National College of Ireland

Enhancing Leukemia Diagnosis with Synthetic Data and Explainable Deep Learning Architectures

MSc Research Project
Data Analytics

Ibrahim Malik
Student ID: x23373385

School of Computing
National College of Ireland

Supervisor: Dr. Bharat Agarwal

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ibrahim Malik
Student ID:	x23373385
Programme:	Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Dr. Bharat Agarwal
Submission Due Date:	11/08/2025
Project Title:	Enhancing Leukemia Diagnosis with Synthetic Data and Explainable Deep Learning Architectures
Word Count:	7302
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th August 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Leukemia Diagnosis with Synthetic Data and Explainable Deep Learning Architectures

Ibrahim Malik
x23373385

Abstract

Leukemia diagnosis through microscopic blood smear analysis remains time-intensive, error-prone, and dependent on expert interpretation. To solve issues of interpretability and data scarcity, this study introduces a first comprehensive framework combining Generative Adversarial Networks (GANs), Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Explainable AI (XAI).

For artificial data augmentation, three GAN variations (DCGAN, WGAN, and cGAN) were used; conditional GANs produced better image quality (FID: 1.1897, SSIM: 0.8869). Using the CNMC dataset, classification showed different architectures responding differently to simulated data. Conventional data quality assumptions were challenged when ViTs showed higher results with WGAN augmentation, whereas CNNs performed ideally with cGAN augmentation. The best ViT-WGAN configuration achieved 74% accuracy, presenting a 167% gain crucial for reducing missed diagnoses, with significant sensitivity improvements from 0.187 to 0.500.

CNN-ViT hybrid architectures are seen to perform worse than individual models, indicating that convolutional and attention methods do not represent features incompatiblely. Explainable AI analysis using Grad-CAM, LIME, and SHAP shows models learned clinically significant morphological characteristics, with focus on cell borders and dispersed properties reflecting pathologist decision patterns.

The framework addresses class imbalance through data generation while providing interpretable predictions essential for adoption. Findings show promise for clinical application, providing pathologists with transparent and precise diagnostic support tackling issues of confidence and performance in medical AI application.

1 Introduction

Leukemia is a type of cancer that leads to malignant proliferation of white blood cells and remains a challenging hematological disorder to diagnose accurately and efficiently. Conventional diagnostic techniques require specialized expertise, are labor intensive, and suffer from inter-observer variability. Given the critical importance of early and precise detection for patient prognosis, there is growing interest in leveraging artificial intelligence (AI) to augment and automate the diagnostic workflow.

Deep learning, particularly Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), has demonstrated remarkable success in medical-imaging tasks. However, two persistent challenges hinder clinical adoption: (1) data scarcity and imbalance in annotated blood-smear datasets, and (2) model interpretability, since black-box predictions

erode clinician trust and complicate regulatory approval. Generative Adversarial Networks (GANs) offer a solution to the first challenge by synthesizing high-fidelity blood-cell images, while Explainable AI (XAI) techniques address the second by visualizing the features driving model decisions.

This thesis presents a hybrid framework that integrates GAN-based data augmentation, a coupled CNN–ViT classification backbone, and intrinsically interpretable modules to deliver both high diagnostic performance and transparent reasoning. We train multiple GAN variants on publicly available datasets to balance class distributions and enrich morphological variability. The augmented data then feed into a custom architecture combining convolutional feature extractors with self-attention blocks to distinguish leukemic from healthy cells. Finally, we embed self-explaining prototypes and attention-map regularization to produce interpretable decision pathways.

Our work directly addresses the central research question: *To what extent can the combined use of Generative Adversarial Networks, Vision Transformers, and Explainable AI improve both the accuracy and interpretability of leukemia classification on microscopic blood-smear images?*

The key contributions of this thesis are threefold:

- Systematic evaluation of GAN variants for generating clinically realistic blood-cell images and mitigating dataset imbalance.
- Implementation of a hybrid CNN–ViT architecture that leverages both local feature hierarchies and global context dependencies.
- Integration of intrinsic interpretability mechanisms that yield transparent explanations without reliance on post-hoc methods.

The remainder of this report is organized as follows: Section 2 reviews related work in AI-driven leukemia detection, generative augmentation, and XAI in medical imaging. Section 3 outlines the methodology involving data preprocessing, GAN training, model architectures, and interpretability mechanisms. Sections 4 and 5 detail on the specifications of design as well as implementation respectively. Section 6 presents experimental results and discusses findings and clinical implications. Finally, Section 7 concludes with achievements and future research directions.

2 Related Work

This chapter surveys state of the art in leukemia diagnostics driven through AI, with a particular focus on (1) deep-learning classification models, (2) generative augmentation techniques, (3) explainability methods, and (4) hybrid architectures that combine these elements. Where relevant, it is highlighted how prior work has informed design choices in the present thesis.

2.1 Deep Learning Architectures in Leukemia Detection

Early efforts in leukemia classification using AI have relied primarily on Convolutional Neural Networks (CNNs), which have shown promise in extracting localized feature hierarchies from blood-smear images. On a blood-smear dataset, for example, Giammarco et al. (2024) showed a CNN ensemble with Class Activation Mapping (CAM) obtaining 94% accuracy, also offering visual explanations via Grad-CAM (Giammarco et al. (2024)).

Recently, Vision Transformers (ViTs) have been applied to hematology imaging, leveraging self-attention to capture global context; Nunna et al. (2024) fused ViT and CNN features in a VISTA model pushing accuracy to 99.96% on ALL-IDB samples (Nunna et al. (2024)). However, pure ViTs often underperform on small medical datasets without extensive augmentation, due to lack of inductive spatial bias.

Vision Transformers (ViTs) employ self-attention to record local and global contextual interactions without a need of convolutional filters, in contrast to CNNs. Almenwer et al. (2024) implemented a ViT-BiLSTM hybrid on genomic datasets for Pleuropulmonary Blastoma (PPB) detection, and went on to achiev superior accuracy and explainability when compared to CNNs, primarily due to the attention mechanism ability of ViTs to link distant cellular features(Almenwer (2024)). However, similar to Nunna et al. (2024), ViT sensitivity to minute morphological changes may be limited due to lack of innate spatial hierarchy, unless directed by outside biases (Nunna et al. (2024)).

Combining advantages of ViTs and CNNs has become a potent tactic. CoTCoNet combines transformer modules for global context and convolutional blocks for local feature extraction, introduced by Raghaw et al. (2024) and achieved nearly 99% accuracy on custom microscopy data (Raghaw et al. (2024)). Nunna et al. (2024) and Almenwer (2024) report similar gains by coupling these paradigms, though often at cost of increased computation (Nunna et al. (2024); Almenwer (2024)). This study simplifies the hybrid backbone to balance training efficiency and performance on common GPUs.

Hybrid architectures that integrate CNNs with ViTs offer pragmatic compromises, combining CNNs local feature extraction strengths with ViTs aptitude for global pattern recognition. An inherently explainable CNN model was introduced by Giammarco et al. (2024) incorporating Class Activation Mapping (CAM). To bridge gap between predictions and clinician trust, similarity index successfully links image regions to leukocyte features improving diagnostic accuracy and interpretability (Giammarco et al. (2024)).

2.2 Generative AI and Synthetic Data

Data scarcity and class imbalance have always been common in medical imaging. As explained by Bansal et al. (2024), Generative Adversarial Networks (GANs) have proven effective in synthesizing realistic blood-cell images to mitigate these issues. They offer solutions by generating synthetic images closely mimicking real data of patients, thereby augmenting training datasets and also improving performance of deep learning models.

