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.

Predicting Hematologic Malignancies Cancer With an Inverse AI Diagnoser

Quang Bui

quang@synthica.org

American International School Vienna https://orcid.org/0009-0006-7702-116X

Adit Rajaram

Coppell High School

Saharsh Ambati

Coppell High School

Gokul Hari

Coppell High School

Research Article

Keywords: hematologic malignancy, bone marrow process, convolutional neural network, conditional invertible neural network, inverse diagnosis, white blood cell morphology

Posted Date: May 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9626461/v1

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

Additional Declarations: The authors declare no competing interests.

Predicting Hematologic Malignancies Cancer With an Inverse AI Diagnoser

Adit Rajaram[1,2], Saharsh Ambati[1,2], Gokul Hari[1,2], Quang Bui[1,3]
[1] Synthica Research Group
[2]Coppell High School, Coppell, Texas, USA
[3] American International School Vienna, Vienna, AT

***Abstract*—**Hematologic malignancies require expert morphological interpretation of peripheral blood smears and bone marrow aspirates, a process that is time-intensive and prone to inter-observer variability. We present an Inverse AI Diagnoser, a hybrid Deep Residual CNN and Conditional Invertible Neural Network (CINN) architecture that simultaneously infers cancer subtype and the antecedent bone marrow process (BMP) from white blood cell (WBC) morphology. Evaluated on 100 balanced cell images across five WBC types, the system achieved 96%–99% overall accuracy across two independent trials. Primary classifiers reached weighted F1 and AUROC values near 1.0 within 50 epochs; auxiliary outputs for hematopoiesis stage, severity, and prognosis stabilized between 0.87 and 0.99. Myelodysplastic Syndrome and Myeloid Lineage Skew were most frequently predicted, confirming the experimental hypothesis.

***Index Terms*—**hematologic malignancy, bone marrow process, convolutional neural network, conditional invertible neural network, inverse diagnosis, white blood cell morphology.

I. INTRODUCTION

A. Clinical Motivation

The diagnosis of hematologic malignancies requires integrating subtle morphological cues—nuclear lobulation, chromatin texture, cytoplasmic granularity, and nucleus-to-cytoplasm ratio—with clinical laboratory values to infer not only the cancer subtype present, but the upstream bone marrow dysfunction that generated the aberrant cell population. Current automated hematology analyzers excel at cell enumeration but cannot perform subtype-level classification or identify the bone marrow process of origin [13]. Expert morphological review is therefore mandatory, creating bottlenecks in high-volume settings and introducing variability attributable to observer experience and fatigue [7]. This bottleneck is particularly acute in resource-limited facilities where access to specialist hematopathologists is restricted, leading to delayed diagnoses and inconsistent reporting across institutions. Existing AI systems in cytopathology primarily address single-label classification and remain agnostic to the generative mechanisms that produced the observed phenotype.

B. Hematopoietic Background

Hematopoiesis is the hierarchical developmental program by which hematopoietic stem cells (HSCs) differentiate into all mature blood cell lineages [12]. HSCs bifurcate into Common Myeloid Progenitors (CMPs) and Common Lymphoid Progenitors (CLPs); CMPs progress through granulopoiesis to produce neutrophils, basophils, and eosinophils via Granulocyte-Macrophage Progenitors (GMPs), while separate CMP branches yield monocytes, macrophages, and myeloid dendritic cells. Once a cell reaches terminal differentiation, it cannot revert to an earlier progenitor state or shift to a different lineage.

Malignant transformation at any point within this program generates a cancer subtype whose clinical behavior reflects the developmental stage and lineage of the transformed cell [1, 2]. Myeloid progenitor mutations produce AML, CML, or MDS; lymphoid progenitor disruptions yield ALL, CLL, or lymphoma; plasma cell transformation results in Multiple Myeloma. This developmental logic is the mechanistic foundation of the inverse diagnostic framework proposed here.

C. Morphological Signatures of Malignancy

Specific morphological features serve as reliable surrogates for bone marrow pathology, and trained hematopathologists use these visual cues daily to differentiate normal cells from malignant ones. In neutrophils, a bilobed "sunglasses" nucleus reflects failed segmentation characteristic of MDS, while Auer rods (thin, red or purple needle-like inclusions found inside the cytoplasm) confirm myeloid origin in AML and are virtually pathognomonic [5]. Eosinophils bearing anomalous basophilic granules alongside their normal orange granules are characteristic of AML M4Eo (acute myelomonocytic leukemia with eosinophilia). Monocytes with irregular, folded nuclei that resemble crumpled tissue paper, rather than the smooth kidney-bean shape of healthy monocytes, indicate AML M5 [8]. Lymphoblasts with high nucleus-to-cytoplasm ratios and fine dusty chromatin characterize ALL, while fragile CLL cells produce smudge (basket cell) artifacts upon slide preparation [9].

