A Coarse-to-Fine Hierarchical Framework for Bone Marrow Cell Recognition: Integrating Morphological Features with Class-Specific Augmentation and Multi-Perspective Explainable AI

by

Abdullah AL Mamun
21301580
Zarin Tasnim Haider
21301380
Sidratul Montaha
21301300
Ismat Jahan
21301541

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material that has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Abdullah Al Mamun
21301580

Zarin Tasnim Haider
21301380

Sidratul Montaha
21301300

Ismat Jahan
21301541

Approval

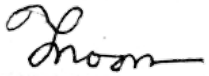
The thesis/project titled "A Coarse-to-Fine Hierarchical Framework for Bone Marrow Cell Recognition: Integrating Morphological Features with Class-Specific Augmentation and Multi-Perspective Explainable AI" is submitted by

1. Abdullah Al Mamun (21301580)
2. Zarin Tasnim Haider (21301380)
3. Sidratul Montaha (21301300)
4. Ismat Jahan (21301541)

Of Summer 2025 have been accepted as satisfactory in partial fulfilment of the requirement for the degree of B.Sc. in Computer Science on October 10, 2025.

Examining Committee:

Supervisor:
(Member)



Jannatun Noor Mukta, PhD
Associate Professor and Director
Data Science Program
United International University
Contractual Associate Professor
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

Hamim Ibne Nasim
Contractual Faculty
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Associate Professor
Department of Computer Science and Engineering
BRAC University

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis has been completed without any major interruption.
Secondly, to our Supervisor, Dr. Jannatun Noor Mukta Ma'am, for her kind support and advice in our work.

Abstract

The classification of bone marrow (BM) cell is essential for the diagnosis of many haematological disorders. Automated cytological analysis still suffers from extreme class imbalance, very high morphological similarity between cell types, and poor interpretability of most artificial intelligence (AI) models, despite advances in medical imaging and deep learning. We present an interpretable, state-of-the-art framework for BM cell classification based on a large-scale dataset with 171,374 single-cell images annotated by experts with 21 classes. Given the high severity of the class imbalance (originally 3678:1 at times), we created a new subset of 95,865 images from the overall dataset through segmentation and feature extraction. Through this process, overrepresented classes were down-sampled, while specific types of augmentation were applied to underrepresented classes to restore balance, resulting in a ratio of 1435:1. Our recursive segmentation approach, based on CMYK (Cyan, Magenta, Yellow, and Black) and HLS (Hue, Saturation, and Lightness) colour spaces, reliably identifies nucleus, cytoplasm, and whole-cell regions. From these areas, we processed 144 biologically motivated shape, colour, texture, and fractal attributes. We build a hierarchical two-stage classification model named HierEff-S2, where an EfficientNet-B4 backbone assigns each cell to one of six morphological groups. Then, group-specific EfficientNet-B3 models perform fine-grained classification within each group. With 21 classes, this architecture obtains 86.1% accuracy and outperforms other models, including VisionMamba, Ensemble Model, and MobileNet. To promote clinical interpretability, we combine two explainable AI methods to visually highlight cell regions that lead to the model predictions: Grad-CAM and LIME. Using XAI, we report 95.2% correctness at the image level, thus providing biologically meaningful attention to the model.

Keywords: Bone marrow, Hierarchical Classification, HierEff-S2, VisionMamba, Augmentation, Deep learning, Class imbalance, Segmentation, Explainable AI (XAI), Grad-CAM, LIME

Table of Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

The following list describes symbols and abbreviations used in this thesis.

ABE Abnormal Eosinophils

AdamW Adam optimizer with Weight Decay, an optimization algorithm

ART Artifactual Cells

BAS Basophils

BLA Blast Cells

BM Bone Marrow

BMC Bone Marrow Cell

CutMix A data augmentation technique that combines parts of images

DL Deep Learning

EBO Eosinophils

FGC Fragmented Granulocytes

Grad-CAM Gradient-weighted Class Activation Mapping

Gradients Gradient calculations for backpropagation in neural networks

HAC Histiocytes

HierEff-s2 Hierarchical Model

KSC Schistocyte Cells

LIME Local Interpretable Model-agnostic Explanations

LYI Immature Lymphocytes

LYI Lymphocyte Immature

LYT Lymphocytes

MixUp A data augmentation technique that mixes images

ML Machine Learning

MMZ Metamyelocytes

MON Monocytes

NGB Neutrophilic Granulocytes

NGS Segmented Neutrophils

NIF Nucleated Interstitial Fibroblasts

PEB Progenitor Erythroblast

PLM Promyelocytes

PMO Promonocytes

XAI Explainable AI

Chapter 1

Introduction

1.1 Motivation

Identification of BM (Bone Marrow) cells is a crucial step in the diagnosis, treatment, and prognosis of numerous haematological diseases, including leukaemia, anaemia, and various other types of blood disorders. Precise categorisation of BM cells is equally important for clarifying the pathogenic process and competent diagnosis. Nevertheless, BM cell classification is inherently challenging due to the severe class imbalance and the similar morphology of cell types. Although pathologists can differentiate these cells based on subtle morphological differences, this is a time-consuming and subjective process with moderate inter-observer variability, with each factor being more pronounced when dealing with rare cell types [1]. Again, this process needs highly skilled and experienced professionals or experts to identify the issues, especially for rare classes [30]. Moreover, earlier works in automated cytomorphologic classification mainly focused on classifying normal cell types or peripheral blood smear samples [31] ,[32] ,[33]. This BM cell classification dataset suffers significantly from a class imbalance problem, with the most significant ratio of samples favouring the majority class reaching as high as 3600:1, which strongly affects model performance and makes the model substantially biased towards the majority class [2]. Recent developments in ML (Machine Learning) and DL (Deep Learning) offer potential approaches for automating BM cells population classification. However, they have struggled to address an imbalanced dataset and, particularly in the case of rare cell types which are crucial for diagnosing disorders, are severely underrepresented in the training dataset, making it difficult for the model to properly learn to discriminate features for these cells [8]. Prior efforts to address the issue of imbalanced datasets using data augmentation, such as geometric transformations and colour jittering, have led to some improvement, and these approaches still have limited effectiveness in addressing severe class imbalance [2]. Additionally, the proper segmentation of cell structures like nucleus and cytoplasm is critical, as they are significant features in the process of classification [7]. However, the actual small or rare cell classes that are morphologically similar to some of the more common ones are among those that state-of-the-art segmentation methods fail to identify correctly [9]. Another significant challenge is the limited number of annotated data, especially for rare cell types like ABE (Abnormal Eosinophils) and KSC (Schistocyte Cells) [35]. It is challenging to obtain a sufficient number of these cells due to their rarity in clinical samples, which leads to the data sparsity issues [8]. Recent

datasets, such as the BoneMarrowWSI-PediatricLeukemia dataset [6] and BaMBo dataset [2], have begun to tackle this problem. However, an absence of strongly labelled datasets having enough high-quality images for rare cell types hinders the construction of strong models in BM cell classification. The primary objective of this paper is to address the challenges of class imbalance, segmentation, data augmentation, and model interpretability in automated BM cell classification and XAI for validation. In this paper, we present models where BM cells are first classified into coarse morphological groups, and then fine-grained classification further categorises samples into individual identity types. This hierarchical strategy aims to reduce classification complexity by simplifying it, thereby improving function approximation and rare-class detection [9]. Using cutting-edge models such as EfficientNet-B4 [10] and EfficientNet-B3 as hierarchical encoders, this work proposes a robust and interpretable system that not only addresses the extreme class imbalance problem using segmentation and augmentation but also enhances diagnostic accuracy for rare yet clinically significant diseases.

1.2 Research Challenges

Accurate characterisation of cell populations in the BM is critical for the diagnosis of several blood diseases, including leukaemia and anaemia. Nevertheless, it addresses a challenging research problem, which is that the BM cells comprise several types and exhibit multi-class imbalance between the classes. This imbalance made it challenging to identify rare cell types that are commonly important for the diagnosis of life-threatening diseases [36]. For example, some rare cell subtypes, such as ABE (8 images) and KSC (42 Images), are substantially undersampled in datasets, making it difficult for machine learning models to identify them successfully [2]. Specifically, those few classes are unfortunately commonly misclassified due to the use of a class-biased model where the majority classes overwhelm the dataset. Most current methods for classifying bone marrow cells have attempted to address these issues. [2], [4], [7], [8], [9]. Even if some solutions, namely data augmentation and image segmentation are available to enhance classification performance. These models remain challenging to implement. First, many models are considered as "black boxes", which is particularly critical for clinicians who must have confidence in the AI's reasoning [18], [19]. Moreover, distinguishing between cells with visually similar morphologies at the fine-grained level remains challenging for most models, especially under conditions of class imbalance and sparse data [9]. At the same time, despite these, such gaps clearly indicate the necessity of an automated classification system that not only addresses rare-class learning but also provides clinically interpretable results. In this paper, our objective is to surmount these difficulties by proposing a hierarchical deep learning model that can manage class imbalance and morphological similarity, while also providing clinical interpretation of the model's prediction. This way is used to mitigate the class imbalance by employing a two-stage classification procedure named HierEff-S2, in which similar sub-cell types are first clustered and then fine-tuned within the clusters. We simplify the problem [21], which allows the model to better differentiate between common cell types that may be rare in some samples, based on other pre-clinical or clinical information and the interpretability of the classification system.

1.2.1 Extreme Class Imbalance

Class imbalance is a frequent problem in BM cell classification. For instance, the rare classes ABE and KSC have at a minimum 8-42 samples, while some of the standard cell types, NGS (Segmented neutrophil) and LYT (Lymphocytes), contain around 30,000 samples [1]. Bone marrow hematopoietic stem cells (HSCs) have 21 classes in total. So, this imbalance significantly degrades the model's ability to learn informative features from rare classes, leading to skewed predictions and inferior performance on the rare cell types essential for disease diagnosis. Additionally, there are morphological analogies between certain cell types, which further complicate their classification. For example, Immature Lymphocytes (LYI) and LYT may look remarkably alike and can be hard for a pathologist to differentiate. To alleviate class imbalance, various methods have been developed, including data augmentation, class weighting, and hierarchical classification. The objective of these approaches is to boost model generalisation and classification performance, particularly for rare cell types that are critical in the diagnosis of haematological disorders [2], [37], [38]. In this study, we use HierEff-S2 that categorises cells into frequency based population to classify cell types. This strategy mitigates the effect of imbalance by giving more intensive attention to smaller subsets during training, thereby enhancing accuracy and ensuring that lower-frequency cell types are treated reasonably when classifying [9]. In this work, we address the class imbalanced problem on a large scale (i.e., an extremely high imbalance rate) and aim to develop a robust and automated system for bone marrow cell classification including XAI. This system will help clinicians diagnose diseases, including leukaemia and anaemia, more reliably and effectively.

1.2.2 Morphological Similarity

The typical morphology among the various cell types found in bone marrow is another important factor to consider when classifying. BM cells are generally homologous, and in terms of structure, dimensions, and internal content, the two kinds of these morphological cell types are very similar, which confuses not only pathologists but also machine learning classifiers [10]. Thus, LYI may be confused with LYT and Metamyelocytes (MMZ) with Myelocyte (MYB). Such subtle morphological differences are often not accounted for in models with limited training data or those that fail to capture shape at a fine scale. Although traditional models utilise expert knowledge and manual feature selection to identify differences between them, deep learning models struggle to differentiate closely related cell types when there is insufficient annotation data. Although U-Net and Fully Convolutional Networks (FCN) [12], [19] have yielded promising results in segmenting important objects such as the nuclei and cytoplasms, they still have some difficulties with rare cells and ambiguous cell morphologies, especially when trained on imbalanced datasets. Recent work, however, demonstrates that Siamese Networks fed with image pairs as input can successfully distinguish morphologically similar cells by learning subtle yet dis-

tinctive changes in shape and texture can create misclassification [9]. In addition, hierarchical classification, which involves categorising cells at a higher order level and then detailing each category, has been effectively applied in modelling morphological similarity. This method simplifies the classification problem by decomposing a complex problem into simpler sub-problems. By considering the morphological similarity between bone marrow cells in a combined manner of segmented refinement and using HierEff-S2, we enhance model discrimination of visually similar cell types, consequently improving diagnosis accuracy.

1.2.3 Feature Extraction

In the BM cell classification process, feature extraction is one of its significant stages to extract meaningful features from raw cell images for the purpose of classifying diverse types of cells. Here, 144×144 features comprising a full field of view (whole cell), nucleus and cytoplasm are extracted. The features contain morphological properties, shape, size, convexity, compactness and elongation to encode the structural properties of a cell. Moreover, colour-based features are extracted from the R, G and B channels as well, such as mean, variance, and entropy that we can use to capture the distribution of colours in the cell [39]. Additionally, texture features are computed using the Grey-Level Co-occurrence Matrix (GLCM), which signifies patterns such as contrast, dissimilarity, and correlation that characterise the surface texture of cells. Lastly, fractal dimension measures the complexity of the cell, mostly in terms of its structures, which helps in achieving shape discrimination of different kinds of cells. The region-first-order features are organised as a 12×12 region-attention embedding matrix that preserves the spatial and regional information, allowing the model to exploit both local and global characteristics of images for better classification.