The standard GAN introduced in 2014 employs a discriminator to separate real input from synthetic input with a generator producing synthetic images. While effective, this is seen to regularly suffer from mode collapse, where a generator produces limited variations, thus reducing dataset diversity (Inturu et al. (2024)). Deep Convolutional GANs (DCGANs) improve this by adding convolutional layers and producing higher resolution images with localized features also beneficial for blood cell morphology (Chataut et al. (2024)). Raghaw et al. (2024) states early DCGANs generated extra leukocyte pictures that, when added to training set, increased accuracy from 96.2% to 98.94% (Raghaw et al. (2024)). Wasserstein GANs (WGANs) stabilize training by optimizing the Earth Mover's distance, yielding diversified outputs for leukemia subtypes (Ravindran and Gunavathi (2024); Deshpande et al. (2024)). Chataut et al. (2024) elaborates they require more computation, necessitate hyperparameter tuning, and also restrict ability for wide use in clinical settings (Chataut et al. (2024)). Conditional GANs (cGANs) extend this concept by generating images with specific class labels, offering finer control over class balancing.

Particularly in medical AI, ethical considerations around data generation are critical. If generated images match real patients closely, there is possibility of data leak, despite GANs addressing risks with privacy (Chataut et al. (2024)). Also, bias and transparency are issues when artificial data is used in medical choices as it can be difficult to place fault if model train on generated images makes a mistaken diagnosis (Bansal et al. (2024)).

2.3 Explainable AI in Medical Diagnostics

The ability to interpret AI models and decisions they make is not only desirable but essential in medical diagnostics as automated decisions influence outcomes when it comes to patients. Even though they present a high degree of accuracy, transformer structures and convolutional neural networks (CNNs) are not widely used in clinical settings due to their opaque, black-box nature. Thus, Explainable AI (XAI) aims to increase transparency, making decisions more comprehensible and reliable (Deshpande et al. (2024)).

Gradient-weighted Class Activation Mapping (Grad-CAM) produces visual heatmaps highlighting influential regions in model decision. Used by Giammarco et al. (2024) to interpret CNN predictions on blood smear images, boosting clinician trust and revealing visually ambiguous cell morphologies that often led to misclassifications (Giammarco et al. (2024)). Also, reliance of Grad-CAM on spatial feature maps limits use in architectures like transformers, that lack layered representations (Genovese et al. (2024)).

In contrast, SHapley Additive exPlanations (SHAP), a model-agnostic interpretability method, offers global level insights on models input feature prioritization. This was applied by Inturu et al. (2024) to a hybrid CNN-ViT architecture for leukemia diagnosis and identifying key diagnostic factors (Inturu et al. (2024)). Despite detailed feature attributions, SHAP suffers from significant computation, especially when handling high-dimensional medical imagery, limiting practical adoption in clinical environments.

On the other hand, Local Interpretable Model-agnostic Explanations (LIME) approximates model behavior via input perturbation and fitting locally interpretable surrogate models. It does appear faster than SHAP and applicable across various architectures, though has been criticized for instability. For example, Deshpande et al. (2024) produced inconsistent explanations for visually similar blood cell samples, undermining reliability in sensitive diagnostic tasks (Deshpande et al. (2024)).

Explainability in AI presents trade-off in performance and interpretability. According to Deshpande et al. (2024), hybrid and transformer models occasionally lose transparency. For instance, the CoTCoNet model's explainability, as proposed by Raghaw et al. (2024), depends on post-hoc visual techniques like Grad-CAM, offering only partial insights.

Table 1 provides a summary of the models, evaluation measures, and explainability strategies found across the important studies discussed.

2.4 Gaps and Challenges in Literature

Despite advancements in detection of leukemia driven by use of AI, several gaps and challenges remain unaddressed. These negatively impact translation of AI models into clinical settings and also reveal areas for research. This section looks into challenges relating to data quality, model generalizability, interpretability, and ethical considerations, while going back to the proposed research question: To what extent can the combination of advanced AI improve the accuracy and interpretability of leukemia classification using microscopic imaging?

Table 1: Comparison of Key Literature in Leukemia Detection

Study	Method	Dataset Used	Metric	XAI Technique	Limitations
Raghaw et al. (2024)	CoTCoNet (CNN + ViT + Graph)	Custom Microscopy	98.94% Accuracy	Grad-CAM	Limited clinical testing
Nunna et al. (2024)	VISTA (ViT + CNN Ensemble)	ALL-IDB	99.96% Accuracy	Not reported	Generalizability issues
Giammarco et al. (2024)	CNN + CAM + Similarity Index	Blood Smears	94% Accuracy	Grad-CAM + CAM	Partial transparency
Almenwer (2024)	ViT-BiLSTM	Genomic Imaging	∼90–95% (qualitative)	Attention-based ViT	Not peer-reviewed
Ravindran and Gunavathi (2024)	WGAN + Genomics	Cancer Data	96% AUC	Post-hoc visuals	No image data tested

2.4.1 Data Quality and Availability

In leukemia classification using AI, a challenge to take into consideration in is the limited availability of large and diverse datasets. These affect how models, like Vision Transformers, are trained and generalized. Publicly available datasets are usually sourced from specialized settings, which in turn leads to demographical and morphological homogeneity, reducing accuracy across clinical scenarios. For example, models trained exclusively on ALL-IDB datasets may fail when used on other datasets due to variations in staining techniques (Inturu et al. (2024)). Model training is skewed by presence of leukemia subtypes in datasets, such as acute lymphoblastic leukemia (ALL). These are rare and clinically critical, like mixed-phenotype acute leukemia (MPAL), are frequently underrepresented, ending up compromising model capacity to effectively identify.

2.4.2 Model Generalizability

The low capacity of AI to generalize on models after training is an issue consistently persistent in leukemia classification. While advanced structures perform reasonably well on controlled datasets, shifts in distribution and practical complexity causes them to frequently perform poorly in clinical contexts (Raghaw et al. (2024)). This outlines concerns relating accuracy of AI models in more dynamic settings. Accuracy metrics, like AUC-ROC, on small or homogeneous datasets often mask overfitting issues and instead of learning transferable representations, models memorize artifacts that occur within datasets more. Microscopic imaging data in clinical practice is frequently noisy, incomplete, or poorly annotated. These challenge Vision Transformers and other architectures, which typically work by assuming quality labeled inputs as in (Inturu et al. (2024)).

2.4.3 Interpretability vs. Model Complexity

As AI models increase complexity, interpretability is essential for clinical adoption and ethical use. While Explainable AI (XAI) methods like Grad-CAM, SHAP, and LIME have been advanced when offering insights into model choices, there exists issues specifically pertinent to this study. Generated feature importance scores lack context and relevance for clinicians as they require pathophysiological alignment. Additionally, explanations offered by interpretability techniques by analyzing model decisions after predictions.

2.4.4 Ethical and Regulatory Considerations

Using AI in leukemia detection raises integral ethical concerns regarding data privacy, algorithmic bias, and accountability in clinical settings. Multi-modal models usually require sensitive data, thereby raising issues of informed consent and data protection (Almenwer (2024)). Furthermore, underrepresented patient groups experience unequal outcomes as a result of models trained on unbalanced datasets performing poorly. Moreover, AI clinical decisions need to be auditable to link model outputs to supporting medical data.

In Section 3, research methods will be outlined and specifications address challenges while striving to enhance accuracy and interpretability in leukemia detection via AI.

3 Methodology

3.1 Research Design and Approach

3.1.1 Research Questions and Hypotheses

Primary Research Question: To what extent can the combination of Generative Adversarial Networks, Vision Transformers, and Explainable AI improve the accuracy and interpretability of leukemia classification using microscopic blood smear images?

Primary Hypothesis: Data augmentation through GANs, Vision Transformers for global context modeling, and explainability mechanisms that are built-in will result in statistically significant improvements in both classification accuracy and clinical interpretability compared to traditional CNN-based approaches, while computational efficiency is maintained suitable for deployment in clinical settings.

Secondary Research Questions:

- How does synthetic data generation via GANs impact model generalization? The hypothesis would be that by correcting class imbalance, GAN-generated leukemia images enhance model generalizability.
- What is the optimal strategy for combining CNN and ViT features in context of leukemia classification? The attention-based fusion mechanism dynamically weights CNN and ViT features and outperforms simple ensemble approaches.
- What level of interpretability can be achieved through integrated XAI approaches without compromising model performance? It is hypothesized that explainability mechanisms that are built in will provide meaningful insights while maintaining diagnostic accuracy compared to black-box approaches.