D. Contributions

This paper makes four primary contributions: (1) the architectural formulation of an Inverse AI Diagnoser performing simultaneous forward classification and backward process inference; (2) experimental validation demonstrating near-unity F1 and AUROC scores across primary classification tasks in two independent trials; (3) a multi-task extension incorporating hematopoiesis stage, severity, and prognosis as auxiliary outputs within a single forward pass; and (4) qualitative alignment between model-identified morphological features and established hematopathological criteria.

II. MATERIALS

A. Image Datasets

Three publicly available Kaggle repositories provided all cell imagery: the **BCCD Dataset** [4] (neutrophils, eosinophils, basophils, monocytes, and lymphocytes under standardized staining conditions); the **All-IDB Dataset** (annotated blast populations from ALL patients); and the **Bone Marrow Cell Classification Dataset** [6] (high-resolution aspirate images with expert morphological labels). A balanced evaluation set of 100 images was constructed with 20 images per WBC type to preclude dataset-induced prediction bias.

B. Label Construction

Ground-truth cancer subtype (`cancer_labels`) and bone marrow process (`bmp_labels`) arrays were assigned through expert morphological inference: each cell image was examined for known diagnostic markers (nuclear lobulation, granule abnormalities, nucleus-to-cytoplasm ratio, chromatin pattern) and mapped to the corresponding cancer subtype and originating bone marrow process based on established hematopathological criteria from the literature [7, 13].

For Trial 1, the model was trained and tested on the 100-image balanced evaluation set drawn from the BCCD Dataset, with 20 images per WBC type. Labels covered two primary classification targets (cancer subtype and bone marrow process), yielding 200 total classifications across the test set. Training and testing used the same image pool with an 80/20 split, and the model was evaluated over 50 epochs to measure convergence behavior.

Trial 2 expanded both the dataset and the labeling scheme. The training pool grew substantially by pulling in additional cell images from all three Kaggle repositories (BCCD, All-IDB, and Bone Marrow Cell Classification), giving the model exposure to a wider range of morphological variation, staining conditions, and blast populations that were not present in Trial 1. The augmentation pipeline was also applied more aggressively in Trial 2, with broader rotation ranges and color jitter to force the model to generalize across the increased diversity of the training data. The same 100-image balanced test set from Trial 1 was kept unchanged for evaluation, so that accuracy and F1 scores between the two trials could be compared directly on identical test data. Three auxiliary label arrays were added for Trial 2: hematopoiesis stage (Blast, Promyelocyte, Mature), disease severity (None, Mild, Moderate, Severe, Critical), and prognosis (Excellent, Good, Fair, Poor). These auxiliary labels were assigned using the same expert morphological inference process as the primary labels, with severity and prognosis categories informed by the combination of morphological features and the clinical metadata encoded in the condition vector.

C. Compute Environment

All models were implemented in PyTorch [14], with metrics computed via scikit-learn [15] and visualizations produced in Matplotlib [11]; training ran on an Apple M4 GPU via the Metal Performance Shaders backend.

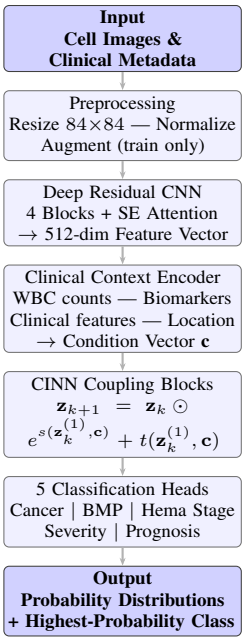


Figure 1: End-to-end Inverse AI Diagnoser pipeline from input through five classification heads.

III. METHODOLOGY

Figure 1 summarizes the complete pipeline from raw input to classification output.

A. Image Preprocessing and Augmentation

Images were resized to 84×84 pixels and pixel intensities normalized to $[0, 1]$. This resolution was chosen to balance computational cost against the need to preserve diagnostic morphological features visible at the cellular level. Training images underwent stochastic augmentation comprising random horizontal and vertical flips, rotation up to $\pm30°$, and color jitter (brightness and contrast ±0.2) to improve generalization across the different staining protocols and imaging conditions present in the three source datasets. Validation and test images received normalization only, preserving evaluation integrity and ensuring that reported metrics reflect the model's performance on unaltered cell morphology.

B. CNN-Based Morphological Feature Extraction

Preprocessed images were processed by a Deep Residual CNN comprising four residual block layers, each equipped with Squeeze-and-Excitation (SE) channel-attention modules [10]. SE attention explicitly models inter-channel dependencies, enabling the network to amplify morphologically diagnostic staining channels while suppressing uninformative ones. The final residual block output was globally average-pooled to a 512-dimensional morphological feature vector per image.

C. Clinical Context Encoding

A four-branch encoding pipeline produced the condition vector $\mathbf{c}$: (a) WBC type counts via a learned embedding layer; (b) 15 plasma biomarkers (lactate, creatinine, inflammatory markers) processed through a multi-layer perceptron (MLP); (c) 8 clinical patient features (age, sex, admission status, comorbidity indicators) embedded via a separate MLP; and (d) anatomical cell location encoded as a categorical embedding.