1.2.4 Clinical Applicability

The organisation of bone-marrow cells is an essential clinical procedure that can be automated. Bone-marrow analysis is a basic diagnostic tool used in the evaluation of haematological disorders. However, this process of identifying through raw images are time-consuming, labour-intensive and prone to high inter-observer variability when manually annotated. Diagnosis can be automated through systems like the one discussed here, thus alleviating the burden on clinicians and eliminating preventable mistakes. This is especially crucial when rare cell phenotypes of clinical interest but not easily distinguishable are involved, due to their low occurrence in training data. To solve the problem of class imbalance and morphological similarity, we used segmentation, data augmentation, and HierEff-S2 to achieve a accurate result in detecting even rare cell types. The model also offers interpretability through XAI, such as Grad-CAM and LIME to allow clinicians to get validation/confirmation that drives the decisions made by the model. This level of transparency is critical to creating trust in the AI-aided diagnostic instruments.

1.3 Research Objectives

The main goal of this study is to create an automatic bone marrow cell classification system that can handle a strong class imbalance and morphological similarity, which makes it difficult to differentiate between types of BM cells. In our experiments, we used a large dataset, consisting of 21 different cell classes, whose ratio could be up to 3,600:1 [2]. To address this significant imbalance in the classes, the study used state-of-the-art methods, such as segmentation, feature extraction, data augmentation which enhanced the models with fewer cell types that are hard to find and detect. HierEff-S2 and VisionMamba again utilise morphological similarities between cell types so that classification is performed reliably on common and unusual classes. Because the accuracy alone is not sufficient to ensure clinical interpretability, so we applied XAI, namely Grad-CAM and LIME, to ensure validation to every decision, which helps pathologists to verify the results. This proposal is to create an automated system to help pathologists provide precise time-saving diagnoses of haematological diseases like leukaemia and anaemia.

1.4 Framework Overview

The approach is built on the connection of traditional cytological knowledge with state-of-the-art deep learning and XAI methods. It works like this: Create a feasible, understandable, mostly accurate and cost-effective way that can be tailored and integrated into clinical diagnostic systems [24]. With this goal in mind, we will conduct research to increase the precision of automated classification, as well as provide a reference to XAI-directed diagnostic systems that should be used to help haematologists in making clinical decisions.

Data Reducing: Down-sampling or major classes, up-sampling minor classes, to decrease variation in and imbalance between the classes.

Segmentation: Recursive segmentation by colour-space transformations to consistently isolate nuclei, cytoplasm and entire cells, in the presence of noisy or non-uniform staining.

Feature Extraction: Computation of various biological features: shape, colour, texture, fractal complexity, over anatomical regions, arranged into a homogeneous region-attention embedding.

Hierarchical Classification: This is a two-step deep-learning framework, with a first step subdividing cells into broad morphological groups, and a second step making predictions about subclasses by specialised models.

XAI: Use of both Grad-CAM and LIME to explain evolving visual parts that contribute to the model prediction, leading to clinical trust and validation.

Evaluation and Validation: Thorough testing of the model's robustness, ability to capture features, and accuracy of explanation, with particular emphasis on rare classes, along with solid interpretation.

1.5 Contributions

The main contributions in this study are:

- A complete automated BM cell classfication with 21 known cell types.
- A recursive segmentation algorithm with a biologically inspired quality-control module robust to the effects of staining artefacts.
- An architecture-driven cytological feature embedding that generalises between human-engineered cytological features and deep-learning designs.
- Hierarchy classification approaches that enhance the accuracy and generalisation of minority classes.
- Validation of model output decision regions created with multi-perspective XAI (Grad-CAM and LIME).

Chapter 2

Related Work

Several problems face the development of the applicability of automated BM cell classification, including difficult morphological patterns, extreme imbalance of classes, the small number of examples of a single type (especially of rare ones), and the need for sensible models that can be directly used by clinicians. This chapter will focus on classical and recent approaches in five major areas: early popular works on BM datasets and BM cell classification, BM segmentation, feature-based methods and augmentation, imbalanced mitigation strategies, architectural or hybrid models, and XAI techniques. We thoroughly elaborate on each work to illustrate its significance, drawbacks, and its relation to our framework.

Paper	**Used Data**	**Model**	**Pretrained**	**Limitations**
[6]	Full dataset, 7 classes	DenseNet121	Yes	Used 7 classes only, ignore rare classes
[4]	30,837 images, 21 classes	Siamese NN (Triplet Loss)	Yes, fine-tuned	Class imbalance unclear, no augmentation, rare classes not reported
[7]	Full dataset, 21 classes	Region-Attention Embedding + DL	Likely pre-trained	No clarity on 171k dataset usage, imbalance handling not discussed
[5]	Full dataset, 21 classes	Siamese NN + GAN	Yes	GAN augmentation unclear, minority class performance not reported
[8]	40,000 images, 8 classes	MTLSO-BMCC	N/A	Methodology clear, handles imbalance effectively
[2]	4,500 images, 3 classes	CNN	Likely pre-trained	Small dataset, poor performance, no augmentation
[9]	Full dataset, 21 classes	RegNet_Y_32gf / ResNeXt-50	Yes	Extreme imbalance not addressed, rare classes likely failed

Table 2.1: Bone Marrow Cell Classification (Key Datasets, Models, Pretrained Use, and Limitations)

Paper	Rare Class Handling	Augmentation	Segmentation / Feature Extraction	Dataset Balancing / Reduction
[2]	Ignored / not reported	No	None	No balancing, small dataset (4.5k)
[4]	Rare classes unclear	No	None	No balancing, full dataset used
[5]	Rare classes likely ignored	GAN-based augmentation	None	No clear balancing, full dataset used
[6]	Rare classes partially handled	Yes (all classes)	None	No reduction, full dataset (7 classes)
[7]	Not explicitly handled	No	None	No clear balancing, full dataset (21 classes)
[8]	Handled effectively	No	None	Balanced dataset, but no rare-class augmentation
[9]	Extreme imbalance not addressed	Yes	None	Full dataset, no reduction, rare classes likely failed

Table 2.2: Comparison of Rare Class Handling, Augmentation, Segmentation, and Dataset Balancing Across Bone Marrow Classification Studies

Well-annotated datasets are the foundation for training in and development of BM cytology mentioned in 2.2. The baseline dataset [1] used by Matek et al. [2] remains the benchmark, especially reducing the dataset 171,374 single-cell images to 4,500 images and also trained their model on only 3 classes among 21. Their performance demonstrated that CNNs can classify among dozens of BM cell morphologies at the human level. However, they did not focus on interpretability or robustness to rare classes. This classification is a form of relaxation, as simplified in some literature. For instance, one IEEE paper with the dataset into a binary classification of normal and abnormal cells [3]. While this technique achieves high scores, it is performed at the cost of anatomical detail and clinical relevance. Similarly, the works of Shaikh et al. were based on a Siamese network [4] that employs the pairwise similarity loss to overcome data scarcity (initially written about 21 classes). Although efficient for few-shot learning, such a framework only deals with a subset of cell classes and does not achieve end-to-end classification on 21 cell types and uses 30,837 samples. Khafaga et al. [5] advocate an Ocotillo-based optimisation algorithm for hyperparameter tuning and feature selection in BM models, achieving close to state-of-the-art accuracy. But rare classes are likely ignored without any

clear balancing and also used GAN, which is not efficient in the medical sector, as it may produce imaginary images or outputs. On the other hand, it does not learn a canonical location map according to anatomical hierarchy, nor is the learning problem framed in terms of hierarchies or interpretability. Attentive transfer learning is investigated in [6], where CNN backbones (e.g., DenseNet, ResNet, Xception) are implanted with attention modules to highlight discriminative parts. Although this method enhances selective focusing, it typically focuses on a truncated set of classes (7 classes) and lacks further interpretability. Tarquino et al. [7] combine segmentation, feature engineering, and embedding fusion. While engineered embeddings improve transparency, the deep learning backbone (Xception/ResNet50) still acts partially as a "black box. It takes hand-crafted features from the nucleus and cytoplasm after they are segmented, which are then fused with the deep feature maps to produce region-aware embeddings. This hybrid approach highlights the tradeoff between explainability and performance, and has a profound impact on how we design region-attention embeddings, but the complexity makes it hard to rely on for real-time output in medical sectors. But again, B, Hardwaj et al. [8] integrate multimodal architectures using snake-guided optimisation for feature and model refinement on 40,000 images and 8 classes. Their approach aims to maximise accuracy by incorporating fusion and metaheuristics, but does not include an interpretable structure or segmentation in its decision-making process. Glüge et al. [9] compare training strategies, including class weighting, oversampling, and learning rate scheduling, on BM images. They claim that poorly configured oversampling tends to overfit minority classes and likely fails when handling rare classes. Thus, the authors encourage a mindful balancing and architecture-oriented augmentation. Morphogo, proposed by Lv et al. [10], is a clinically used commercial cytology system. It can achieve high diagnostic accuracy on real-life bone marrow images. Even though some methods are black boxes, with no publicly available interpretation and segmentation elements, this field is enriched by segmentation-related databases. For instance, consider the 185 bone-marrow specimen images on BaMBo [11], which have been expertly labelled by professionals in the field to precisely categorise the cells and also enable extensive morphological analyses. Additionally, BoMBR [12] utilises 201 annotated images to focus on reticulin fibre segmentation. Although these datasets refer to tissue-wide tasks and not to single-cell cytology, they demonstrate that segmentation-based studies of haematological imaging receive increasing attention. Finally, early methods of analysis, such as the BM smear detection dataset (230 labelled smears) [16], demonstrate the first attempts to implicate cell detection and classification at a small scale, providing an invaluable examination of the complexity of the cytology BM pipeline.

2.1 Segmentation and Feature-Driven Methods

Targeted morphological analysis relies on segmentation. Specifically, Tarquino et al. [7]segment the nucleus, cytoplasm, and cell boundaries to generate region-specific descriptors, which are then combined with CNN embeddings. Coupling the CNN-based model with structured anatomical features proves to be particularly useful for classification. Moreover, more generalised segmentation-classification networks, such as HoVer-Net, which was initially developed for histology, propose holistic nucleus

labelling protocols that utilise horizontal/vertical distance maps in the segmentation and classification processes [34]. These protocols have inspired a trend of integrated pipelines across various imaging fields. Conversely, the most intuitive layer, which is an explainability network (in our case, segmentation-agnostic) suggests that traditional attribution methodologies restrict interpretability for various highly organised, dense outputs. This fact further indicates that the segmentation-attributed explanation should be an integral part rather than an independent post-processing step. This paper is a culmination of these principles: we have implemented the multiple rounds of recursive segmentation utilizing CMYK and HLS color-space transform integration, with parameters re-thresholding at each step, alongside the biologically reasonable assumptions, such as nucleus-to-cell ratio and single domain mask component to achieve robust region masks in the staining-varying setting and this is a fundamental requirement that the current BM cell Classification models do not widely address.

2.2 Rare class and Imbalance Mitigation

BM cell cytology is riddled with a severe amount of class imbalance, which is difficult to rectify in naive ways. The binary-collapse [3] approach circumvents the imbalance issue, but at the cost of losing diagnostically crucial distinctions. Optimisation-based methods, including Ocotillo [5] and snake fusion [8], optimise feature sets or model parameters to improve minority-class accuracy. However, these methods treat the imbalance as a tuning problem rather than the inherent structural issue. They do not reconfigure the space for classification and/or explicitly make room for rare classes. Glüge et al. [9] demonstrate that navigating oversampling or class weighting may result in overfitting, but propose using balanced augmentation, curriculum scheduling, and carefully tuning the hyperparameters to enhance minority-class robustness. To delve deeper, our hierarchical approach first classifies into morphological superclasses (i.e., high-frequency, medium, rare), allowing each subclassifier to become specialised. Such a configuration suppresses the rivalry between rare and frequent classes, thus improving the internal representation capability and mitigating class confusion.

2.3 Architectures and Hybrid Strategies

Deep learning architectures, conventional feature-based methods, and combination strategies between both methods have been crucial in making automated classification of bone marrow cells. Initial research has been done on traditional image processing and manual feature detection, such as nucleus-plasma segmentation, shape, and texture-based features to classify cells [1], [3], [36], [38]. These methods could give interpretable features, although they tended to have difficulties with complicated morphological variations and overlapping cells. CNNs have enhanced the accuracy and strength of classification. Matek et al. [2], [31] showed recognition of human-level blast cells in acute myeloid leukaemia with the help of CNNs, and Kimura et al. [33] and Lv et al. [10] used deep learning to differentiate myelodysplastic syndromes (MDS) and other haematological disorders. Other architectures, such as the Morphogo [10] architecture and the HoVer-Net architecture [34] focus on concurrent nuclei segmentation and classification, which makes it possible to

learn several tasks simultaneously in histopathology images. Hybrid strategies are now hybrid, integrating handcrafted features with DL embeddings that have already been improved in terms of performance as well as interpretability. Tarquino et al. [7] proposed the implementation of engineered feature embeddings and deep neural networks that enhance the model transparency, shape, colour, texture, and fractal features. Equally, Fazeli et al. [35] and Guo et al. [36] combined visual representations and feature-based descriptors to deal with morphological complexity and class imbalance. Attention-based transfer learning in fact, has also been investigated, as in Alshahrani et al. [6] where attention is applied to important parts of bone marrow images to utilise existing pre-trained networks. In order to enhance the performance of the models, metaheuristic optimisation has been used and hybrid frameworks. As an example, Khafaga et al. [5] adopted an Ocotillo optimisation-based deep learning model, whereas transfer adaptation with optimisation algorithms was described in Scientific Reports [8]. Such hybrid methods are best in selecting features and optimising model parameters at the same time, thereby enhancing heterogeneous dataset classification. Explainability and clinical relevance have also been given importance in recent work. Explainable AI methods, like Grad-CAM, have been applied to visualise significant diagnostic areas, which allows model outputs to be viewed in a pathologist-style interpretation. This is in line with the current trend of combining both the feature-based and deep learning approaches to develop hybrid systems that are more accurate, transparent, and clinically applicable [7], [18], [35]. The current literature suggests that hybrid approaches that include DL and handcrafted features are favoured in the strong and interpretable BM cell classification. These methods combine the strengths of both the traditional feature engineering and modern deep learning models and open the pathway towards accurate haematological diagnostics that are explainable. On the other hand, Several recent studies have explored transfer learning for hematologic image classification. For instance, the study [41]fine-tuned pretrained CNNs to classify leukaemia images, showing that generic feature extractors can be effectively adapted to medical imaging tasks. However, most of these approaches focus on broad classification and do not address the challenge of distinguishing fine-grained subtypes or recognising rare class challenges that are central to our hierarchical and rare-class-aware framework. Guo et al. [42]developed a version of MobileNetV3 enhanced with an attention mechanism and multi-stage feature fusion to classify eight types of breast cancer pathology images. Their approach achieves high accuracy while keeping the model lightweight and efficient, making it suitable for practical medical applications. Also MobileVit a lightweight vision transformer, for classifying bone marrow blood cells. Their model [43] achieved approximately 97% accuracy without relying on transfer learning, surpassing traditional CNNs like ResNet and DenseNet in classification performance. This indicates that newer transformer-based models can perform well without transfer learning, offering a promising direction for bone marrow cell classification.