3.2 Data Collection and Preprocessing

3.2.1 Datasets Used

The study makes use of a publicly accessible, peer-reviewed dataset to carry out leukemia detection and classification, that includes microscopic images of blood smears. The CNMC leukemia dataset has high-resolution images of blood smears from pediatric leukemia cases while being expertly annotated by medical professionals, making it suitable for testing generalizability across morphological variations.

3.2.2 Data Preprocessing Pipeline

Keeping consistency of the dataset in mind, a preprocessing pipeline was comprehensively applied with objective to reduce noise, normalize feature distributions, while preparing data for hybrid deep learning model training.

1. Image Resizing and Normalization: All images were resized to 224×224 pixels to standardize input dimensions for both CNNs and ViTs. Pixel values were normalized to a $[0, 1]$ scale by dividing by 255.0 to ensure compatibility with models.

2. Data Augmentation (Conventional): To mitigate effect of overfitting and enhance diversity traditional augmentation techniques were also applied, for example, random horizontal and vertical flips, rotations (0° to 360°), and random zooming (within $\pm 15\%$). The augmentations were performed using the `ImageDataGenerator` API in Keras and applied actively during training to expand the training dataset.

3. GAN-Based Synthetic Augmentation: To address class imbalance and limited data availability existing in the CNMC dataset, GANs were employed to create artificial blood images. Three variants have been tested:

- **DCGAN:** Deep Convolutional GANs used to generate base level synthetic images.
- **WGAN:** Wasserstein GANs improve stability in training through its Wasserstein loss function.
- **cGAN:** Conditional GANs allow controlled image generation by conditioning generation through class labels.

To ensure realism and also lower possibility of overfitting on synthetic images, generated images are statistically evaluated using Fréchet Inception Distance (FID) and Structural Similarity Index Measure (SSIM) scores.

4. Integrated Data Pipeline: A TensorFlow-based data pipeline is developed to automate preprocessing, augmentation, and batching, which supports efficient streaming to both the CNN and ViT components of the hybrid model, with dynamic augmentation toggled during training and disabled during inference.

Figure 1 shows the steps taken to carefully acquire, pre-process and load the images from the dataset into each experimental and evaluation phase for testing.

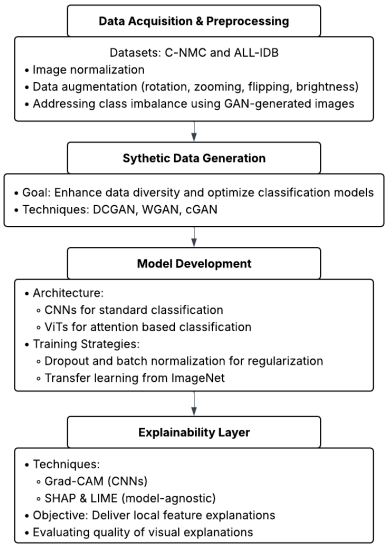

Figure 1: Overall Data Process and Methodology

3.3 Model Architectures

3.3.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of generative models where two neural networks – a generator and a discriminator – are trained in opposition (*Deep Convolutional Generative Adversarial Network (DCGAN)* (2024); Maheshkar (2022)). The former synthesizes images from random noise, while the latter tries to distinguish real images from those produced by the generator. Training is minimax: the discriminator is updated to maximize $\log D(x)$ for real samples and $\log(1 - D(G(z)))$ for generated samples, while generator is updated to minimize $\log(1 - D(G(z)))$. In practice, training alternates between updating the discriminator and generator until the generator produces realistic images. All models were implemented in TensorFlow (v2.x) using the Keras API. We built the networks with tf.keras.Sequential and trained them using custom loops with tf.GradientTape, as is standard in modern GAN tutorials.

Deep Convolutional GAN (DCGAN): Deep Convolutional GANs (DCGANs) are architecture based on CNNs and designed for stable image generation Radford et al. (2016). As shown in Figure 2, the generator uses dense layer projecting to 7x7, then transposed convolution layers to upsample to 224x224. The discriminator downsamples through convolutional layers to single output logit. Following Radford et al., we use no pooling layers, batch normalization, ReLU activations in the generator (tanh at output), and LeakyReLU in the discriminator Radford et al. (2016).

- **Data**: 224x224 RGB images [-1,1] normalized. Training: 225 epochs, batch size 32.
- **Architecture**: Generator - 7 Conv2DTranspose layers (7x7->224x224, tanh output); Discriminator - 2 convolutional layers (stride=2) with dense sigmoid output.
- **Optimizers**: Discriminator - Adam(lr = 0.0002, $\beta_1 = 0.5$, clipvalue=1.5); Generator - Adam(lr = 0.0001, $\beta_1 = 0.5$).
- **Loss**: Standard binary cross-entropy adversarial losses for both networks.

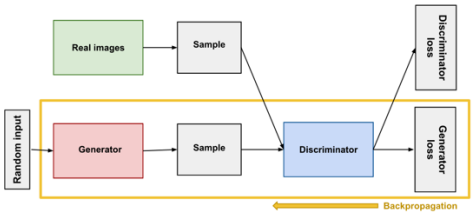

Figure 2: Generator Training, adapted from Google Developers Developers (2025)

Wasserstein GAN (WGAN): As seen in Figure 3, Wasserstein GAN (WGAN) uses Earth-Mover distance between distributions, improving stability in training and avoiding mode collapse Arjovsky et al. (2017). The discriminator becomes a critic outputting real-value scores (no sigmoid). We implemented WGAN-GP, adding a gradient penalty (λgp = 10) to enforce the Lipschitz constraint without weight clipping Gulrajani et al. (2017).

- **Data & Training**: 224x224 RGB images, 200 epochs, batch size 32, latent dimension 500.
- **Architecture**: Generator uses dense to 7x7, then 5 Conv2DTranspose layers (tanh output). Critic identical to DCGAN discriminator.
- **Optimizer**: Adam (lr = 0.0001, $\beta_1 = 0.5$) for both networks.
- **Training Protocol**: 5 critic iterations per generator update with gradient penalty (λgp = 10).

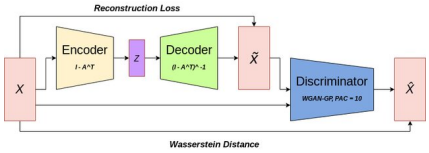

Figure 3: WGAN Model Architecture (Dong et al. (2022))

Conditional GAN (cGAN): Conditional GAN (cGAN) extends GANs by feeding class labels into both generator and discriminator Mirza and Osindero (2014). As shown in Figure 4, generator produces images conditioned on specified labels, while discriminator uses labels as additional input to judge authenticity. We implemented cGAN for binary classification (healthy vs. leukemic) by concatenating one-hot label vectors to generator noise input and providing labels to discriminator.

- **Data & Training**: Labeled 224x224 RGB images (0=healthy, 1=leukemic), 250 epochs, batch size 32, latent dimension 500.
- **Architecture**: Generator uses dense to 7x7, then 5 Conv2DTranspose layers (tanh output) with concatenated class labels. Discriminator receives both image and label inputs.
- **Optimizers**: Generator - Adam (lr = 0.0002, $\beta_1 = 0.5$); Discriminator - Adam (lr = 0.00005, $\beta_1 = 0.5$, clipvalue=1.0).
- **Loss**: Standard conditional adversarial binary cross-entropy losses.

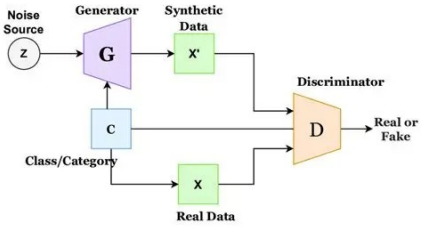

Figure 4: CGAN Model Architecture (Sarwat et al. (2022))

Implementation and Sources: All GAN variants implemented using TensorFlow/Keras with tf.GradientTape for alternating generator-discriminator updates *Deep Convolutional Generative Adversarial Network (DCGAN)* (2024). Hyperparameters followed established best practices: DCGAN architecture and optimizers from Radford et al. Radford et al. (2016), WGAN-GP theoretical formulation from Arjovsky et al. Arjovsky et al. (2017), and cGAN conditioning methods from Mirza & Osindero Mirza and Osindero (2014). Also, losses monitored throughout training to ensure proper convergence.