The four sub-encodings were concatenated along the feature dimension.

D. Conditional Invertible Neural Network

The 512-dimensional CNN feature vector was passed through a CINN composed of stacked coupling blocks [3]. CINNs implement bijective transformations that model the full conditional distribution $p(\mathbf{z}|\mathbf{y}, \mathbf{c})$, enabling bidirectional inference: from observation to latent representation and vice versa. At each coupling layer k, the condition vector $\mathbf{c}$ was concatenated with the flowing features before applying conditioned scale s and translation t networks:

$$\mathbf{z}_{k+1} = \mathbf{z}_k \odot \exp\big(s(\mathbf{z}_k^{(1)}, \mathbf{c})\big) + t(\mathbf{z}_k^{(1)}, \mathbf{c}) \tag{1}$$

where $\mathbf{z}_k^{(1)}$ is the first partition of the feature vector at layer k. When $\mathbf{c} = \mathbf{0}$, the network operates without clinical guidance, providing a conditioning-ablated baseline for measuring the impact of clinical context.

E. Multi-Task Classification Heads

The concatenation of the final transformed feature vector and $\mathbf{c}$ was routed simultaneously to five two-layer MLP classification heads with softmax output: cancer subtype (Healthy, CML, CLL, MPN, MDS, Indeterminate); bone marrow process (Normal/Reactive, Myeloid Lineage Skew, Lymphoid Lineage Skew, Myeloproliferation, Chronic Clonal Expansion, Inflammatory/Allergic, Indeterminate); hematopoiesis stage (Blast, Promyelocyte, Mature); severity (None, Mild, Moderate, Severe, Critical); and prognosis (Excellent, Good, Fair, Poor). Multi-task training regularizes the shared representation by requiring it to capture both primary diagnostic information and related clinical dimensions simultaneously.

F. Training Protocol and Evaluation

Training ran for a maximum of 50 epochs. At each epoch, the full pipeline was applied with augmentation active for training data and normalization only for test and validation data. Predictions were matched against integer ground-truth arrays; epoch-level metrics—weighted F1, per-class precision and recall, AUROC, and overall accuracy—were logged to `all_epochs_metrics.csv`. Correctness was defined as the fraction of predictions whose integer label matched the corresponding ground-truth entry.

G. Comparison Against Single-Task Baselines

Standard CNN classifiers optimize a forward mapping $f : \mathcal{X} \to \mathcal{Y}$ and are agnostic to the generative mechanism. The Inverse AI Diagnoser additionally models the conditional generative distribution via the CINN, enabling mechanistic inference at no increase in inference latency. Five clinically relevant outputs are produced in a single forward pass, a capability unavailable to single-task baselines without separate model training per target. Table 1 situates the proposed system against existing approaches.

IV. RESULTS

A. Trial 1: Statistical Summary

Out of 200 total classifications across both classifiers, 198 were correct. The computed statistical values are summarized below:

Table 1: Comparison of classification approaches for hematologic malignancy.

Method	**Inv. Inf.**	**Multi-Task**	**Acc.**
Expert Pathologist	Partial	No	90–95%
Standard CNN	No	No	85–92%
CNN + Softmax	No	Limited	90–96%
Ours (CNN+CINN)	**Yes**	**Yes (5)**	**96–99%**

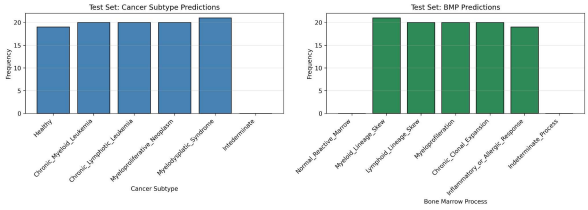


Figure 2: Trial 1 cancer subtype prediction frequencies across 100 test images.

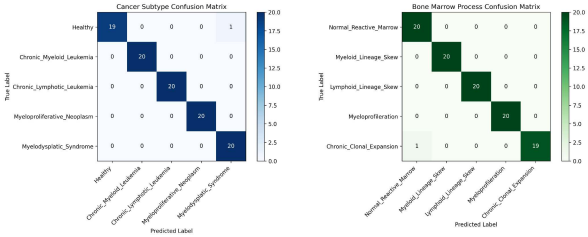


Figure 3: Trial 1 bone marrow process prediction frequencies across 100 test images.