2.4 XAI: Visualisation, Attribution, and Trust

Clinical deployment demands transparency. Bukhari et al. [18] present an extensive review of XAI approaches in medical image analysis, discussing how methodology combinations (saliency, perturbation, concept-based) provide better explanations. To date, Grad-CAM (Selvaraju et al.) [9] is a cornerstone for saliency. It calcu-

lates gradient-weighted class activation maps for visualising discriminative regions in images. In haematology, Zhou et al. [20] applied Grad-CAM to detect nucleus-level differences in AML subtypes, comparing model attention with the knowledge of pathologists. The MDPI ALL classification CAD system [21] combines segmentation and Gradi-CAM-based explanation to produce understandable outputs that are insightful for white blood cell classification, demonstrating that both segmentation and saliency can complement each other. In the "Pathologist-like Explanations" architecture [22], explanation generation modelling is conceived under the framework of human judgments, focusing on cell regions in a clinically interpretable manner rather than relying on saliency maps. Segmentation-explainability problems are also mentioned in the survey by [23], which claims that dense attribution (for segmentation) frequently does not work for the naive application of classification-based methods.

Chapter 3

Dataset

This study started with the collection of a Bone Marrow Cell dataset [1] shown in . The image samples were then recursively segmented to identify regions of interest, and 144 morphological and texture-based features were obtained. The extracted features were normalised to pre-process the features, making them uniform across the data. A subset of the data set was developed to balance the imbalance in classes according to the mask quality and feature. Uncommon classes with small samples were singled out and selectively increased to improve its representation. Three models, HierEff-S2 and a Vision-Mamba and an Ensemble model were then trained using these features, and two pre-trained convolutional structures, MobileNet V2 with attention mechanism fine-tuned on an augmented dataset to benchmark performance. Lastly, Explainable Artificial Intelligence (XAI) techniques were used to visualise and interpret model predictions in a way that the learned representations were consistent with domain-relevant biological patterns and overall model reliability was assessed 3.1.

3.1 Dataset Processing

The performance of automated cytological classification strongly depends on the quality, dispersion, and pre-processing of the image data used. In this work, we used a subset of a pre-processed bone marrow dataset, focusing on a small yet carefully curated part of a large, publicly available BM cytology dataset reported by Matek et al. [3]. The complete dataset consists of 171,374 single-cell images, with each cell manually labelled by an expert as one of the 21 morphological and clinically relevant hematopoietic classes [1]. These range across the entire gamut of erythroid, myeloid, lymphoid, and progenitor cell subsets. All samples originated from 945 patients, thus comprising a broad spectrum of morphologic and staining divergence. All images were taken using 40× oil immersion bright-field microscopy and stained following either the May-Grünwald-Giemsa or the Pappenheim method, thereby ensuring a uniform colour contrast within all recordings. All images are resized to 250 × 250 pixels and can be embedded easily into convolutional and segmentation pipelines. The dataset is vast and rich, yet it has significant inter-class imbalance, which makes model generalisation a challenging task. Namely, cell types have frequencies spread over three orders of magnitude, with the most common class covering almost 29,424 samples and the rarest type, ABE, numbering only 8. This results in a starting maximum imbalance ratio of about 3,600:1, which has

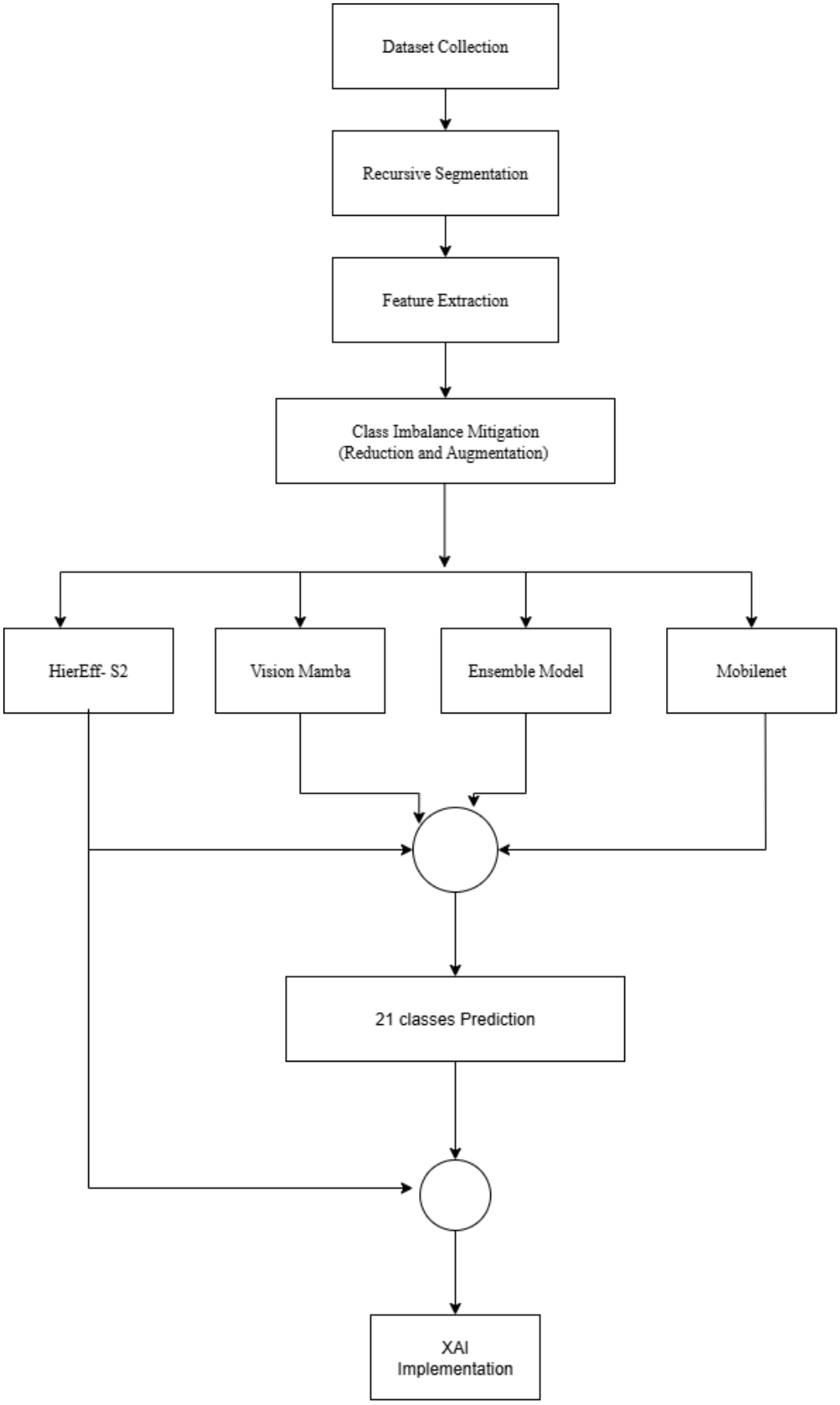


Figure 3.1: Workflow

been demonstrated to strongly bias baseline classifiers towards the majority class (primarily through cross-entropy loss).

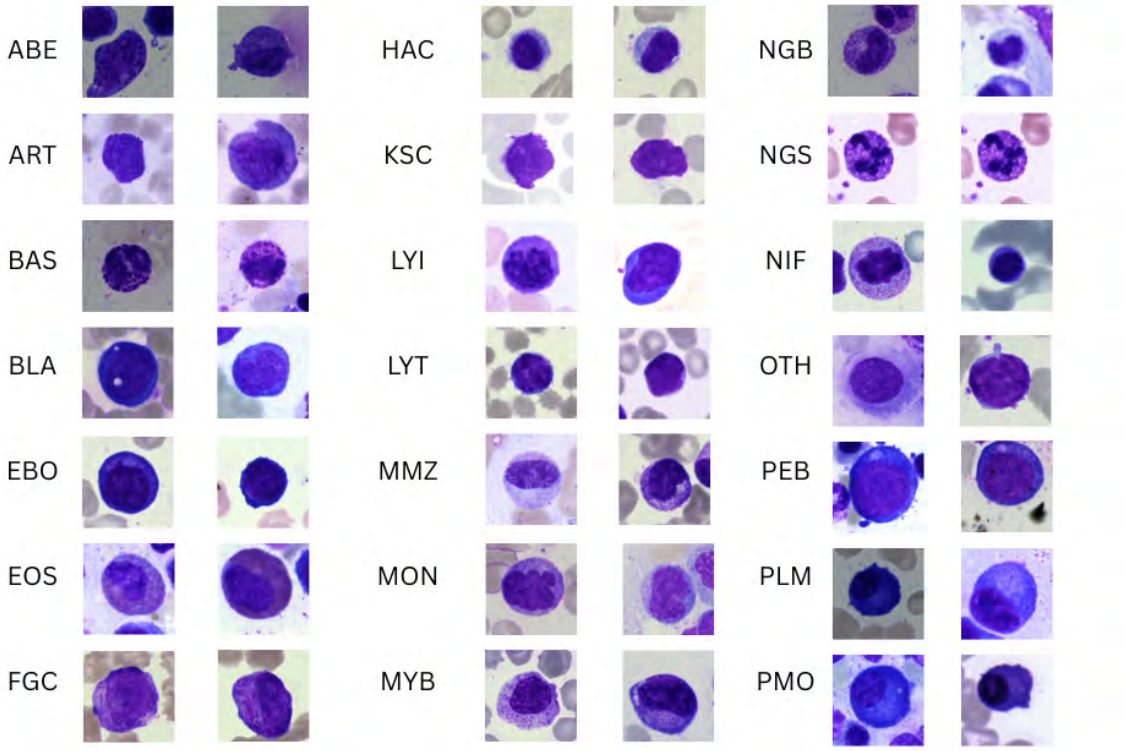


Figure 3.2: 21 different classes on BM

3.1.1 Segmentation

Segmentation is a critical step in the bone marrow image analysis pipeline, forming the foundation for accurate feature extraction . The primary goal is to isolate three biologically relevant regions from each microscopic image: the nucleus, cytoplasm, and the whole cell. To ensure robustness against variations in staining, illumination, and cellular morphology, a recursive segmentation approach was adopted [53]. Before segmentation, each image undergoes color balancing to standardise illumination and enhance visual contrast among cellular components. A simple white-balancing algorithm normalizes the mean intensity of each color channel using an adjustable intensity factor, ensuring consistent color representation across the dataset [49]. Following preprocessing, the color-balanced image is transformed into multiple color spaces to improve discriminability between cellular regions. The CMYK color space highlights variations in pigment absorption, particularly in the magenta and black (K) channels, which correlate with nuclear staining, while the HLS color space provides hue and lightness information that aids in distinguishing cytoplasm from the background [57].
Nuclear segmentation is performed by enhancing contrast between the K and M channels (black and magenta) to emphasize regions rich in nuclear dye. A soft segmentation map is generated using the difference between these channels, followed by Otsu thresholding to produce a binary mask. Small regions are removed, holes are