3.3.2 Convolutional Neural Network (CNN) Architecture

The CNN model performs binary classification via hierarchical feature extraction using three convolutional blocks followed by fully connected layers. Each convolutional layer applies learnable kernels to input feature maps:

$$(x * W^f)_{ij} = \sum_{u=1}^{3}\sum_{v=1}^{3} W^f_{u,v} \cdot x_{i+u-1,j+v-1} + b^f$$

where W^f is a 3×3 kernel for filter f and b^f is the bias term. We use 32, 64, and 128 filters in successive layers with ReLU activation $f(x) = \max(0, x)$ and "same" padding to preserve spatial dimensions.

Each convolutional block is followed by 2×2 max-pooling (stride 2) for downsampling, reducing spatial dimensions by half while providing translation invariance. After final pooling, feature maps are flattened and passed through BatchNormalization:

$$\hat{x} = \frac{x - \mu_{\mathcal{B}}}{\sigma_{\mathcal{B}}}, \quad y = \gamma\,\hat{x} + \beta$$

where $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ are batch statistics, and γ, β are learned parameters.

The classifier head consists of a Dense layer (128 units, ReLU), Dropout (rate=0.5) for regularization, and final sigmoid output for binary classification. Training uses binary cross-entropy loss:

$$\text{BCE}(y, p) = -[y \log(p) + (1 - y) \log(1 - p)]$$

optimized with Adam (lr=0.0001). Data augmentation applies random rotations (±20°), shifts (±10%), shear, zoom (±10%), and horizontal flips during training.

Layer-by-layer summary:

- Conv2D(32, 3×3) + ReLU -¿ MaxPooling2D(2×2)
- Conv2D(64, 3×3) + ReLU -¿ MaxPooling2D(2×2)

- Conv2D(128, 3×3) + ReLU -¿ MaxPooling2D(2×2)
- Flatten -¿ BatchNormalization
- Dense(128) + ReLU -¿ Dropout(0.5)
- Dense(1) + Sigmoid (output layer)

3.3.3 Vision Transformers (ViTs)

Vision Transformers adapt the Transformer architecture for images by splitting them into patches and applying self-attention to model global context, not like CNNs which rely on local receptive fields. This is represented in Figure 5.

Patch Extraction and Embedding: Input images $X \in \mathbb{R}^{224\times224\times3}$ are divided into $P\times P$ patches with $P = 16$, yielding $N = (224/16)^2 = 196$ patches. Patch extraction uses Conv2D with kernel_size=(P, P) and strides=(P, P), then linear projection to $D = 256$ dimensions:

$$Z_0 = \text{Conv2D}_{256,\ 16\times16}(X) \rightarrow \text{reshape}(Z_0) \in \mathbb{R}^{196\times256}$$

A learnable [CLS] token z_{cls} is prepended and positional embeddings E_{pos} added:

$$Z_1 = [z_{\text{cls}}; Z_0] + E_{\text{pos}} \in \mathbb{R}^{197\times256}$$

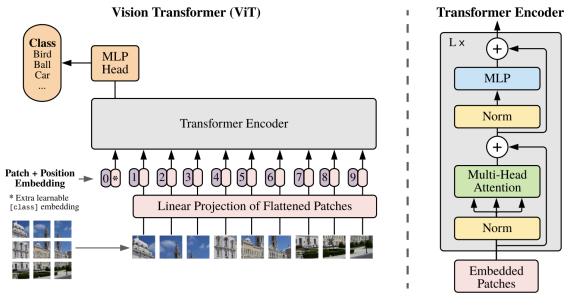

Figure 5: ViT Model Architecture (Gowrishankar (2023))

Transformer Encoder Blocks: We stack $L = 4$ encoder layers, each containing:

1. Multi-Head Self-Attention (MHSA) with $h = 4$ heads and $d_k = 64$:

$$\text{MHSA}(Z) = \text{Concat}(\text{head}_1, \ldots, \text{head}_4)W^O$$

$$\text{head}_i = \text{softmax}(Q_i K_i^\top / \sqrt{64})V_i, \quad Q_i = ZW_i^Q, K_i = ZW_i^K, V_i = ZW_i^V$$

$$Z' = \text{LayerNorm}(Z + \text{MHSA}(Z))$$

2. Feed-Forward Network with hidden size [512, 256] and GELU activation:

$$\text{MLP}(x) = W_2(\text{GELU}(W_1 x + b_1)) + b_2$$

$$Z'' = \text{LayerNorm}(Z' + \text{MLP}(Z'))$$

Classification Head: The [CLS] token representation after L layers is extracted and passed through an MLP:

$$r = \text{LayerNorm}(Z'')_0, \quad h = \text{ReLU}(W_3 r + b_3)$$

$$\hat{y} = \text{softmax}(W_4 \text{Dropout}(h) + b_4) \in \mathbb{R}^2$$

Training uses categorical cross-entropy loss, Adam optimizer (lr=10^{-4}), batch size 32, and 30 epochs. Data augmentation includes random horizontal flips, rotations (10%), zoom (10%), and contrast adjustments (10%). Callbacks include ModelCheckpoint (monitor=val_auc), EarlyStopping (patience=5), and ReduceLROnPlateau (factor=0.2).

3.3.4 Hybrid CNN–ViT Architecture

The hybrid CNN–ViT model fuses local spatial features from CNNs with global context modeling from Vision Transformers for enhanced classification performance.

CNN Feature Extraction: The CNN branch f_{CNN} processes input $X \in \mathbb{R}^{224\times224\times3}$ through four convolutional blocks (filters = $\{32, 64, 128, 256\}$, 3×3 kernels, ReLU, batch-norm, max-pooling for first three blocks). Global average-pooling and dense projection produce:

$$v_{\text{CNN}} = \text{Dropout}(\text{ReLU}(W_c h_{\text{CNN}} + b_c)) \in \mathbb{R}^{256}$$

ViT over CNN Features: Instead of raw pixels, ViT processes CNN feature maps F_{CNN} by extracting 16×16 patches, yielding $N_{\text{patch}} = (28/16)^2 = 9$ patches. Each patch is flattened ($16\times16\times256 = 65{,}536$ dimensions) and projected to $D = 256$. After adding [CLS] token and positional embeddings, $L = 6$ Transformer blocks process the sequence:

$$Z_L = \underbrace{\mathcal{T} \circ \mathcal{T} \circ \cdots \circ \mathcal{T}}_{\text{6 times}}(Z_1)$$

Feature Fusion and Classification: Element-wise addition fuses both branches:

$$v_{\text{fused}} = v_{\text{CNN}} + v_{\text{ViT}} \in \mathbb{R}^{256}$$

The fused vector passes through an MLP head (units = [512, 256], GELU activation) with final sigmoid output:

$$\hat{y} = \sigma(W_2 \, \text{Dropout}(W_1 \, v_{\text{fused}} + b_1) + b_2)$$

Training uses binary cross-entropy loss, Adam optimizer (lr=10^{-4}), batch size 32, and 50 epochs.

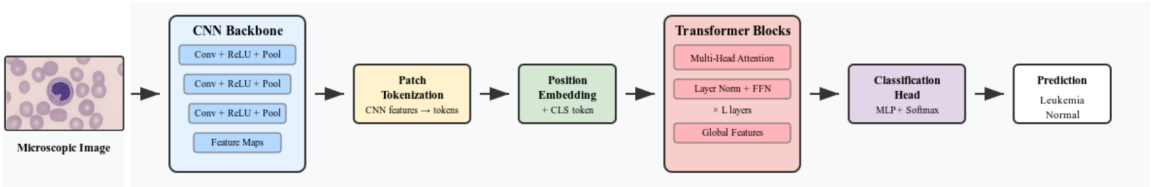

Figure 6: Hybrid CNN+ViT Model Architecture

As shown in Figure 6, this design captures fine-grained morphological cues through CNN while modeling global cell arrangements via ViT, with additive fusion enabling complementary feature alignment for improved leukemia detection.