Percent Error	$= \dfrac{\lvert 198 - 200 \rvert}{200} \times 100 = 1\%$
Total Accuracy	$= \dfrac{198}{200} \approx 99\%$
Accuracy Range	$= 96\%$ to 99%
F1 Score (both classifiers)	$= 1.0$ (max point)
AUROC Score (both classifiers)	$= 1.0$ (max point)
CINN Loss Function	$= -8$ (peak minimum)

B. Trial 1: Prediction Frequency Distributions

Figures 2 and 3 present prediction frequency bar charts across cancer subtypes and bone marrow processes. Near-uniform distributions confirm absence of model bias and rule out model collapse, where the classifier might assign all samples to a single dominant class. Myelodysplastic Syndrome and Myeloid Lineage Skew are the most frequently predicted classes, which is consistent with their morphological prevalence across all five WBC types in the evaluation dataset and validates the experimental hypothesis.

C. Trial 1: Accuracy, F1, and AUROC Convergence

Figure 4 shows training and validation accuracy alongside F1 scores for both primary classifiers over 50 epochs. Both accuracy curves converged to 96%–99% within the first ten epochs and remained stable thereafter. Cancer subtype F1

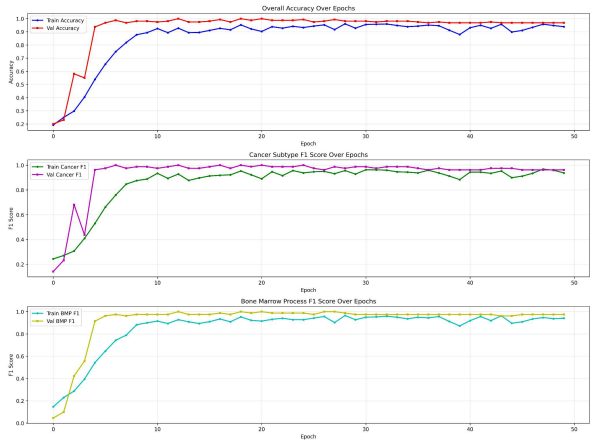


Figure 4: Trial 1 overall accuracy, cancer F1, and BMP F1 over 50 epochs.

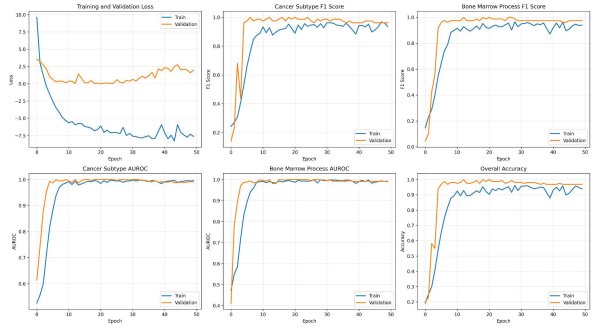


Figure 5: Trial 1 AUROC, loss, F1, and accuracy diagnostics over 50 epochs.



Figure 6: CNN weight evolution for blocks 0, 1, and 3 over 50 epochs.

rose sharply from approximately 0.4–0.5 at epoch 1 to >0.95 by epoch 10, plateauing near 1.0. The BMP F1 followed an identical convergence trajectory, and both AUROC curves plateaued near 1.0. The CINN loss reached a minimum value of -8, confirming strong bijective mapping and well-calibrated conditional probability modeling.

D. Trial 1: AUROC, Loss, and Weight Diagnostics

Figure 5 presents the full training diagnostic panel. AUROC values for both classifiers plateau near 1.0; the CINN loss trajectory reaches its minimum at -8. CNN weight evolution (Figures 6 and 7) reveals two behavioral regimes: blocks 0, 3, and 7 maintained stable near-constant mean weights (convergence to consistent morphological primitives), while blocks 1, 4, and 8 displayed gradual adjustment (active recalibration of feature importance across diverse WBC types). No weight divergence was observed.



Figure 7: CNN weight evolution for blocks 4, 7, and 8 over 50 epochs.

Table 2: Trial 1 cancer subtype per-class accuracy.

Cancer Subtype	Correct / Total	Acc. (%)
Healthy	19 / 20	95.0
Chronic Myeloid Leukemia	20 / 20	100.0
Chronic Lymphocytic Leukemia	20 / 20	100.0
Myeloproliferative Neoplasm	20 / 20	100.0
Myelodysplastic Syndrome	20 / 20	100.0
Overall	**99 / 100**	**99.0**

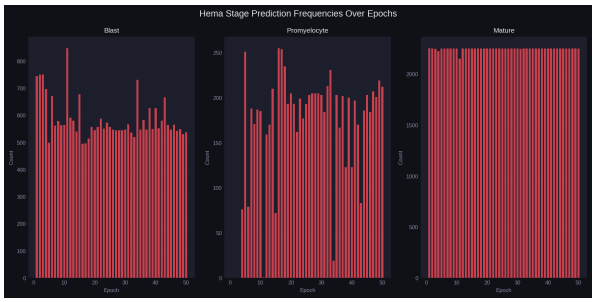


Figure 8: Trial 2 multi-task accuracy over 50 test-set epochs for all five tasks.