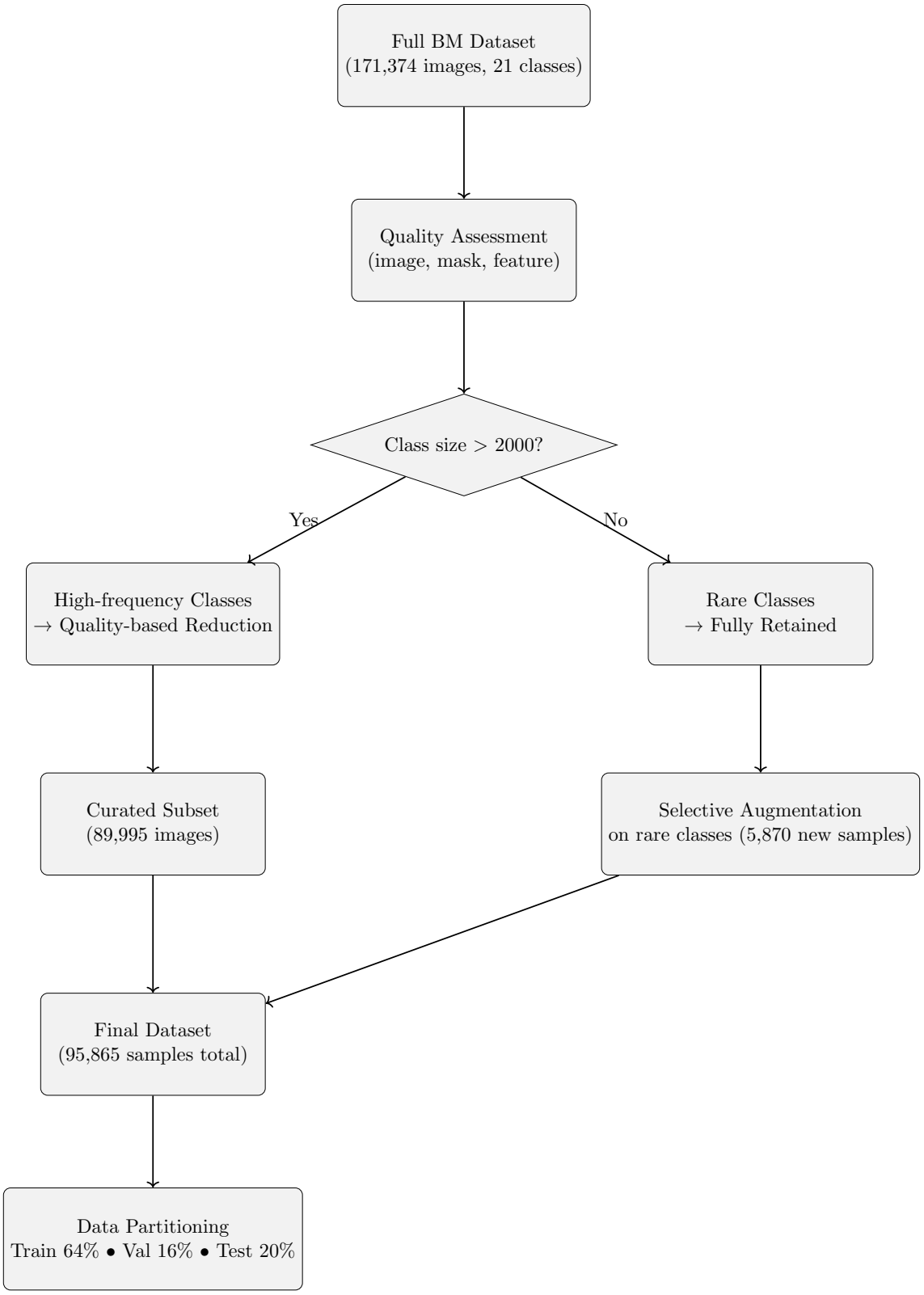


Figure 3.3: Dataset Workflow: Curated Subset, Selective Augmentation, and Final Dataset

filled, and only the largest connected component is retained to ensure morphological accuracy [54]. Segmentation depends on the difference between the yellow and saturation channels. Also, it uses thresholding, which can be either global thresh-

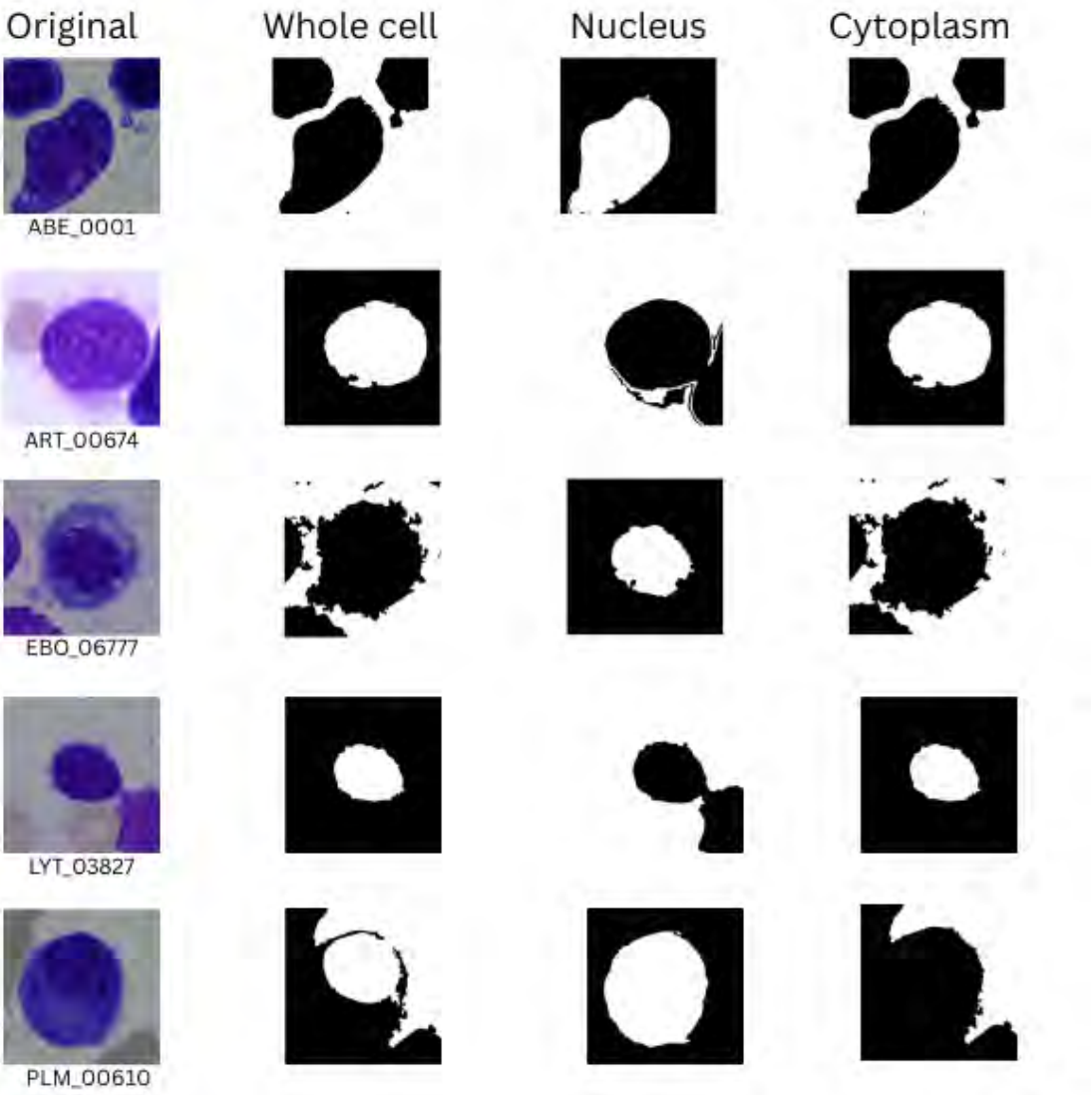


Figure 3.4: Segmentation

olding or adaptive Gaussian thresholding, which is then followed by morphological post-processing and connected component analysis to isolate the main cell body [50]. Cytoplasm is derived by subtracting the nucleus mask from the whole-cell mask using a logical XOR operation, ensuring that cytoplasmic pixels are correctly localized outside the nucleus but within the cell boundary [57],[58]. This is a strategy that is in line with. past research in which cytoplasm is prepared by the subtraction of the nucleus in the case of the former research on whole-cell mask [57].
To handle variability in staining and cell morphology, segmentation is performed recursively with up to three parameter configurations. Each iteration adjusts the color balance intensity and thresholding strategy until satisfactory masks are produced [54]. Iterative refinement of segmentation is a well-established strategy in medical imaging, ranging from auto-context modeling to modern recursive-refinement networks [52],[53]. Segmentation quality is evaluated using region-level checks, including nucleus-to-cell area ratio, component connectivity to ensure single-cell representation, and the presence and proportion of cytoplasmic regions [55]. The masks that do not meet the requirements are re-executed using new parameter sets to enhance the robustness of the mask to various imaging conditions.

Feature Extraction

Following the extraction of the segmentation masks, a reduced number of features are isolated out of each region nucleus, cytoplasm and cell body, which will give a detailed quantitative account of the cellular structure and composition. Shape features include area, perimeter, compactness, roundness, convexity, solidity, eccentricity, and elongation, capturing size, boundary complexity, and morphological asymmetry [57]. Color features are computed from each RGB channel, including mean, variance, skewness, kurtosis, and entropy, characterizing chromatic composition and heterogeneity, which aids in detecting pathological differences in staining uptake or pigment density [49]. Texture features are extracted using the Gray-Level Co-occurrence Matrix (GLCM) at four orientations (0°, 45°, 90°, 135°), with derived metrics such as contrast, dissimilarity, homogeneity, energy, and correlation, describing micro-level spatial patterns in chromatin and cytoplasm. Grayscale intensities are normalized to 64 levels to balance computational efficiency and resolution [51]. Finally, fractal dimension is estimated using the Minkowski–Bouligand box-counting method, capturing boundary complexity and self-similarity; smaller box sizes capture fine structural irregularities, while larger boxes describe global shape, providing a quantitative estimate of morphological intricacy in nuclear and cytoplasmic contours [52],[57].

3.1.2 Selective Augmentation Strategy

In view of the heavy class imbalance within the dataset, we followed an exclusive parameter augmentation approach that concentrated on only the six worst-performing classes. The main goal of this strategy was to augment the proportion of rare cell types in the dataset, but without over-representing those that were already represented. For the rarest classes, as Abnormal Eosinophils (ABE), Schistocyte Cells (KSC) and Fragmented Granulocytes (FGC), we have employed heavy augmentation with a 10x multiplier. This is so that these classes were over-sampled and

increased to ten times the amount of data, which eventually will support the model learn from a wider range of images. For the minority classes, such as Basophils (BAS) and Histiocytes (HAC), these classes were somewhat underrepresented but not to the extent of the very rare classes, so we used moderate augmentation with a 5x multiplier. The 15 classes that were well represented in the data did not receive any augmentation to maintain the natural distribution of these classes.
The augmentation consisted of various geometric transformations in addition to colour jittering, blurring and noise injection. In the case of the super-rare classes, intense augmentations were used: horizontal and vertical flips, 90-degree rotations, and shift-scale-rotate transformations, thus modelling distortions similar to those experienced in pictures of microscopes. Further modifications, such as motion blur, Gaussian blur, median blur and elastic distortions were added to simulate variability in image quality. Colour jittering was also used to vary image brightness, contrast and saturation. These strong augmentation measures made sure that enough representation of the rarest classes was in the training set, thus alleviating the risk of bias in the training set to more common cell types. For the smaller-sized classes, a more conservative augment strategy was used (using horizontal flips, 90 rotation, brightness/contrast change) to generate some synthetic data without overfitting.
In contrast, classes that had a sufficient amount of samples were not augmented to remain true to the distribution. Over-sampling these classes can cause data artefacts and imbalance the model performance.

3.1.3 Augmentation and Dataset Composition Results

After the application of the selective augmentation strategy, the final composition of the data sets was:
Original samples: 89,995 samples (93.88%)
Augmented samples: 5,870 samples (6.12%)
The attempted balanced dataset contains a total of 95,865 samples, which aided in solving the class imbalance problem to a great extent without losing much information from various sectors. The extensive augmentation applied to the rare classes and the moderate augmentation applied to small classes increased the model's sensitivity in detecting rare cell types without altering the natural distribution of more frequent classes.

3.1.4 Data Partitioning

A two-stage stratified random sampling process was used to separate the entire dataset into training, validation, and test subsets that maintain representative class distributions in all partitions and guarantee data independence.

Table 3.1: Dataset Composition and Selective Augmentation Results

Class Group	**Classes**	**Original Samples**	**Aug Applied**	**Final Samples**
Very rare classes	4	162	Heavy augmentation ×10	1,970
Small classes	2	850	Moderate augmentation ×5	5,100
Medium classes	6	8,593	None	8,593
Medium-large classes	3	17,697	None	17,697
Large classes	6	142,498	None, reduced Uses image, mask, and feature quality metrics to select the best samples	62,505
Total Dataset	21	171,343	Selective augmentation only	95,865

Table 3.2: Dataset Reduction, Selective Augmentation, and Partition Summary

Stage	**Description**	**Samples**	**Percentage (%)**
Original Dataset (Matek et al. [3])			
Total Samples	21 hematopoietic cell classes from 945 patients	171,374	100.00
Most Common Class (e.g., NGS)	Maximum frequency	29,424	17.17
Rarest Class (e.g., ABE)	Minimum frequency	8	0.005
Initial Class Imbalance Ratio	–	∼3,600:1	
Curated Subset and Augmentation Strategy			
Original (Unchanged) Samples	Real high-quality images retained	89,995	93.88
Augmented Synthetic Samples	Added for six rarest classes	5,870	6.12
Final Curated Dataset Used	**(Original + Augmented)**	**95,865**	**100.00**
Post-Curation Class Ratio	Maximum-to-minimum frequency balance	∼1,466:1	
Final Dataset Partition (Stratified Sampling)			
Training Set	Used for model optimisation	61,354	64.00
Validation Set	Used for tuning and checkpointing	15,338	16.00
Test Set	Held out for final evaluation	19,173	20.00
Total (Used in Study)	–	**95,865**	**100.00**

Chapter 4

Models and XAI

4.1 HierEff-S2 (Hierarchical Model)

We follow a transfer learning approach based on models pre-trained on large-scale visual recognition tasks (ImageNet, 14M images, 1000 classes). The vector of network weights, which captures ascent on a hierarchy of visual representation learned from diverse natural images, was used as the initial point for our target domain-specific medical image classification task. This approach capitalises on the fact that low- to mid-level visual representations (edges, textures, and shapes) can be transferred across domains; therefore, the low-level feature was fixed, whereas higher-level representations were optimised to bone-marrow cell features.

