Smears in a Real-World Cohort,” 2025.

[22]. A. Genovese et al., “DL4ALL: Multi-Task Cross-Dataset Transfer Learning for Acute Lymphoblastic Leukemia Detection,” IEEE Access, vol. 11, pp. 65222–65237, 2023.