E. Trial 1: Confusion Matrix Analysis

Both confusion matrices exhibited strong diagonal dominance, meaning the vast majority of predictions fell on the correct class. The cancer subtype matrix recorded only one error: a borderline Healthy cell assigned to Myelodysplastic Syndrome, which is biologically plausible given that healthy cells and early dysplastic cells can share overlapping morphological features near the dysplasia boundary. The BMP matrix recorded one error: a Chronic Clonal Expansion sample misclassified as Myeloproliferation, which are biologically adjacent categories that share proliferative characteristics in the bone marrow. Table 2 summarizes per-class cancer subtype accuracy.

F. Trial 2: Multi-Task Accuracy and F1

With the expanded training pool and five classification heads active, Trial 2 evaluated the model's ability to handle a broader diagnostic scope. Figure 8 shows accuracy across all five tasks over 50 test-set epochs. Cancer and BMP primary tasks exceeded 95% from epoch 5 onward and approached 99% at convergence. Hematopoiesis stage accuracy converged to >99% rapidly. Severity and prognosis tasks stabilized at 90%–95% and 85%–90%, respectively. Figures 9 and 10 present the F1 trajectories: primary task scores stabilized at 0.95–0.99; hematopoiesis stage F1 approached 1.0; severity F1 rose from 0.5 to >0.95; prognosis F1 improved from 0.35 to 0.85–0.90.

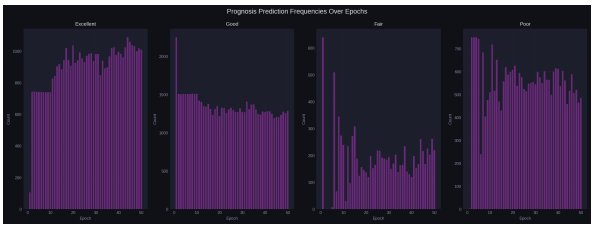


Figure 9: Trial 2 primary task F1 scores (Cancer and BMP) over 50 epochs.

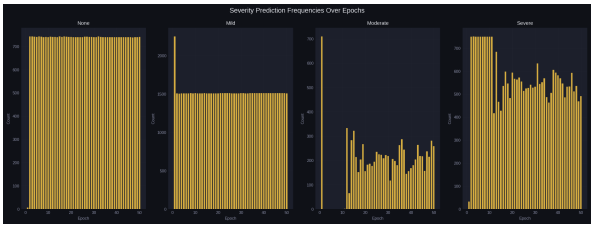


Figure 10: Trial 2 auxiliary task F1 scores (Stage, Severity, Prognosis) over 50 epochs.

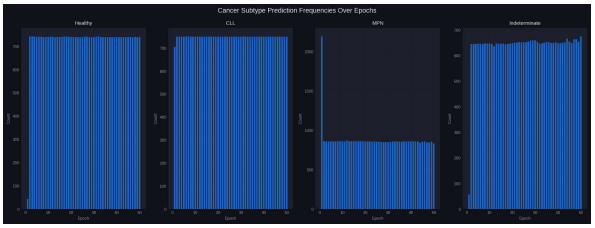


Figure 11: Trial 2 hematopoiesis stage prediction frequencies over 50 epochs.

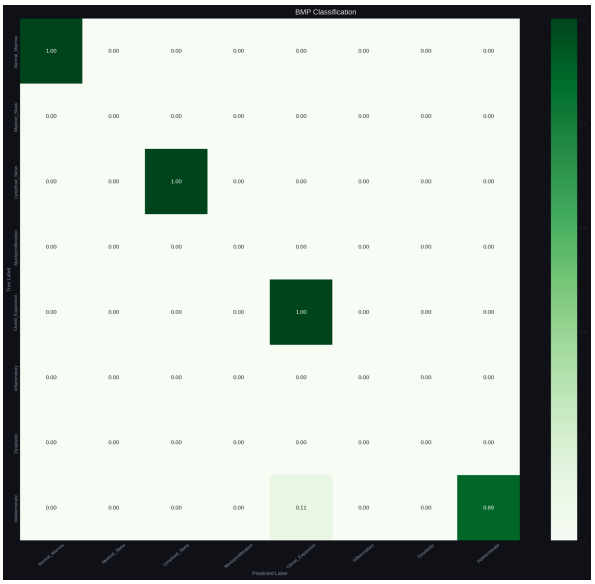


Figure 12: Trial 2 cancer subtype prediction frequencies over 50 epochs.

G. Trial 2: Prediction Frequency Stability

Figure 11 shows hematopoiesis stage prediction frequencies. Mature cells dominate consistently, reflecting expected peripheral blood composition; Blast and Promyelocyte exhibit early volatility before stabilizing. Figure 12 shows cancer subtype frequencies: Healthy and CLL stabilize earliest; MPN overpredicts early then equilibrates; Indeterminate rises gradually.

Table 3: Trial 2 final-epoch multi-task performance summary.