4.1.1 Rationale and Motivation

In tackling the issue of class imbalance, the two-stage hierarchical classification scheme was used, whereby a classification is done in terms of the frequency distribution. This strategy has a few advantages:

- Breaking down the problem into smaller sub-problems: The classification process is simplified by breaking the problem into more manageable sub-problems.
- Balanced learning process: The method creates equal distributions of groups, giving categories equal learning opportunities.
- Minor class discrimination: Small-category instances should be pooled into special buckets to increase the model's capability to discriminate minority classes.
- Class-specific feature learning: This approach allows learning unique features of classes that have similar sample distributions.

4.1.2 Grouping Methodology

The 21 classes were grouped into 6 large groups according to their population frequency in the dataset:
Group 1 (Very Rare Cells): ABE, KSC, FGC and LYI. Includes rarest cell types with the fewest numbers represented.

Group 2 (Less Populated): BAS, HAC, OTH. Contains cell types with smaller numbers, with slightly more representation than in Group 1.
Group 3 (Medium Population – Type A): PEB, MMZ, NIF, MON. Well-represented classes with similar frequency tendencies.
Group 4 (Medium Population – Type B): EOS, MYB, PLM. Another group of sparsely spread classes.
Group 5 (Intermediates – Polymorphic Populations): NGB, BLA, PMO. Well-represented classes approaching high-frequency categories.
Group 6 (High-Density Population): NGS, EBO, LYT, ART -the top cell types in the dataset.

Stage-Specific Training Parameters

Stage 1 (Group Classification): 25 epochs, initial learning rate $LR = 10^{-3}$, using a higher learning rate at the beginning since it makes it easier to focus on the coarser-grained group classification task.
Stage 2 (Specialist Training): 35 epochs for specialists with a smaller learning rate of 5×10^{-4} to allow fine-grained discrimination in each class.

4.1.3 Optimisation Strategy

We utilized the AdamW optimiser along with a weight decay of 1×10^{-4} for L2 regularisation. The learning rate scheduling used was CosineAnnealingLR, resulting in the smooth annealing of the learning rate from its initial value to near zero during training epochs, thereby avoiding sharp jumps that may affect convergence. To avoid exploding gradients, we applied gradient clipping with a maximum norm of 1.0. The loss function was plain cross-entropy without class weights at the group level, as the grouping procedure inherently deals with class imbalance and creates balanced groups.

4.1.4 Checkpointing and Training Resumption

We implemented a checkpointing system to store the state of our model every 5 epochs, including the model weights, optimiser states, scheduler states, training metrics, and epoch number. We maintained separate checkpoint directories for Stage 1, Stage 2 (per specialist), and final combined models. This allowed us to resume training from any stage, ensuring progress was not lost due to interruptions.

4.1.5 Stage 1: Group Classification Network Architecture and Training

Backbone Network - EfficientNet-B4

In the figure 4.1, The feature extraction backbone is a pretrained EfficientNet-B4 on ImageNet1K, where the initial classification head was removed and replaced with a localizer for high-dimensional features. EfficientNet-B4 was chosen because it uses compound scaling to scale up network depth (29 layers), width (channel multiplier of 1.4), and resolution of input images (default 380×380, adapted 224×224 for efficiency). The backbone processes $224 \times 224 \times 3$ RGB input images with a series of

mobile inverted bottleneck convolution (MBConv) blocks optimised with squeeze-and-excitation.

The EfficientNet-B4 Backbone Architecture

Initial Stem Layer: A 3×3 convolution with stride 2, reducing image dimensions from 224×224 to 112×112 and increasing the number of channels from 3 to 48.
MBConv Block Stages: Seven sections of MBConv blocks with the number of layers in each set as follows:
Stage 1: MBConv1 blocks ($k = 3 \times 3$, expansion ratio 1) on feature maps of size 112×112.
Stage 2: MBConv6 ($k = 3 \times 3$, expansion ratio 6) with down-sampling to 56×56 resolution.
Stage 3–7: Progressive MBConv6 blocks with squeeze-and-excitation (SE) ratio = 0.25, down-sampling spatial dimensions to 7×7 and up-scaling the channel number to 1792.
Final Convolution: A 1×1 convolution expanding channels to 1792 dimensions.
Global Average Pooling: Reduces the $7 \times 7 \times 1792$ tensor to a 1792-dimensional feature vector.
The pretrained weights provide strong initial feature representations learned from 1.2 million ImageNet images and are suitable for transfer learning to the bone marrow cell domain. During Stage 1 training, all backbone parameters were kept learnable to enable domain-specific adaptation.

Group Classification Head

Classification head: It converts the 1792-dimensional feature vectors extracted from the backbone into 6-class group-wise predictions as follows:

Input Dropout Layer ($p = 0.3$)

This was applied to the 1792-dimensional feature vector so as to counteract the co-adaptation of features while also improving generalisation. During training, 30% of the feature activations are randomly set to zero, encouraging the network to rely on multiple redundant features rather than memorising specific patterns.

Fully Connected Layer 1 (1792 $\rightarrow$ 256)

This linear layer diminishes the feature dimensionality from 1792 to 256:

$$y = Wx + b, \quad W \in \mathbb{R}^{256 \times 1792}, \quad b \in \mathbb{R}^{256}.$$

This bottleneck transformation forces the model to learn compact, discriminative representations suitable for group-level classification.

Batch Normalisation (256 features)

Normalises activations across the batch dimension to stabilise training. For each mini-batch, it computes the mean μ_B and standard deviation σ_B, and applies:

$$\hat{x} = \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}, \quad y = \gamma\hat{x} + \beta,$$

where γ and β are learnable scale and shift parameters. Batch normalization facilitates higher learning rates, mitigates internal covariate shift, and enchances convergence stability.

ReLU Activation (in-place)

Applies the rectified linear unit function element-wise:

$$f(x) = \max(0, x).$$

The `inplace=True` operation overwrites input tensors to save memory, which is important for larger batch sizes.

Second Dropout Layer ($p = 0.15$)

The dropout rate is reduced due to the lower-dimensional feature space, still providing regularisation while allowing more information flow towards the final classification layer.

Output Fully Connected Layer ($256 \to 6$)

The final fully connected layer projects the 256-dimensional features to 6 logits, one per group. An activation function is not employed here, as the logits are directly consumed by the cross-entropy loss, which internally performs softmax normalisation.

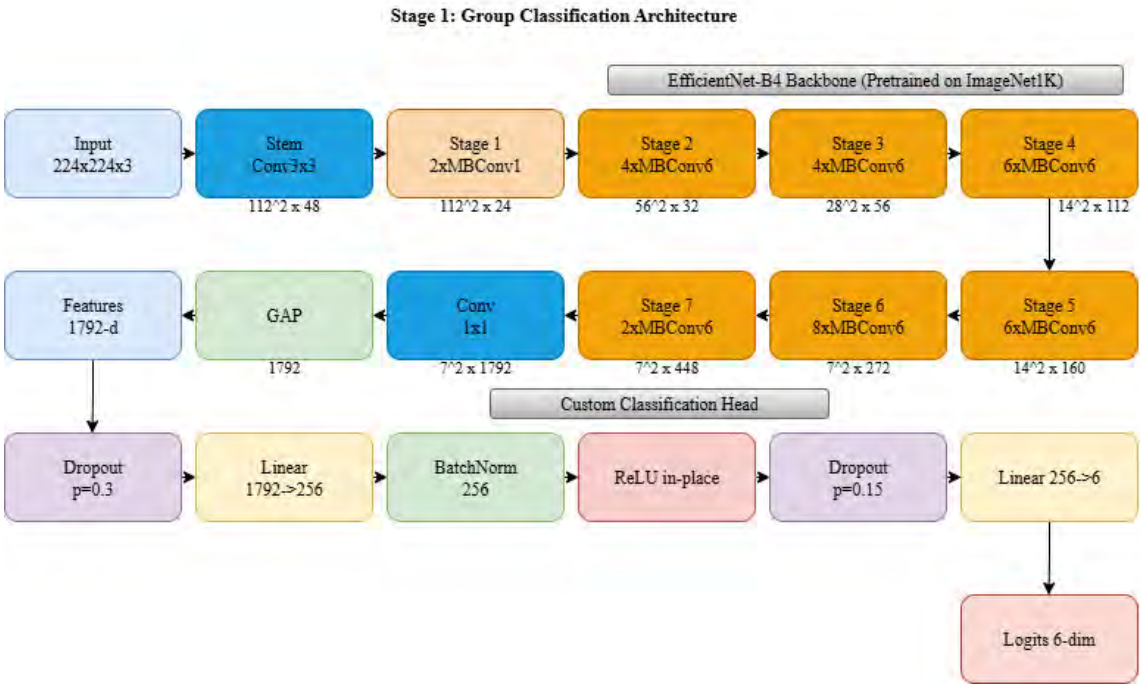


Figure 4.1: Stage 1: Group Classification Architecture

4.1.6 Architecture Dimensions and Information Flow

- Input: $224 \times 224 \times 3$ RGB image block
- After backbone: Batch $\times$ 1792 feature vectors
- Post FC1 + BN + ReLU: Batch $\times$ 256 activated volume
- Ends: Batch $\times$ 6 unnormalized logits

During forward, each image is passed through a backbone network to obtain multi-scale features capturing cellular morphology, texture, and colour information. The classification head applies a mapping between these features and group-specific representations, with cross-entropy loss against one-hot encoded group labels. The softmax function transforms logits into probability distributions during testing, with the arg max operation yielding the predicted group.

4.1.7 Stage 2: Specialist Network Architecture and Training

After conducting group-level classification in stage 1, it moves to the next stage of utilizing six specialist networks which are specialized to do classification within its assigned group. This stage moves from coarse-grained group discrimination to accurate identification at the class level with each specialist optimized for the particular morphological characteristics and class distribution of the group it is targeting.

Backbone Network - EfficientNet-B3

In the figure 4.2, Each member specialist uses a distinct EfficientNet-B3 architecture that has been shown to strike an appropriate balance between computational efficiency and strong feature extraction capability. The architectural down-scaling from EfficientNet-B4 to B3 is strategically designed to reduce computational load without sacrificing accuracy. Since six parallel task-specific specialists are required, adopting EfficientNet-B3 instead of B4 significantly decreases the total memory footprint and allows efficient training on GPUs with limited VRAM capacity. This enables simultaneous fine-tuning of multiple specialists without exceeding memory constraints. As each specialist network tackles a smaller sub-problem (discriminating between 3–4 classes) rather than the original 21-class classification, the reduced model capacity of EfficientNet-B3 (compound coefficient $\phi = 1.2$) is better aligned with task complexity and mitigates overfitting on smaller, group-specific datasets. The compact architecture further encourages learning of discriminative intra-group features instead of memorization, which is particularly valuable when handling rare cell types with limited samples. Structurally, EfficientNet-B3 follows the same design principles as EfficientNet-B4 beginning with a 3×3 convolutional stem, followed by seven stages of MBConv blocks with squeeze-and-excitation SE mechanisms. The network progressively downsamples to 7×7 feature maps, expanding the channel depth to 1536. A global average pooling layer then produces 1536-dimensional feature vectors used by the specialist classification heads. Similar to the group classifier, the specialist backbones are initialized with pretrained weights from large-scale natural image recognition tasks and undergo domain adaptation during Stage 2 training to

specialize in fine-grained discrimination among groups of morphologically similar cell types.

Specialist Classification Head Architecture

In the figure 4.2, The specialist classification heads share the same design principle as the group classifier but are tailored for specialist-specific needs. Each head operates on 1536-dimensional feature vectors as follows:
Input Dropout Layer ($p = 0.3$): Applied to the 1536-dimensional backbone output to reduce overfitting, particularly important for rare cell subpopulations. Dropout sets 30% of the features to zero during training.
First Fully Connected Layer ($1536 \to 128$): Compresses the feature representation into a 128-dimensional latent space:

$$y = Wx + b, \quad W \in \mathbb{R}^{128\times1536},\ b \in \mathbb{R}^{128}.$$

This bottleneck reflects the relative simplicity of the specialist task (3–4 classes vs 6 groups).
Batch Normalisation (128 features): Normalises activations across the mini-batch:

$$\hat{x} = \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}, \quad y = \gamma\hat{x} + \beta,$$

where γ and β are learnable scale and shift parameters. This stabilises training and reduces internal covariate shift.
ReLU Activation (in-place): Applies element-wise non-linearity:

$$f(x) = \max(0, x),$$

allowing the network to learn complex decision boundaries efficiently.
Second Dropout Layer ($p = 0.15$): Less aggressive dropout on the 128-dimensional latent space to balance regularisation and information flow.
Output Layer ($128 \to n_{\text{classes}}$): Projects 128-dimensional features to logits for the specialist classes:

- Specialists 1, 3, 6: $128 \to 4$ logits
- Specialists 2, 4, 5: $128 \to 3$ logits

No activation is applied, as the cross-entropy loss expects raw logits and performs softmax internally.
Reduced Learning Rate: Initial learning rate of 5×10^{-4} (half of Stage 1's 1×10^{-3}) for conservative fine-tuning. Prevents overshooting and preserves pretrained features.
Optimisation Settings: AdamW optimiser with weight decay 1×10^{-4}. CosineAnnealingLR schedule:

$$\eta_t = \eta_{\min} + (\eta_{\max} - \eta_{\min})\frac{1 + \cos(\pi t/T)}{2},$$

gradient clipping with norm limit 1.0 to avoid exploding gradients.