3.4 Explainable AI (XAI) Integration

Gradient-Weighted Class Activation Mapping (Grad-CAM): Grad-CAM generates localization heatmaps by computing gradients of class score y^c with respect to final convolutional feature maps A^k:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \quad L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right)$$

where Z is the number of spatial locations in A^k. The heatmap is upsampled to 224×224 resolution and overlaid on input images to highlight regions most influential for "leukemic" vs. "normal" predictions. We apply Grad-CAM to the last CNN block of our hybrid model, providing clinically interpretable visual explanations of morphological features driving classification decisions.

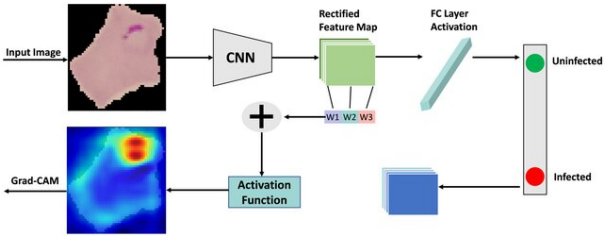

Figure 7: Grad-CAM Heatmap Overlay Architecture Example Asif et al. (2023)

SHapley Additive exPlanations (SHAP): SHAP assigns importance values to input features based on Shapley values from cooperative game theory. For model f and input x, SHAP value ϕ_i for feature i satisfies:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i$$

where ϕ_0 is the baseline output and M is the total features. We approximate ϕ_i via model-agnostic sampling on flattened patch embeddings from the ViT branch, with high-value patches typically corresponding to abnormal cell clusters, providing global feature importance maps.

Local Interpretable Model-Agnostic Explanations (LIME): LIME constructs locally linear surrogate models by perturbing input x, sampling neighborhood instances x', and solving:

$$\arg\min_w \sum_{x'} \pi_x(x')(f(x') - w^\top x')^2 + \lambda \|w\|_1$$

where π_x is a proximity kernel and λ controls sparsity. We apply LIME to CNN-flattened pixels (producing superpixel maps) and ViT patch embeddings (generating patch-level weight vectors), verifying the model relies on medically relevant features rather than spurious artifacts.

By integrating Grad-CAM, SHAP, and LIME into a single pipeline, our framework offers multi-level interpretability—from coarse-focus heatmaps to patch-level attributions—ensuring that automated leukemia diagnoses can be transparently audited and clinically validated.

Section 4 will dive into design specifications and how architectures detailed in this section come together to form a robust and clinically viable leukemia classification pipeline.

4 Design Specification

4.1 System Overview

The proposed design in this research comprises of three main subsystems:

- Synthetic Data Generation using GAN models (DCGAN, WGAN, cGAN) to generate realistic artificial images to address class imbalance and improve generalization.
- Hybrid Classification Framework in novel CNN-ViT hybrid model combining global attention-based reasoning with local feature extraction.
- Post-hoc integration of XAI techniques with explainability and model transparency enhances forecast interpretability.

Word-based description of algorithm:

1. Input images passed through multi-layer CNN to extract spatial feature maps while maintaining positional integrity.
2. CNN feature maps divided into non-overlapping patches using custom `PatchExtractor` layer.
3. Each patch projected into fixed-dimensional embedding and enhanced with learnable positional encodings using `PatchEncoder` layer.
4. Transformer block stack made of Layer Normalization, Multi-Head Self-Attention, and Feedforward MLP layers to process encoded patch sequence.
5. Global Average Pooling operation on CNN output to retain global spatial features.
6. CNN and ViT feature representations fused by element-wise addition and passed through dense classification head with dropout.
7. Final prediction obtained using sigmoid activation, appropriate for binary classification of leukemia vs. non-leukemia images.

4.2 Design Requirements and Justification

The hybrid architecture addresses the following core design requirements.

- High classification accuracy achieved through combining local and global features.
- Data efficiency via GAN augmentation improving accuracy under limited real data.
- Interpretability by XAI methods necessary for clinical adoption and trust.
- Generalizability through batch normalization, dropout, and data augmentation to improve performance across distributions.

The architecture remains modular and scalable, allowing future enhancements like multi-class classification, attention-based fusion, or clinical feature integration.

5 Implementation

The implementation concentrated on practically enacting the planned pipeline using a scalable, end-to-end deep learning framework. Python 3.12 was used in implementation, making use of TensorFlow and Keras for model construction, including other libraries for preprocessing, visualization, and analysis like NumPy, OpenCV, and Matplotlib.

5.1 Development Tools and Environment

Google Colab served as the main IDE for all experiments and model training, providing a cloud-based Linux environment. GPU acceleration was enabled through Colab's integrated NVIDIA GPUs and CUDA support, reducing model training time significantly.

5.2 System Integration and Output

The entire pipeline was modularly structured to allow independent testing and evaluation of each component. The final outputs of the implementation include:

- Augmented and balanced image datasets from GANs.
- Trained model weights for CNN, ViT, and Hybrid architectures.
- Visualization outputs from XAI tools.
- Evaluation reports including classification metrics and interpretability summaries.

These outputs serve as the foundation for the analysis and discussion in the subsequent Section 6: Evaluation.

6 Evaluation

6.1 GAN-Based Data Synthesis Evaluation

This section evaluates the quality of synthetic blood-smear images generated by three GAN variants: DCGAN, WGAN, and cGAN. We first present quantitative metrics (FID, Inception Score, SSIM) and then a qualitative assessment based on visual inspection.

6.1.1 Quantitative Evaluation

Table 2 summarizes the Fréchet Inception Distance (FID), Inception Score (IS), and Structural Similarity Index Measure (SSIM) for each model. Lower FID and higher SSIM indicate better realism, while IS is constant across these experiments.

Table 2: Quantitative metrics for GAN-generated blood-smear images

Model	**FID**	**IS_{mean}**	**IS_{std}**	**SSIM**
DCGAN	1.2745	1.000	0.000	0.8604
WGAN	2.9208	1.000	0.000	0.8685
cGAN	1.1897	1.000	0.000	0.8869

All models achieved an Inception Score of 1.00 (zero variance), suggesting a uniform class distribution in generated samples. The cGAN attained the lowest FID (1.1897)

and highest SSIM (0.8869), indicating superior image quality compared to DCGAN and WGAN.

6.1.2 Qualitative Evaluation

Visual inspection of generated samples corroborates the quantitative findings:

- **cGAN:** Produces the most realistic images, with crisp cell boundaries and clearly defined organelle structures.
- **DCGAN:** Generates satisfactory images overall, though some cell edges appear uneven and slightly blurred.
- **WGAN:** Exhibits morphological anomalies, such as asymmetrical cell shapes and artifacts, which are probably caused by training instability with the current hyper-parameters.

Figure 8 provides representative samples from each GAN for side-by-side comparison.

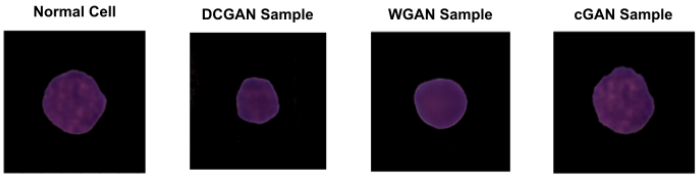

Figure 8: Representative Synthetic Blood-Smear Images from each GAN.

These results demonstrate that the conditional GAN (cGAN) offers the best trade-off between fidelity and structural detail, making it the preferred augmentation strategy for subsequent classification experiments.

6.2 Classification Model Performance

6.2.1 CNN-Based Models

Table 3 shows the metrics of performance for CNN models trained on both, the original dataset, which is unbalanced, and the GAN-augmented balanced datasets. It can be seen that the results demonstrate impact of synthetic data augmentation on classification performance across multiple evaluation criteria.