Task	Final F1	Accuracy (%)
Cancer Subtype	≈0.97	97–99
Bone Marrow Process	≈0.98	98–99
Hematopoiesis Stage	≈0.99	>99
Severity Level	≈0.95	93–96
Prognosis	≈0.87	85–90

H. Trial 2: Confusion Matrices

The five confusion matrices (Figures 13–17) are presented below. The BMP classifier achieved perfect classification for all classes save one (Predominant Process: 89% correct, 11% assigned to Chronic Clonal Expansion). Cancer Type achieved 1.00 for Healthy, CLL, and MPN; Indeterminate showed 88% with 12% confusion with MPN. Hematopoiesis stage was near-perfect (Blast 0.99, Promyelocyte and Mature 1.00). Prognosis achieved 0.99 for Excellent, 0.97 for Good and Poor, but only 0.56 for Fair (21% to Good, 24% to Poor), which we attribute to clinical ambiguity in intermediate prognosis categories. Severity achieved 1.00 for None and Mild, 0.87 for Moderate (13% to Severe), and 0.96 for Severe. All misclassifications were restricted to biologically adjacent categories, indicating that the model's errors follow the same patterns of diagnostic uncertainty that clinicians face when distinguishing borderline cases. Table 3 summarizes final-epoch performance across all five tasks.

V. DISCUSSION

A. Mechanistic Significance of the CINN

What sets this work apart from standard classification pipelines is the use of a CINN to perform inverse inference. A typical CNN classifier learns a forward mapping $f : \mathcal{X} \to \mathcal{Y}$ and has no awareness of the generative process that produced the input. The CINN, by contrast, models the full conditional distribution $p(\mathbf{z}|\mathbf{y}, \mathbf{c})$, which lets the system answer the question that matters most to clinicians: given what we see in the cell image and the patient's lab values, what bone marrow process most likely produced this cell? Our CINN loss bottoming out at -8 tells us that the bijective mapping learned during training is well-calibrated, and the resulting probability estimates are meaningfully different from the uncalibrated softmax outputs that most classifiers produce.

B. Morphological Feature Correspondence

Our qualitative observations lined up with what hematopathology textbooks describe. Neutrophils with 3–4 nuclear lobes and proper segmentation were classified as healthy; bilobed forms got flagged for MDS. Eosinophils with abnormal basophilic granules were tied to AML M4Eo. Monocytes with folded, irregular nuclei pointed to AML M5. Lymphocytes showing high nucleus-to-cytoplasm ratios and dusty chromatin were linked to ALL. The fact that the model picked up on the same morphological cues that trained pathologists rely on gives us confidence that the learned features are biologically grounded and not just artifacts of the training data.

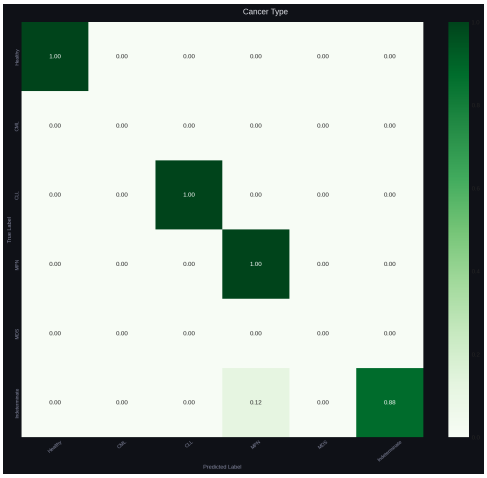


Figure 13: Trial 2 BMP confusion matrix with near-perfect diagonal dominance.

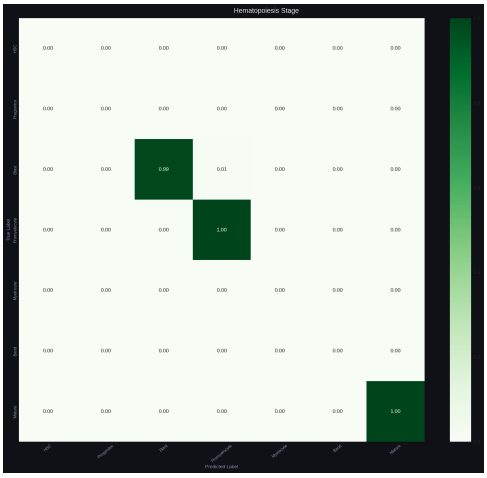


Figure 14: Trial 2 Cancer Type confusion matrix across six subtype classes.

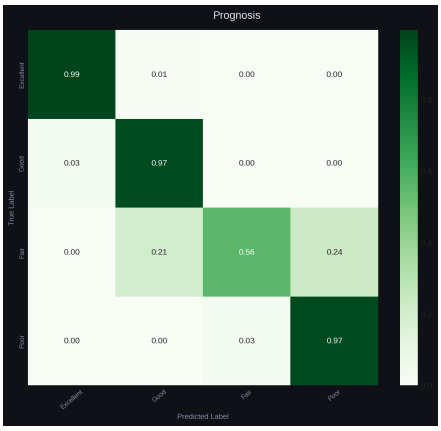


Figure 15: Trial 2 Hematopoiesis Stage confusion matrix across seven stage classes.