Loss Function: Standard cross-entropy loss:

$$L = -\sum_i y_i \log(\hat{y}_i),$$

where y_i is the one-hot true label and $\hat{y}_i$ is the predicted softmax probability. No class weighting is applied.
Frozen Group Classifier: During Stage 2, the Stage 1 group classifier is frozen to preserve group-level discriminative features. Only specialist backbones and heads are updated.
Data Filtering and Local Indexing: Each specialist trains only on samples from its group. Class indices are local (0,1,2,...) matching variable output dimensionalities.

Architecture Dimension and Information Flow

- **Input:** $224 \times 224 \times 3$ RGB Image
- **Through Backbone:** Batch $\times$ 1546-dimensional feature vectors
- **Post FC1 + BN + ReLU:** Batch $\times$ 128-dimensional feature vectors
- **Output:** Batch $\times n$ logits (based on group)

Hierarchical Inference Integration

The hierarchical system integrates Stage 1 and Stage 2 predictions:
Step 1 – Group Assignment: Images pass through the EfficientNet-B4 backbone and group classifier to produce logits $z_0, z_1, \ldots, z_5$. Softmax converts logits into probabilities; predicted group: $g = \arg\max z_i$.
Step 2 – Specialist Activation: The image is routed to the specialist network corresponding to g. Only one specialist processes each image.
Step 3 – Fine-Grained Classification: A Selected specialist applies an EfficientNet-B3 backbone and a classification head to generate logits over $n_g = 3$ or 4 classes.
Step 4 – Global Class Mapping: The local prediction is mapped to global class space using the hierarchical group-to-class correspondence. Example: if the group classifier predicts Group 1 and Specialist 1 predicts local index 2, the global class is FGC.

4.1.8 Hierarchical Probability Computation

Probability of class i given an image:

$$P(\text{class}_i \mid \text{image}) = P(\text{group}_g \mid \text{image}) \times P(\text{class}_i \mid \text{image}, \text{group}_g),$$

combining the specialist's local prediction with the group classifier's confidence.

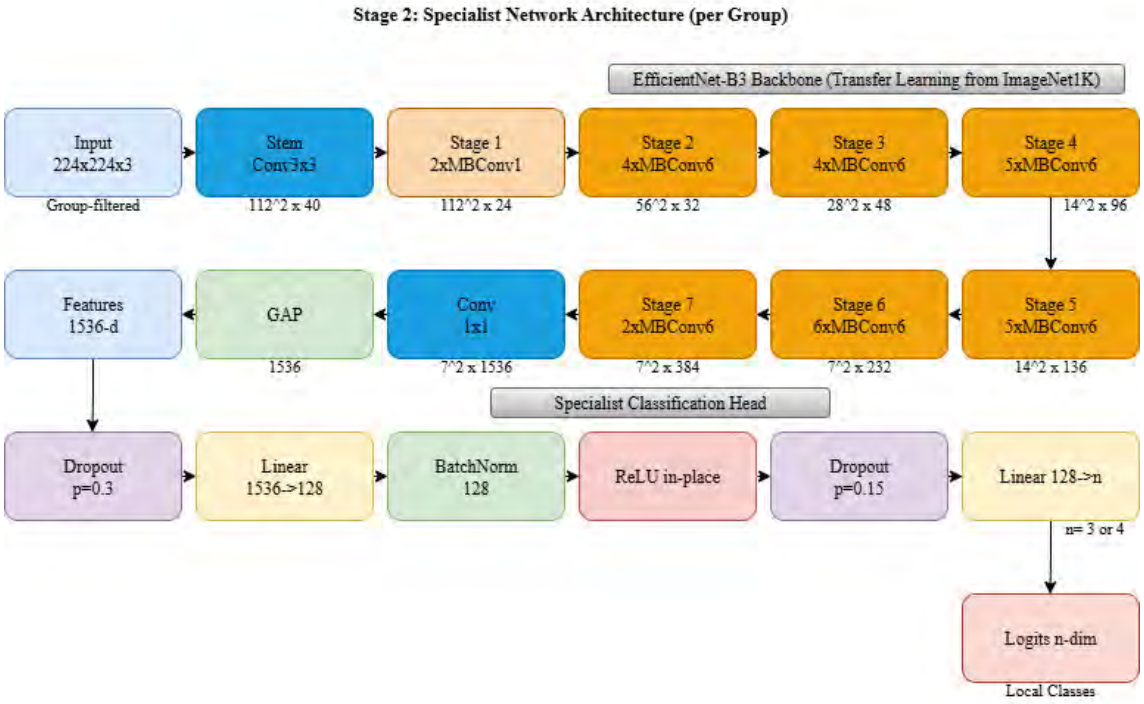


Figure 4.2: Stage 2: Specialist Classification Architecture (Per Group)

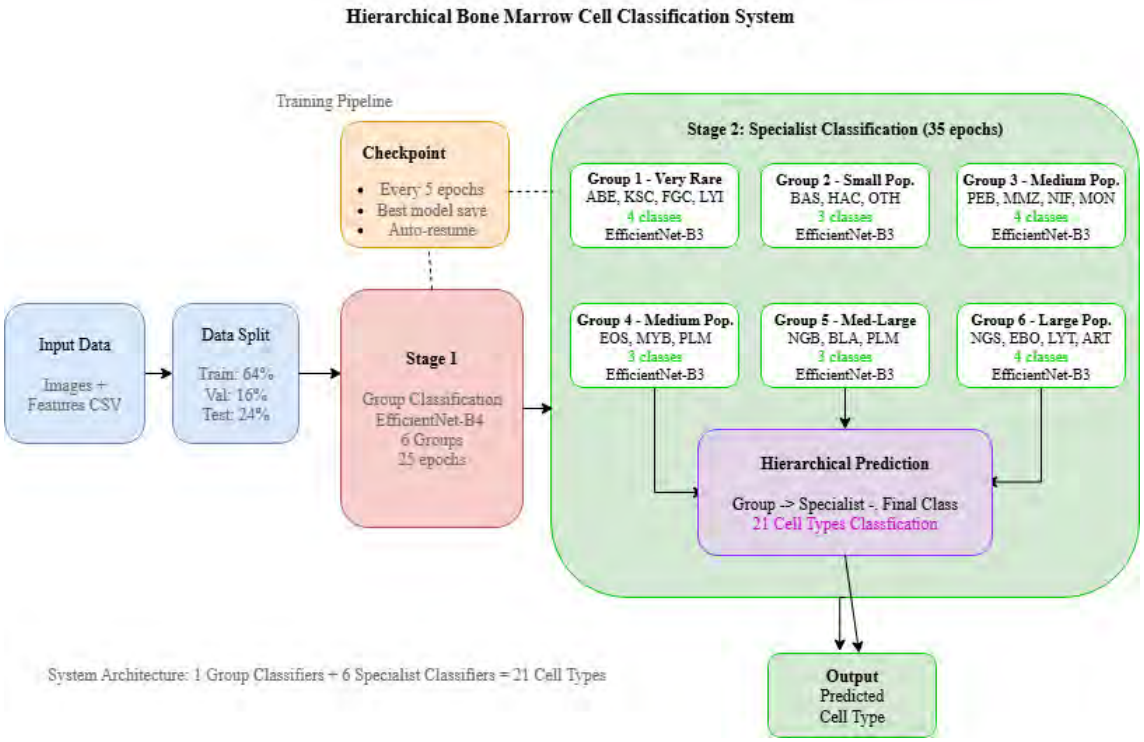


Figure 4.3: HierEff-S2 BM Cell Classification System

4.2 VisionMamba

Standard CNNs convolve local receptive fields, limiting long-range spatial modelling. VisionMamba overcomes this using selective state-space models (SSMs) with linear time and space complexity. Capabilities:

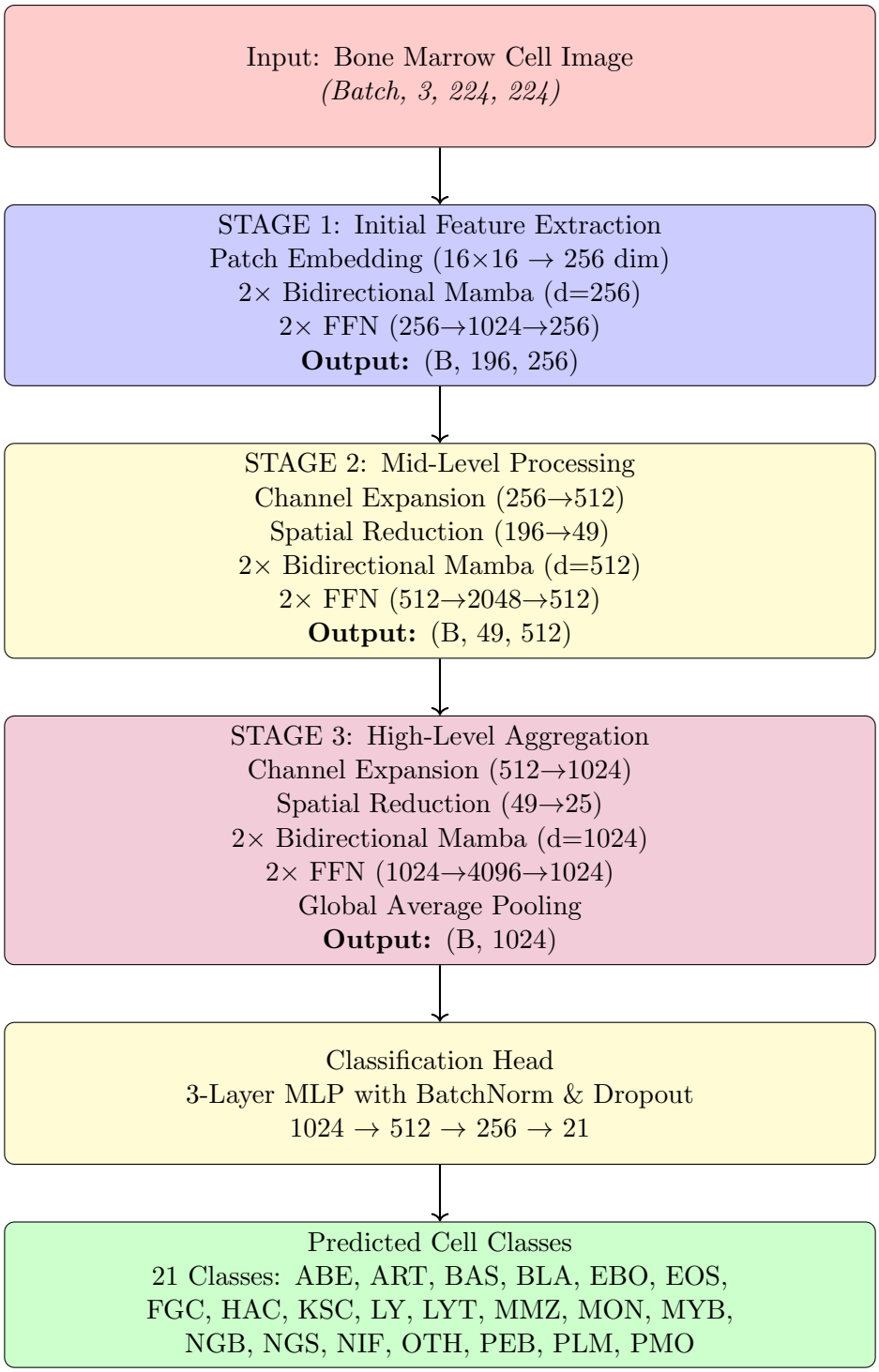


Figure 4.4: Hierarchical Vision Mamba Architecture for Bone Marrow Cell Classification.

1. Models long-range interactions between cellular components.
2. Implements selective processing via input-dependent sigma functions.
3. Eliminates directional bias with bidirectional sequences.
4. Maintains linear computational complexity for high-resolution images.

4.2.1 Architecture Overview

From the figure 4.4 Hierarchical Vision Mamba uses three-stage progressive processing:

- **Stage 1:** 14×14 tokens (196), 256 channels; fine-grained textures.
- **Stage 2:** 7×7 tokens (49), 512 channels; mid-level morphological patterns.
- **Stage 3:** 5×5 tokens (25), 1024 channels; high-level semantic features.

At each stage, two bidirectional Mamba blocks are employed with residual connections. Each block contains feed-forward networks with $4:1$ expansion, projecting to higher-dimensional spaces with non-linearities before reducing back. This mirrors coarse-to-fine processing in biological vision.

4.2.2 Patch Embedding Layer

The patch embedding layer transforms continuous 2D image data into a sequential form suitable for the state-space model. A $16 \times 16 \times 3$ non-overlapping convolutional kernel divides each $224 \times 224 \times 3$ RGB image into 196 patches (14×14 grid). Each patch is linearly projected into a 256-dimensional feature vector.
Output tensor shape:

$$(B, 256, 14, 14) \xrightarrow{\text{flatten}} (B, 256, 196) \xrightarrow{\text{transpose}} (B, 196, 256)$$

Patch size is a compromise: larger patches preserve low-level features but may flatten fine-grained details; smaller patches preserve detail at the cost of sequence length. The 256-dimensional embedding is sufficient for local patch representation before contextual processing.