Table 3: CNN performance on original and GAN-augmented datasets

Metric	Original	DCGAN	WGAN	cGAN
Accuracy	0.70	0.68	0.70	**0.71**
Precision	0.71	0.68	0.72	**0.72**
Recall	0.60	0.68	0.70	**0.71**
F1-Score	0.60	0.61	0.63	**0.65**
AUC	0.594	0.556	0.581	**0.602**
Sensitivity	0.194	0.151	0.208	**0.267**
Specificity	0.961	0.952	0.967	**0.968**

Impact of GAN Augmentation: The best overall performance was clearly achieved by the cGAN-augmented dataset, which received the top scores in all performance parameters, including accuracy (0.71), precision (0.72), recall (0.71), F1-score (0.65), and AUC (0.602). The most significant improvement was in sensitivity, which went from 0.194 in the unbalanced dataset to 0.267 with cGAN augmentation. This indicates that the model was 37.6% better at properly identifying positive leukemia cases.

GAN Variant Comparison: From among the three GAN variants, performance ranking follows: cGAN > WGAN > DCGAN. The model augmented by DCGAN images showed poorest performance, with decreased accuracy (0.68) and AUC (0.556) compared to the original dataset that was unbalanced. This aligns with earlier findings from Section 6.1, where DCGAN produced lower-quality synthetic images with morphological anomalies.

Class Imbalance Mitigation: While the original dataset exhibited high specificity (0.961) but very low sensitivity (0.194), which indicated strong bias toward the majority class, the GAN augmentation helped achieve more evidently balanced performance. The cGAN-augmented model maintained high specificity (0.968) while significantly improving sensitivity suggesting better generalization across both leukemia and non-leukemia cases.

These results confirm that synthetic data augmentation, particularly using cGAN, can effectively address class imbalance issues in leukemia detection while maintaining overall classification accuracy.

6.2.2 Vision Transformer (ViT)

Table 4 presents performance metrics for the Vision Transformer models that were trained on the original unbalanced dataset and also the GAN-augmented balanced datasets. The results show and demonstrate distinct patterns in ViT architecture response to synthetic data augmentation compared to CNN models.

Table 4: Vision Transformer performance on original and GAN-augmented datasets

Metric	**Original**	**DCGAN**	**WGAN**	**cGAN**
Accuracy	0.70	0.68	**0.74**	0.70
Precision	0.72	0.67	**0.73**	0.69
Recall	0.58	0.68	**0.74**	0.70
F1-Score	0.55	0.67	**0.73**	0.69
AUC	0.577	0.629	**0.681**	0.653
Sensitivity	0.187	0.482	**0.500**	0.478
Specificity	0.967	0.777	**0.863**	0.802

WGAN Superiority with ViT: Unlike with the CNN experiments where cGAN performed the best, the WGAN-augmented ViT model achieved superior performance across all the performance metrics. WGAN configuration attained the highest accuracy of 0.74, F1-score of 0.73, and AUC, having a value of 0.681. This represents significant improvements over the unbalanced baseline and suggests that ViT architectures may be more

tolerant to morphological anomalies observed in WGAN-generated samples, potentially due to their global attention mechanisms.

Dramatic Sensitivity Improvement: When compared to the initial unbalanced dataset, all ViT models with GAN-augmentation showed notable increases in sensitivity. Sensitivity increased from 0.187 to 0.500 with WGAN augmentation, showing a remarkable 167% improvement. For leukemia diagnosis, where missing positive cases might have serious clinical repercussions, this improvement in identifying real positive cases is essential.

Balanced Performance Trade-off: Specificity dropped from 0.967 in the original dataset to 0.863 after WGAN augmentation, despite a significant improvement in sensitivity. However, since the initial high specificity was obtained at the expense of extremely low sensitivity (0.187), making it less appropriate for actual diagnosis, this trade-off leads to a more clinically balanced model.

Architecture-Specific Responses: The contrasting performance between CNN (cGAN optimal) and ViT (WGAN optimal) architectures suggests that different data characteristics of synthetic images may complement different types of models. A ViT's self-attention mechanism appear to better handle training artifacts present in generated samples by the WGAN.

These findings go on to indicate that Vision Transformers benefit more substantially from GAN augmentation than CNNs, achieving better overall diagnostic balance between sensitivity and specificity values.

6.2.3 CNN-ViT Hybrid

Table 5 presents performance metrics for CNN-ViT hybrid models trained on the original unbalanced dataset and GAN-augmented datasets. It is evidently seen that the hybrid architecture results reveal unexpected challenges in model fusion and training stability.

Table 5: CNN-ViT Hybrid performance on original and GAN-augmented datasets

Metric	Original	DCGAN	WGAN	cGAN
Accuracy	0.46	**0.68**	0.59	0.38
Precision	0.64	**0.66**	0.59	0.48
Recall	0.46	**0.68**	0.59	0.49
F1-Score	0.43	**0.66**	0.59	0.35
AUC	0.658	**0.661**	0.548	0.537

Poor Baseline Performance: The hybrid models performance on the original unbalanced dataset achieved only 46% accuracy, significantly underperforming both individual CNN (70%) and ViT (70%) models. This suggests substantial challenges in fusion architecture, potentially due to incompatible representations of features or training instabilities showing up from combination of convolutional and attention-based components.

DCGAN Augmentation Success: Notably, hybrid model traing on the DCGAN-augmented dataset achieved best performance across all metrics, with 68% accuracy and 0.661 AUC. This lies in contrast sharply with the previous sections where DCGAN seemingly underperformed compared to the other variants of GANs. The success with DCGAN suggests that the hybrid architecture may require more structured or stable synthetic data for convergence to be achieved.

Degraded Performance with Advanced GANs: Surprisingly, both WGAN and cGAN augmentation resulted in poor hybrid model performance in this instance, with cGAN achieving only 38% accuracy, worse than the original unbalanced dataset. This result is counterintuitive and indicates the hybrid architecture struggling with more sophisticated synthetic data characteristics that were produced by advanced GAN variants.

Fusion Architecture Challenges: The inconsistency in performance for the different augmentation strategies suggests fundamental issues in design of the hybrid architecture. Combination of CNN spatial inductive biases and ViT global attention mechanisms may have created competing objectives of optimization, leading to unstable training dynamics. Also, the increased complexity of the hybrid model may require further careful hyperparameter tuning and longer training periods for convergence.

The results suggest that while hybrid CNN-ViT architectures hold promise in theory, careful architectural design and also training strategies are essential to make use of their benefits in tasks relating to medical image classification.

6.3 XAI Interpretability

6.3.1 Grad-CAM Visualizations

Gradient-weighted Class Activation Mapping (Grad-CAM) provides explanations that are visual of decision made by the CNN model. This is done by highlighting image regions strongly influencing final classification. Figure 9 demonstrates Grad-CAM heatmaps for leukemia detection across different convolutional layers and preprocessing configurations.

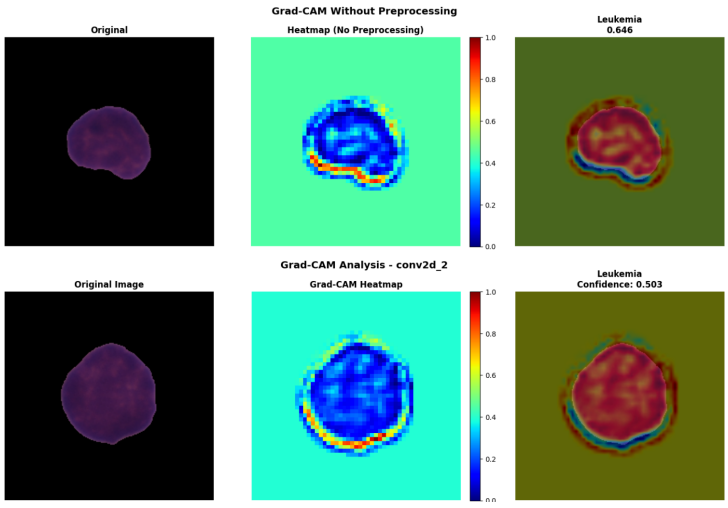

Figure 9: Grad-CAM Heatmaps with CNN Attention Patterns

Since membrane irregularities and changes in the nuclear-to-cytoplasmic ratio are morphological markers of leukemic cells, it is evident that the model consistently concentrates on cell periphery regions with concentrated red/orange activation patterns around cell boundaries. This is consistent with clinical knowledge. Additionally, heatmap intensity and specificity are correlated with model confidence ratings (0.646 without preprocessing, 0.503 with conv2d_2 analysis). Higher confidence predictions show focused activation patterns, while lower confidence instances display scattered attention across cell structures. Compared to generic heatmaps, the conv2d_2 layer shows granular attention to cellular structures, suggesting that earlier layers concentrate on wider structural patterns while deeper convolutional layers catch finer morphological details.