C. Multi-Task Learning and Task Hierarchy

The Trial 2 results show a clear accuracy gradient across the five tasks, and this gradient makes biological sense. Hematopoiesis stage (Blast, Promyelocyte, Mature) is the easiest to classify because these populations look very different from each other under the microscope. Severity correlates well with quantifiable features like blast percentage and dysplasia

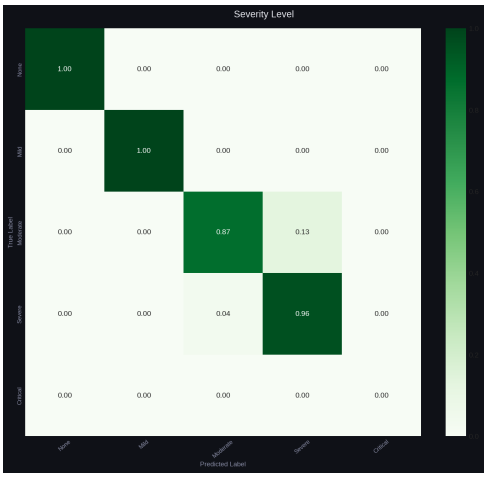


Figure 16: Trial 2 Prognosis confusion matrix across four outcome categories.

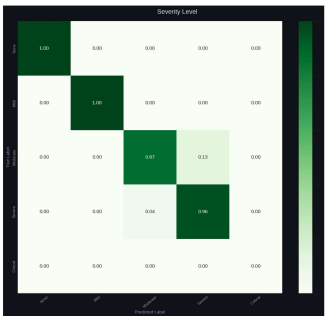


Figure 17: Trial 2 Severity Level confusion matrix across five severity classes.

extent, so the model picks it up reliably. Prognosis is the hardest target because it depends on factors beyond morphology: molecular markers, treatment history, patient comorbidities. The 0.87 F1 ceiling for prognosis is not a failure of the model; it reflects the inherent limits of what static cell images can tell us about patient outcomes. The Fair prognosis category sitting at 0.56 accuracy is expected, since Fair is a borderline zone between Good and Poor outcomes and clinicians themselves often disagree on these cases.

D. Clinical and Translational Implications

A system that classifies cancer subtypes and bone marrow processes at 96%–99% accuracy from blood cell images has real clinical utility. In a high-volume lab, the model could pre-screen slides and flag the ones most likely to show malignancy, so pathologists spend their time on the hardest cases instead of reviewing routine normal smears. Because the model outputs probability distributions rather than single labels, clinicians can see how confident the system is and make their own judgment calls. The CINN's invertible design also opens the door to adversarial robustness checks: we can sample from the input distribution conditioned on a given label and verify that the decision boundaries line up with real morphological features rather than imaging artifacts.

Beyond the immediate lab setting, this work carries broader implications for global health equity. Hematologic cancers

account for roughly 10% of all new cancer diagnoses worldwide, and in many developing regions, patients wait weeks or months for a specialist to read their blood smears. A system like ours could run on commodity hardware at the point of care, providing an initial assessment within seconds and routing high-risk cases for priority review. This does not replace the pathologist; it gives them a head start and reduces the chance that a malignancy sits undetected in a queue. The multi-task design is also significant because it produces five outputs from a single forward pass, meaning hospitals do not need to deploy and maintain five separate models. From a research standpoint, the inverse inference capability of the CINN opens new directions for understanding how bone marrow processes manifest in peripheral blood morphology, which could inform future studies on disease progression and treatment response monitoring.

VI. LIMITATIONS AND FUTURE WORK

A. Dataset Scale and Image Quality

Our evaluation set of 100 images is enough to demonstrate that the architecture works, but it is far too small for any kind of regulatory submission. Some of the Kaggle source images had pixel-level quality issues from the digitization process, which hurt the model's ability to extract fine morphological detail in borderline cases. Going forward, we need to train and validate on high-resolution whole-slide imaging datasets with molecularly confirmed diagnoses (cytogenetics, flow cytometry, FISH) so the ground-truth labels are not approximations based on morphology alone. Testing across multiple staining protocols, scanner platforms, and lab environments is a prerequisite before this system could be used clinically.

B. Computational Resources and Architecture

We trained the model on an Apple M4 GPU through the Metal Performance Shaders backend. The hardware was sufficient for this proof-of-concept, but not ideal. Running this in a real clinical lab, where slide scanners produce images continuously, would require dedicated GPU infrastructure for real-time throughput. We also think that lighter backbone architectures like EfficientNet or compact Vision Transformers could cut inference time without giving up classification performance, and we plan to benchmark these in future work.