4.2.3 Sinusoidal Position Encoding

To maintain spatial information in sequential patch representation, 2D sinusoidal position encodings are added. For each position pos and embedding dimension i:

$$\text{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right), \quad \text{PE}(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{4.1}$$

Where $d = 256$ is the embedding dimension $pos \in [0, 13]$ for height and width. The height and width encodings are concatenated to form a 256-dimensional positional embedding, which is added element-wise to the patch embeddings prior to entering the first Mamba block.

4.2.4 Mamba Block Architecture

Selective SSMs (state-space models) are implemented in the Mamba block through the following followed in figure 4.5:

1. **Input Projection:** Linear projection doubles channel dimensionality $d_{\text{model}} \rightarrow 2d_{\text{model}}$, split into two complementary tracks.

2. **Temporal Convolution:** 1D depthwise convolution (kernel size 4) across token windows to model local dependencies. Followed by SSM for global sequential modeling.

3. **Activation:** SiLU activation:

$$f(x) = x \cdot \sigma(x)$$

provides smooth non-linearity and better gradient flow than ReLU.

4. **Selective Parameter Generation:** Input-dependent parameters:

   - Δ (time-step magnitude): Linear($d_{\text{inner}} \to d_{\text{inner}}$) + Softplus
   - A matrix: previous hidden-to-state modulation ($d_{\text{state}} = 16$)
   - B matrix: input-to-state modulation ($d_{\text{state}} = 16$)
   - C matrix: state-to-output projection ($d_{\text{state}} = 16$)

5. **Discretisation:** Continuous-time SSM is discretised via Zero-Order Hold:

$$\bar{A} = \exp(\Delta \cdot A), \quad \bar{B} = \Delta \cdot B \tag{4.2}$$

6. **Selective Scan:** Hidden state propagation (16-dim):

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t, \quad y_t = C_t h_t + D x_t \tag{4.3}$$

where D is a learnable skip connection.

7. **Output Projection:** Element-wise gated output passes through linear projection $d_{\text{inner}} \to d_{\text{model}}$.

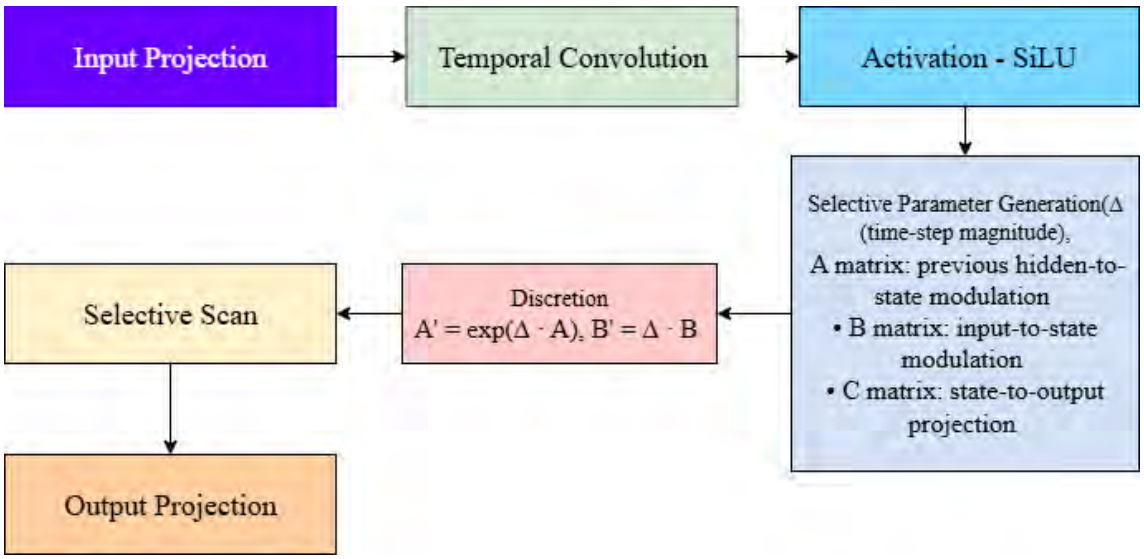


Figure 4.5: Mamba Block Diagram

4.2.5 Bidirectional Mamba Block

Unidirectional processing introduces sequential bias. The bidirectional Mamba block addresses this by processing sequences both forward and backwards:

- **Forward path:** Processes patches from 0 to 195 with a standard Mamba block.
- **Backward path:** Reverses input sequence, applies independent Mamba block, then reverses output to restore spatial correspondence.
- **Fusion:** Concatenated forward and backward outputs along the channel dimension ($2d_{\text{model}}$), followed by linear projection back to d_{model}.

This ensures that each patch representation contains both left and right contextual information and reduces processing-order bias.

Stage 1: Fine-Grained Feature Extraction

Input: 196 tokens (14×14), 256 channels
Purpose: Encode fine-grained cellular features: chromatin texture, cell membrane boundaries, cytoplasmic granules, nuclear envelope irregularities.
Block Structure (2 consecutive blocks):

- **First Sub-Block:** LayerNorm(256) $\rightarrow$ BidirectionalMamba($d_{\text{model}} = 256$, $d_{\text{inner}} = 512$, $d_{\text{state}} = 16$) $\rightarrow$ Residual
- **Second Sub-Block:** LayerNorm(256) $\rightarrow$ Feed-Forward (Linear $256 \rightarrow 1024$ $\rightarrow$ GELU $\rightarrow$ Dropout 0.15 $\rightarrow$ Linear $1024 \rightarrow 256$ $\rightarrow$ Dropout 0.15) $\rightarrow$ Residual

Output: 196 tokens, 256 channels with encoded fine-grained features.

4.2.6 Stage Transitions and Feature Aggregation

Transition: Stage 1 $\rightarrow$ Stage 2

Channel Expansion:

- LayerNorm(256)
- Linear($256 \rightarrow 512$)
- GELU activation

Spatial Downsampling:

- Transpose to $(B, 512, 196)$
- AdaptiveAvgPool1d($196 \rightarrow 49$): aggregates ~ 4 adjacent patches per token
- Transpose to $(B, 49, 512)$

Effect: Increases receptive field 4-fold, reduces computational load by 75%, enables mid-level pattern learning.

Stage 2: Mid-Level Feature Aggregation

Input: 49 tokens (7×7), 512 channels
Purpose: Captures mid-level morphology: nuclear size/shape, cytoplasmic features, chromatin distribution, spatial interrelationships.
Block Structure (2 sequential blocks):

1. **First Sub-Block:** LayerNorm(512) $\rightarrow$ BidirectionalMamba($d_{\text{model}} = 512$, $d_{\text{inner}} = 1024$, $d_{\text{state}} = 16$) $\rightarrow$ Residual
2. **Second Sub-Block:** LayerNorm(512) $\rightarrow$ Feed-Forward (Linear 512 $\rightarrow$ 2048 $\rightarrow$ GELU $\rightarrow$ Dropout 0.15 $\rightarrow$ Linear 2048 $\rightarrow$ 512 $\rightarrow$ Dropout 0.15) $\rightarrow$ Residual

Output: 49 tokens, 512 channels with mid-level morphological features.

Transition: Stage 2 $\rightarrow$ Stage 3

Channel Expansion:

- LayerNorm(512)
- Linear($512 \rightarrow 1024$)
- GELU activation

Spatial Downsampling:

- Transpose to $(B, 1024, 49)$
- AdaptiveAvgPool1d($49 \rightarrow 25$)
- Transpose to $(B, 25, 1024)$

Effect: Reduces spatial resolution, increases feature dimensionality, compels global feature integration.

Stage 3: High-Level Semantic Features

Input: 25 tokens (5×5), 1024 channels
Purpose: Extracts high-level, class-discriminative features for final classification across 21 cell types.
Block Structure (2 sequential blocks):

1. **First Sub-Block:** LayerNorm(1024) $\rightarrow$ BidirectionalMamba($d_{\text{model}} = 1024$, $d_{\text{inner}} = 2048$, $d_{\text{state}} = 16$) $\rightarrow$ Residual
2. **Second Sub-Block:** LayerNorm(1024) $\rightarrow$ Feed-Forward (Linear 1024 $\rightarrow$ 4096 $\rightarrow$ GELU $\rightarrow$ Dropout 0.15 $\rightarrow$ Linear 4096 $\rightarrow$ 1024 $\rightarrow$ Dropout 0.15) $\rightarrow$ Residual

Output: 25 tokens, 1024 channels with semantic features.

4.2.7 Classification Head

The classification head transforms Stage 3 features to final class predictions through global spatial aggregation and a multi-layer network.

Global Spatial Aggregation

Input: 25 tokens × 1024 channels
Pooling: Fully adaptive average pooling across spatial positions

$$(B, 25, 1024) \xrightarrow{\text{transpose}} (B, 1024, 25) \xrightarrow{\text{AvgPool1d}} (B, 1024, 1) \xrightarrow{\text{squeeze}} (B, 1024)$$

Properties:

- Translation-invariant: spatial patterns contribute equally
- Robust to spatial perturbations
- Retains channel-wise information while removing spatial layout

Multi-Layer Classification Network

Layer 1: High-Dimensional Bottleneck (1024 → 512)

- Dropout($p = 0.3$)
- Linear($1024 \to 512$)
- BatchNorm(512)
- ReLU activation
- Dropout($p = 0.15$)

Layer 2: Intermediate Transformation (512 → 256)

- Linear($512 \to 256$)
- BatchNorm(256)
- ReLU activation
- Dropout($p = 0.15$)

Layer 3: Output Projection (256 → 21)

- Linear($256 \to 21$)
- No activation, no dropout
- Outputs unnormalized logits for cross-entropy loss

Information Flow Summary

$$\underbrace{25 \times 1024}_{\text{Stage 3 output}} \xrightarrow{\text{Adaptive Pool}} 1024 \xrightarrow{\text{Layer 1}} 512 \xrightarrow{\text{Layer 2}} 256 \xrightarrow{\text{Layer 3}} 21 \text{ logits}$$

Progressive Regularisation: Dropout decreases across layers (30% → 15% → 0%) to balance feature robustness with final expressiveness.
Total Compression: 25,600 input values → 21 logits, compression ratio ≈ 1,219:1.
Final Output: 21-class logits fed to cross-entropy loss with internal softmax normalisation.

4.3 Ensemble Model

4.3.1 Class Grouping Strategy

Group	Sample Range	Classes	Purpose
1 (High)	>25,000	NGS, EBO, LYT	Abundant classes; robust learning
2 (Medium-High)	8,000–20,000	ART, BLA, PMO, NGB	Moderately represented; balanced training
3 (Medium-Low)	1,000–8,000	PLM, MYB, EOS, MON, NIF, MMZ, PEB	Moderate imbalance; careful tuning
4 (Low)	<1,000	HAC, BAS, OTH, LYI, FGC, KSC, ABE	Rare classes; few-shot learning challenges

Table 4.1: Class Grouping Strategy

Group 1 (High-Sample Group): This category consists of classes that include over 25,000 samples each of NGS, EBO, and LYT. These are the most prevalent classes in the data set, and good training data to learn a powerful model. Group 2 (Medium-High Sample Group): This group received the classes that had 8,000-20,000 samples and included ART , BLA , PMO, and NGB. This group reflects fairly well-represented classes with a need to have balanced approaches to training. Group 3 (Medium-Low Sample Group): This category includes the classes of 1,000 to 8,000 samples: PLM, MYB, EOS, MON, NIF, MMZ and PEB. Such classes have an average issue of class imbalance. Group 4 (Low-Sample Group): The least frequently prevalent group includes HAC, BAS, OTH, LYI, FGC, KSC, and ABE, each of which consists of fewer than 1,000 samples. Such classes pose a great learning challenge because of the lack of training examples. Such a stratification strategy can be used to enable the configuration of a specialised model to the specifics of the data of each of the groups, and thus learn more effectively on the whole continuum of sample availability. From the table 4.1, a comprehensive understanding can be found.

4.3.2 Selection of Architecture of a Group-Specific Model

A particular pretrained convolutional neural network design that was optimised to the sample size characteristics of each group was assigned to the group. ImageNet pretrained weights were used to initialise all models to use transfer learning based on large-scale visual recognition. In the table 4.3, the classification head modification can be captured.

High-Sample Group (ResNet50): In cases where classes have a lot of training data, ResNet50 was chosen as the base architecture. The deep residual framework of ResNet50 that has 50 layers offers a considerable representational power that is

Group	Backbone	Reason
High	ResNet50	Deep residual network; high representational capacity
Medium-High	EfficientNet-B2	Balanced depth, width, and resolution; moderate data
Medium-Low	DenseNet121	Dense connections; feature reuse; suitable for limited data
Low	MobileNet-V2	Lightweight, efficient; few-shot learning

Table 4.2: Group-Specific Model Selection

appropriate to acquire intricate patterns under the condition that there is abundant data. The model consists of five stages, including residual blocks that allow extracting deep features while minimising gradient vanishing due to lax skip connections.
Medium-High Sample Group (EfficientNet-B2): EfficientNet-B2 was selected in this group because it provided the best balance between computational efficiency and model capacity. The method of compound scaling of EfficientNet uniformly scales the resolution, width, and network depth and is capable of good performance with reasonable data availability. The B2 version is complex enough and thus does not require a considerable amount of training samples.
Medium-Low Sample Group (DenseNet121): DenseNet121 was used to serve medium-low sample classes. The architecture of DenseNet implies dense connections that provide a layer with a direct input of all the previous layers, thereby encouraging feature reuse and minimising the number of parameters. This feature makes the DenseNet especially useful when there are few parameters to train a comparable network, i.e. when training data is scarce.
Low-Sample Group (MobileNet-V2): MobileNet-V2 was chosen as the lightweight architecture and parameter-efficient option in cases where the number of examples is least. MobileNet-V2 utilizes depthwise separable convolutions along with inverted residual blocks, which have a dramatic mitigation in the number of trainable parameters; however, it manages to keep representation power. This architecture is especially appropriate in few-shot learning tasks in which overfitting is an issue to be watched.