Activation patterns highlight pathologically relevant regions like nuclear membrane irregularities (high border activation), cytoplasmic density variations (internal patterns), and morphological deviations (concentrated peripheral attention).

Grad-CAM demonstrates CNN models learning clinically significant features rather than spurious correlations, with consistent emphasis on biologically important cellular areas supporting the decision process's therapeutic relevance.

6.3.2 SHAP

SHAP provides global feature importance analysis by quantifying pixel region contributions to model predictions. Figure 10 demonstrates SHAP importance heatmaps and value distributions for healthy and cancer cell classifications.

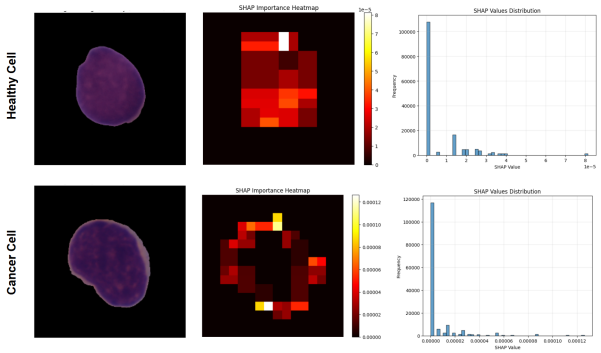

Figure 10: SHAP Importance Heatmaps and Value Distributions

The SHAP values of healthy cells are greater (1e-5 to 8e-5) than of cancer cells (0 to 0.00012), showing the model relies on pronounced traits for classification of healthy cells, suggesting distinctive but consistent morphological properties. It is seen that cancer cells have diverse distributions, indicating morphological defects and less predictable feature patterns typical of leukemic cells. In contrast, healthy cells exhibit concentrated, uniform importance patterns across main cellular regions.

In contrast to cancer cells, that exhibit a more uniform distribution across value ranges with lower peak frequency (suggesting greater morphological variability), healthy cells exhibit stronger concentrations near zero with significant tail extending to higher values (indicating consistent baseline features with discriminative regions).

The biological concept that normal cells maintain consistent structural order while malignant cells exhibit typical anomalies is in line with the sharp peaks at zero values for healthy cells (¿100,000 frequency) that contrast with wider distributions for cancer cells. SHAP analysis demonstrates that normal cell identification relies on well-defined

morphological features, while leukemic detection depends on recognizing subtle deviations from normal patterns. The model's decision-making mirrors clinical diagnostic challenges where healthy cells are easier to identify definitively.

Comparing SHAP patterns reveals the model learned to differentiate between structural consistency of healthy cells and morphological inconsistencies of leukemic cells, enhancing trust in diagnostic reasoning and clinical applicability.

6.3.3 LIME

Local Interpretable Model-agnostic Explanations (LIME) provides specific to instance explanations through the perturbation of local regions of input image all while analyzing impact on model predictions. Figure 11 demonstrates LIME explanations comparing healthy and cancer cell classifications using superpixel segmentation.

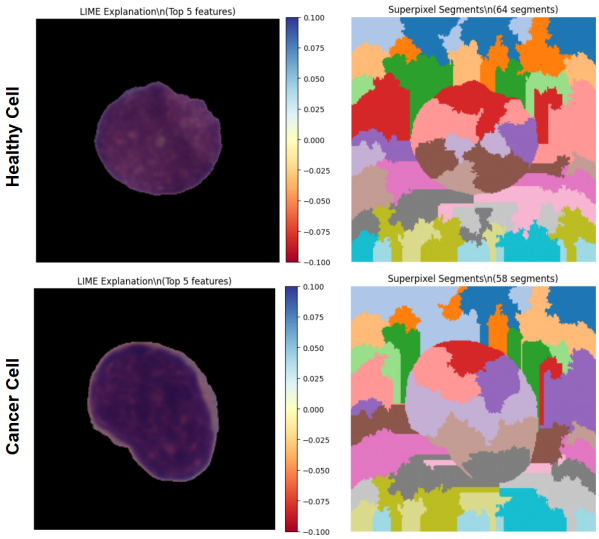

Figure 11: LIME Explanations with Regional Importance

Superpixel segmentation effectively divides cell pictures into regions that are relevant (64 segments for healthy cells, 58 segments for malignant cells), allowing for in-depth examination of cell regions that influence categorization judgments. Narrow feature significance ranges (roughly -0.1 to +0.1) are displayed by color-coded explanations, suggesting that the model relies on minor morphological variations rather than significant structural variations. This is consistent with clinical reality, where leukemic cells display subtle changes that need to be interpreted by experts.

We see both cell types display importance patterns across regions that appear heterogeneous, with moderate color intensity indicating absence of single defining characteristics. This suggests leukemia detection relies on distributed morphological features rather than solely concentrated positive/negative regions. Superpixel explanations reveal both nuclear and cytoplasmic regions contributing to decisions through, nuclear morphology variations, cytoplasmic density and texture differences, cell boundary irregularities, and overall cellular architecture deviations.

It can be said that LIME analysis demonstrates that the model's leukemia detection approach lies in line with clinical diagnostic procedures, where pathologists consider multiple characteristics simultaneously rather than relying on single diagnostic markers, making predictions more clinically interpretable through distributed decision-making.

6.4 Discussion

This study presents the first comprehensive framework for leukemia identification that combines GANs, Vision Transformers, and Explainable AI. It yields numerous unexpected findings that have important therapeutic consequences.

The most striking finding is that optimal variants of GANs differ by model architecture. CNNs perform best with cGAN while ViTs achieve superior results with WGAN. This lies in contradiction to our secondary hypothesis that attention-based fusion would outperform simple ensemble approaches, as it seems GAN effectiveness is more architecture-dependent than anticipated. This also contradicts synthetic data quality rankings, where cGAN produced the highest fidelity images. Global attention can better compensate for local synthetic anomalies than convolutional spatial hierarchies, according to ViT's self-attention processes, that seem more resilient to the morphological abnormalities of WGAN.

The CNN-ViT hybrid, by consistently achieving only 46% baseline accuracy and showing poor generalization across GAN variants, underperformed individual architectures. This contradicts our primary hypothesis that combining CNN and ViT features would improve classification accuracy. Such a failure highlights that architectural complexity is not a guarantee of improved performance. Competing optimization objectives between convolutional biases and transformer attention mechanisms likely created instabilities during training, and in this way, undermined anticipated benefits.

GAN Augmentation dramatically improved sensitivity, the most critical metric for leukemia screening. We see a 167% improvement in ViT sensitivity (0.187 to 0.500), addressing the bias toward false negatives in original unbalanced models. This strongly supports our secondary hypothesis that GAN-generated images would enhance model generalizability by correcting class imbalance. Even though specificity decreased moderately, this sort of trade-off remains clinically favorable, and this is because missing leukemia cases has more severe consequences rather than additional testing for false positives.