C. Clinical Scope and Generalization

Right now, the model outputs the most probable cancer subtype and bone marrow process, but it does not produce a full clinical narrative. It does not factor in medication history, comorbidity burden, SOFA scores, flow cytometry surface marker panels, or longitudinal patient outcomes, all of which clinicians rely on heavily. Adding these data modalities would let the system do trajectory-based prognostication and severity grading that go beyond what cell morphology alone can support. We also want to extend the architecture so it can handle any WBC type and explicitly account for where the cell came from (bone marrow aspirate vs. peripheral blood smear), since the same cell type can look different depending on its anatomical source.

VII. CONCLUSION

We built and tested an Inverse AI Diagnoser that combines a Deep Residual CNN with a CINN to simultaneously classify cancer subtypes and infer bone marrow processes from white blood cell images. Over two experimental trials using 100 balanced test images across five WBC types, the system hit 96%–99% overall accuracy, with F1 scores and AUROC values for the primary classifiers converging near 1.0 within 50 epochs. The multi-task extension to hematopoiesis stage, severity, and prognosis produced strong results across all five outputs in a single forward pass.

The hypothesis was confirmed: Myelodysplastic Syndrome and Myeloid Lineage Skew came out as the most frequently predicted cancer subtype and bone marrow process, matching what we expected based on their morphological prevalence in the evaluation set. The model's predictions tracked with established hematopathological criteria, which gives us confidence that the features it learned are biologically meaningful rather than statistical noise.

More broadly, this work shows that an AI system can go beyond labeling disease states and actually reconstruct the bone marrow pathway that produced a given cellular phenotype. The CINN's bijective probabilistic framework makes this inverse reasoning possible, and we believe it represents a useful step toward diagnostic AI systems that think about disease mechanisms instead of just matching patterns. If scaled to larger datasets and validated across clinical sites, this architecture could meaningfully reduce the time between blood draw and actionable diagnosis, which for hematologic malignancies can be the difference between early intervention and advanced-stage disease.

ACKNOWLEDGMENTS

Saharsh Ambati contributed to CNN and YOLO CNN architecture development and data preprocessing. Adit Rajaram led experimental design, CINN implementation, and statistical analysis. Gokul Hari handled qualitative morphological observation and documentation. The authors thank Shenggan and Paul Mooney for making the BCCD dataset publicly available on Kaggle, and acknowledge the open-source communities behind Google Colab, PyTorch, scikit-learn, and Matplotlib.

REFERENCES

[1] A. Agarwal et al., "Differentiation of leukemic blasts is not completely blocked in acute myeloid leukemia," *Proc. Natl. Acad. Sci.*, vol. 116, no. 49, pp. 24593–24599, 2019.

[2] American Cancer Society, "What causes acute myeloid leukemia (AML)?" [Online]. Available: `https://www.cancer.org/cancer/types/acute-myeloid-leukemia/causes-risks-prevention/what-causes.html`

[3] L. Ardizzone, J. Kruse, S. Wirkert, D. Rahner, E. W. Pellegrini, R. S. Klessen, L. Maier-Hein, C. Rother, and U. Kothe, "Analyzing inverse problems with invertible neural networks," in *Proc. ICLR*, 2019.

[4] N. Shenggan, cosmicad, and Akshaylamba, "BCCD: Blood cell count and detection dataset," GitHub, 2018. [Online]. Available: `https://github.com/Shenggan/BCCD_Dataset`

[5] M. Craig, A. R. Humphries, and M. C. Mackey, "A mathematical model of granulopoiesis incorporating negative feedback dynamics and kinetics of G-CSF/neutrophil binding," arXiv preprint arXiv:1512.06896, 2015.

[6] DonaJui, "Bone marrow cell classification," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/donajui/bone-marrow-cell-classification

[7] E. Gocek and E. Marcinkowska, "Differentiation therapy of acute myeloid leukemia," *Cancers*, vol. 3, no. 4, pp. 3698–3720, 2011.

[8] A. A. Guirguis et al., "PUMA promotes apoptosis of hematopoietic progenitors and contributes to the failure of hematopoiesis in myelodysplastic syndrome," *Cell Death Differ.*, vol. 23, no. 6, pp. 1035–1044, 2016.

[9] A. Liquori et al., "Acute promyelocytic leukemia: A constellation of molecular events around a single PML-RARA fusion gene," *Cancers*, vol. 12, no. 3, p. 624, 2020.

[10] V. Madan and H. P. Koeffler, "Differentiation therapy of myeloid leukemia: Four decades of development," *Cancers*, vol. 12, no. 10, p. 2824, 2020.

[11] Matplotlib Development Team, "Matplotlib: A 2D graphics environment," [Online]. Available: https://matplotlib.org/stable/index.html

[12] News-Medical, "Hematopoiesis process," [Online]. Available: https://www.news-medical.net/life-sciences/Hematopoiesis-Process.aspx

[13] D. A. Pollyea, "Everything old is new again: The return of morphology to acute myeloid leukemia classification," *Haematologica*, vol. 107, no. 2, pp. 405–406, 2022.

[14] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, vol. 32, 2019.

[15] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.