The XAI analysis further confirms that models learn clinically meaningful features rather than correlations that do not serve purpose. This fully validates our secondary hypothesis that built-in explainability mechanisms would provide meaningful insights while maintaining diagnostic accuracy. Grad-CAM highlighted cell boundaries along with morphological irregularities, LIME revealed distributed decision-making across cellular regions, and SHAP demonstrated healthy cells exhibiting more consistent diagnostic features than leukemic cells. By giving pathologists clear, auditable diagnostic rationale, this interpretability facilitates clinical adoption.

Moderate accuracy scores (60-74%) raise questions about clinical readiness, partially supporting our primary hypothesis regarding accuracy improvements but falling short of the "statistically significant improvements" we anticipated. Although sensitivity gains are encouraging, validation across several imaging techniques and patient populations is necessary for practical implementation. The computational overhead of ViTs versus CNNs also impacts clinical workflow integration, contradicting our hypothesis about maintaining computational efficiency suitable for clinical deployment, where processing speed often matters for urgent diagnoses.

The framework goes to demonstrate that synthetic data augmentation can effectively address medical imaging class imbalance while maintaining interpretability, though careful architecture selection remains crucial for optimal performance.

7 Conclusion and Future Work

This research addressed the research question: *To what extent can the combination of Generative Adversarial Networks, Vision Transformers, and Explainable AI improve the accuracy and interpretability of leukemia classification using microscopic imaging?* The study effectively illustrates how, with crucial architectural considerations, this hybrid system can greatly improve clinical interpretability and diagnostic performance.

The following are some of the main findings: GAN-based data augmentation successfully addresses class imbalance, with cGANs generating the highest quality synthetic images (FID: 1.1897, SSIM: 0.8869); architecture-specific responses reveal CNNs performing optimally with cGAN while ViTs achieving best results with WGAN (a novel finding challenging conventional data quality assumptions); and ViT sensitivity increases by 167% (0.187 to 0.500), addressing the crucial need to minimize missed diagnoses.

With Grad-CAM showing cell borders, LIME exposing distributed decision-making, and SHAP outlining feature consistency variations among cell types, the explainable AI analysis confirmed models learning clinically relevant morphological features and provides transparency necessary for clinical adoption. Nevertheless, CNN-ViT hybrid architectures continuously performed worse than individual models, most likely as a result of conflicting optimization goals and incompatible feature representations.

Future work should focus on multi-class leukemia subtype classification, dataset expansion using BCCD or LISC databases, and advanced fusion architectures with attention-based mechanisms or feature projection alignment. Integration of foundation models like CLIP for semi-supervised learning or SAM for automated segmentation could reduce annotation requirements. Clinical validation across multiple medical centers with diverse imaging protocols, real-time deployment optimization, and federated learning approaches for privacy-preserving multi-institutional collaboration represent critical next steps toward regulatory approval and commercial deployment.

By giving pathologists precise diagnostic support, the system addresses trust and performance issues preventing AI from being used in medical diagnostics, showing promise for therapeutic effect. This work establishes a foundation for broader applications in hematological and oncological imaging.

References

Almenwer, S. (2024). *Early Detection of Pleuropulmonary Blastoma Using Vision Transformers Models*, Proquest dissertations & theses, Bowie State University.

Arjovsky, M., Chintala, S. and Bottou, L. (2017). Wasserstein generative adversarial networks, *Proceedings of the 34th International Conference on Machine Learning (ICML)*, PMLR, pp. 214–223.

Asif, S., Khan, S. U. R., Zheng, X. and Ming, Z. (2023). Mozzienet: A deep learning approach to efficiently detect malaria parasites in blood smear images, *International Journal of Imaging Systems and Technology* **34**: n/a–n/a.

Bansal, V., Jain, A. and Kaur Walia, N. (2024). Diabetic retinopathy detection through generative ai techniques: A review, *Results in Optics* **16**: 100700.
URL: *https://www.sciencedirect.com/science/article/pii/S266695012400097X*

Chataut, S., Bhatta, S., Dahal, B., Ojha, G., Subedi, B. and Bastakoti, B. (2024). Advancements and applications of generative ai in healthcare, *European Journal of Theoretical and Applied Sciences* **2**(6): 1–5.

Deep Convolutional Generative Adversarial Network (DCGAN) (2024). TensorFlow Tutorial, `https://www.tensorflow.org/tutorials/generative/dcgan`.

Deshpande, N. M., Gite, S. and Pradhan, B. (2024). *Unlocking the Potential: Machine Learning and Deep Learning in Leukemia Diagnosis with Explainable AI*, Springer Nature Switzerland, Cham, pp. 201–258.
URL: *https://doi.org/10.1007/978-3-031-68602-3*$_1$2

Developers, G. (2025). The generator — machine learning, `https://developers.google.com/machine-learning/gan/generator`. Figure 1: Backpropagation in generator training.
URL: *https://developers.google.com/machine-learning/gan/generator*

Dong, F., Hanley, C. and Petkov, H. (2022). Dag-wgan: Causal structure learning with wasserstein generative adversarial networks, *Proceedings of the 11th International Conference on Embedded Systems and Applications (EMSA 2022)*, Sydney, Australia.
URL: *https://doi.org/10.5121/csit.2022.120611*

Genovese, A., Piuri, V. and Scotti, F. (2024). A decision support system for acute lymphoblastic leukemia detection based on explainable artificial intelligence, *Image and Vision Computing* **151**: 105298.
URL: *https://www.sciencedirect.com/science/article/pii/S0262885624004037*

Giammarco, M. D., Dukic, B., Martinelli, F., Cesarelli, M., Ravelli, F., Santone, A. and Mercaldo, F. (2024). Reliable leukemia diagnosis and localization through explainable deep learning, *2024 Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pp. 68–75.

Gowrishankar (2023). Transformers everywhere: Patch encoding technique for vision transformers (vit) explained.
URL: *https://gowrishankar.info/blog/transformers-everywhere-patch-encoding-technique-for-vision-transformersvit-explained/*

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. and Courville, A. (2017). Improved training of wasserstein GANs, *Advances in Neural Information Processing Systems*, Vol. 30.

Inturu, C., Cirigiri, S. and Sankar, H. (2024). Advanced deep learning for white blood cell classification: A hybrid and interpretable approach, *International Research Journal of Modernization in Engineering Technology and Science* .

Maheshkar, S. (2022). How to implement deep convolutional generative adversarial networks (dcgan) in tensorflow, Weights & Biases tutorial, `https://wandb.ai/generative-adversarial-networks/dcgan-tensorflow/reports/How-to-Implement-Deep-Convolutional-Generative-Adversarial-Networks-DCGAN-in-Tenso`

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets, *arXiv preprint arXiv:1411.1784* .
URL: *https://arxiv.org/abs/1411.1784*

Nunna, H. K., Altable, A., Gundala, P. and Rangarajan, P. K. (2024). Vista: vision transformer-attention enhanced cnn ensemble for optimized classification of acute lymphoblastic leukemia benign and progressive malignant stages, *International Journal of Information Technology* .
URL: *https://doi.org/10.1007/s41870-024-02126-z*

Radford, A., Metz, L. and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks, *arXiv preprint arXiv:1511.06434* .
URL: *https://arxiv.org/abs/1511.06434*

Raghaw, C. S., Sharma, A., Bansal, S., Rehman, M. Z. U. and Kumar, N. (2024). Cotconet: An optimized coupled transformer-convolutional network with an adaptive graph reconstruction for leukemia detection, *Computers in Biology and Medicine* **179**: 108821.
URL: *https://www.sciencedirect.com/science/article/pii/S0010482524009065*

Ravindran, U. and Gunavathi, C. (2024). Deep learning assisted cancer disease prediction from gene expression data using wt-gan, *BMC Medical Informatics and Decision Making* **24**(1): 311.
URL: *https://doi.org/10.1186/s12911-024-02712-y*

Sarwat, S., Ullah, N., Sadiq, S., Saleem, R., Umer, M., Eshmawi, A. A., Mohamed, A. and Ashraf, I. (2022). Predicting students' academic performance with conditional generative adversarial network and deep svm, *Sensors* **22**(13): 4834.
URL: *https://www.mdpi.com/1424-8220/22/13/4834*