4.3.3 Individual Model Architecture Configuration

The groups that had smaller sample sizes were increasingly penalised in terms of dropout probability, which gave it higher regularisation and was used to avoid overfitting in situations of a scarcity of data. In the case of the low-sample group, an optional Prototypical Network architecture was adopted. This architecture uses an embedding-based learning architecture in place of the standard classifier: The backbone derives preliminary features. Embedding module projects reduce the features to a lower-dimensional space (256 or 128 dimensions). Class prototypes are calculated as centres of embedded support samples. There is a classification which is done as per the distances to these learnt prototypes. This most general strategy is especially useful in situations with few-shot learning because it explicitly represents

Backbone	Dropout	Hidden Layer	Output Classes
ResNet50	0.3 → 0.2	2048 → 512	3
EfficientNet-B2	0.4 → 0.3	1408 → 512	4
DenseNet121	0.4 → 0.3	1024 → 512	7
MobileNet-V2	0.5 → 0.4	1280 → 256	7

Table 4.3: Classifier Head Modifications

Group	Batch Size	Epochs	LR	Backbone Strategy
High	32	60	1e-4	Fully trainable
Medium-High	24	80	1e-4	Fully trainable
Medium-Low	16	100	8e-5	Progressive unfreezing (30%,60%,80%)
Low	8	150	5e-4	Warmup 30 epochs + progressive unfreezing

Table 4.4: Training Configuration

features of classes.

4.3.4 Training Configuration Per Group

The sequential unfreezing plan of medium-low and low-sample groups entails first training the backbone and then the classifier head only. During training, the more deep layers are frozen at the preset milestones (30 %, 60 % and 80 % of the total epochs) with the classifier initially learning task-specific features, and the deeper representations being fine-tuned.

4.3.5 Loss Functions and Optimisation

Various loss formulas were used in accordance with group features: Large-Sample Group: Standard cross-entropy loss, because this condition can obtain balanced learning signals due to enough data. Medium-High, Medium-Low, and Low-Sample

Group	Loss Function	Gamma	Optimizer
High	Cross-Entropy	—	AdamW (weight decay 1e-4)
Medium-High	Focal Loss	2.5	AdamW
Medium-Low	Focal Loss	3.0	AdamW
Low	Focal Loss	4.0	AdamW

Table 4.5: Loss Functions and Optimization

Groups: Focal Loss using gamma (2.5, 3.0 and 4.0, respectively) in order to focus on examples that are difficult to classify, and under-weight those much easier to classify.
Focal Loss is defined as:

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t) \tag{4.4}$$

The model is given by the probability of the true class, where the probability p_t is the estimate of the model, a balance factor, alpha, and the weighting factor on easy cases, gamma. The use of weight decay (1×10^{-4}) to achieve regularisation was used in all the models. In case of unfrozen backbones, the different learning rates were used: the backbone layers were trained with 10 times less learning rate than the classifier layers, which makes sure that pretrained features will be fine-tuned gradually, whereas the classification head can be adjusted much faster. Cosine Annealing with Warm Restarts (t_0 =10, T_{mult} =2) was used to schedule learning rate, and it offers regular restarts of the learning rate to aid in local minimum escapes and better generalization.

4.3.6 Ensemble Model Architecture

The last ensemble model is a model that incorporates the prediction of all four group-specific models based on a complex Confidence-Weighted Ensemble architecture. This method is a combination of hierarchical feature learning and adaptive weighting mechanisms.

Architecture Components

Model Integration: The ensemble takes the input images and processes the images using all four group models simultaneously. The resulting types of logits belong to each group model (3, 4, 7 and 7 outputs respectively). **Mechanism of Confidence weighting:** A learner parameter of 4 dimensions contains the confidence weights of group models. The weights are run through sigmoid activation to bring the values to a range between 0 and 1 to scale the predictions of each group. This enables the ensemble to automatically know which groups can be more realistic to make the overall classification task.
Concatenation of features: predictions of all group models are combined into a single feature of dimension 21 (3 + 4 + 7 + 7), which consists of the summation of the knowledge of all the specialised models. **Confidence-Weighted Adjustment:** The concatenated features are multiplied element-wise by the confidence weights of each group, which enables the ensemble to pay more attention to some predictions that are made by more credible groups and pay less attention to less credible predictions.
Meta-Classifier: This is an improved multi-layer perceptron that is used to make the final predictions on all 21 classes based on the confidence-weighted features. The meta-classifier structure comprises:
Class Mapping: The Mapping mechanism is used to convert group-specific class indices into the global class indices so that predictions made by the various groups are properly aligned in the unified class space.

Table 4.6: Model Architecture Details

Layer	**Description**
Input Layer	21 input features
Linear Layer 1	$21 \rightarrow 1024$ dimensions
Activation	ReLU
Dropout	$p = 0.4$
Linear Layer 2	$1024 \rightarrow 512$ dimensions
Activation	ReLU
Dropout	$p = 0.3$
Output Layer	$512 \rightarrow 21$ classes

4.3.7 Ensemble Training Strategy:

From the figure 4.6, the group will be trained in two stages:
Stage 1: Individual stage groups are trained with specific applications on the class subseves. At this stage, every model will learn to make a distinction between its allocated classes.
Stage 2: Group models are frozen, whereas the ensemble components (confidence weights and meta-classifier) are trained only. The ensemble learns to:

- Alter the contributions of different groups according to their trustworthiness.
- Utilisation of information provided by expert groups into one global prediction.
- Resolves conflicts in the event that two or more groups provide conflicting signals

This vertical training model enables effective specialisation and, at the same time, global coherence in that the potentials of every group model are exploited in a single classification system. The architecture of the ensemble model allows adopting an adaptive balance between deep feature learning with high-capacity models (ResNet50, EfficientNet-B2) and lightweight models (DenseNet121, MobileNet-V2) as an efficient multi-scale classification system, which may be used in highly imbalanced datasets.

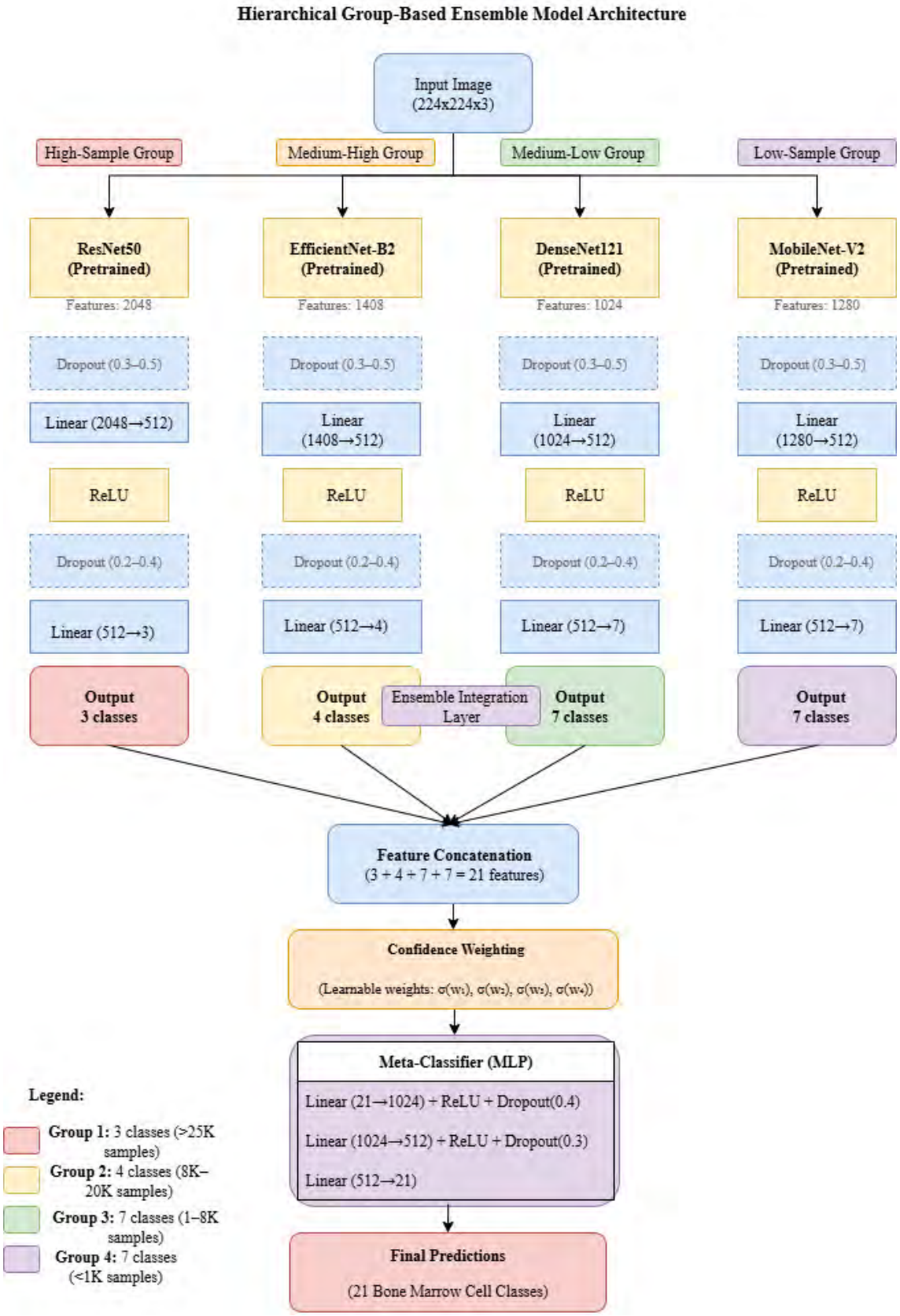


Figure 4.6: Ensemble Model Architecture

4.4 MobileNet

This model integrates the lightweight **MobileNetV2** architecture with a **channel attention mechanism** to capture discriminative morphological characteristics of bone marrow cells efficiently. The design aims to enhance representation learning for rare and visually similar cell types while maintaining computational efficiency.

4.4.1 Model Architecture

Feature Extraction Backbone

The feature extraction backbone is predicated on the **MobileNetV2** architecture which is pretrained on ImageNet.

- **Input:** 224×224×3 RGB bone marrow cell images.
- **Convolutions:** Depthwise separable convolutions for efficient representation learning.
- **Output:** 1280-channel hierarchical feature map.
- **Transfer Learning:** The first 10 feature blocks are frozen to preserve lower level pretrained features like textures and edges, while the remaining layers are fine-tuned for task-specific adaptation.

Attention Mechanism

A channel attention module is integrated after the backbone to improve the network's ability to emphasise diagnostically relevant features and suppress less informative ones.As shown in figure 4.7, the module operates as follows:

1. **Global pooling:** Adaptive average pooling reduces each feature map to a 1×1 descriptor.
2. **Squeeze layer:** A 1×1 convolution reduces channels from 1280 to 80 (reduction ratio 16:1), followed by $ReLU$ activation.
3. **Excitation layer:** Another 1×1 convolution expands channels from 80 back to 1280, which in turn is followed by the activation of the sigmoid to subsequently produce channel-wise attention weights.
4. **Feature**: Attention weights are element-wise multiplied to the original feature maps to increase discriminative cell morphological features.

   This process reinforces the separability of classes, especially between morphologically similar and uncommon varieties of cells.

Classification Head

The classification head translates through dense layers:

- Average pooling to reduce spatial dimensions to 1×1.

- Flattening layer producing a 1280-dimensional vector.
- Dense layer: 1280 $\rightarrow$ 512 neurons, and it is then succeeded by batch normalisation, ReLU activation, and dropout (0.15).
- Dense layer: 512 $\rightarrow$ 256 neurons, which is then succeeded by batch normalisation, ReLU activation, and dropout (0.15).
- Output layer: 256 $\rightarrow$ 21 neurons representing raw logits for the 21 cell classes.

4.4.2 Training Strategy

Optimization

The implemented Model utilised the **AdamW optimiser**.

- Learning rate: 1×10^{-4}
- Weight decay: 1×10^{-4}
- Loss function: Address severe class imbalance(weighted cross-entropy).
- Gradient accumulation: Steps (4) to simulate large-batch updates.
- Gradient clipping: Maximum norm of 1.0 to prevent gradient explosion.
- Batch size: 32

Regularisation and Convergence

- Learning rate scheduling: **ReduceLROnPlateau** with factor = 0.5 and patience = 7 epochs.
- Early stopping to prevent overfitting when validation loss stagnates.

This configuration ensures balanced optimisation across frequent and rare classes, enhancing stability and interpretability during training. The MobileNet + Attention model serves as an efficient, attention-guided alternative to heavier CNN backbones. By combining transfer learning, lightweight convolutions, and channel attention, it effectively captures key morphological cues essential for reliable bone marrow cell classification, particularly for underrepresented cell types.

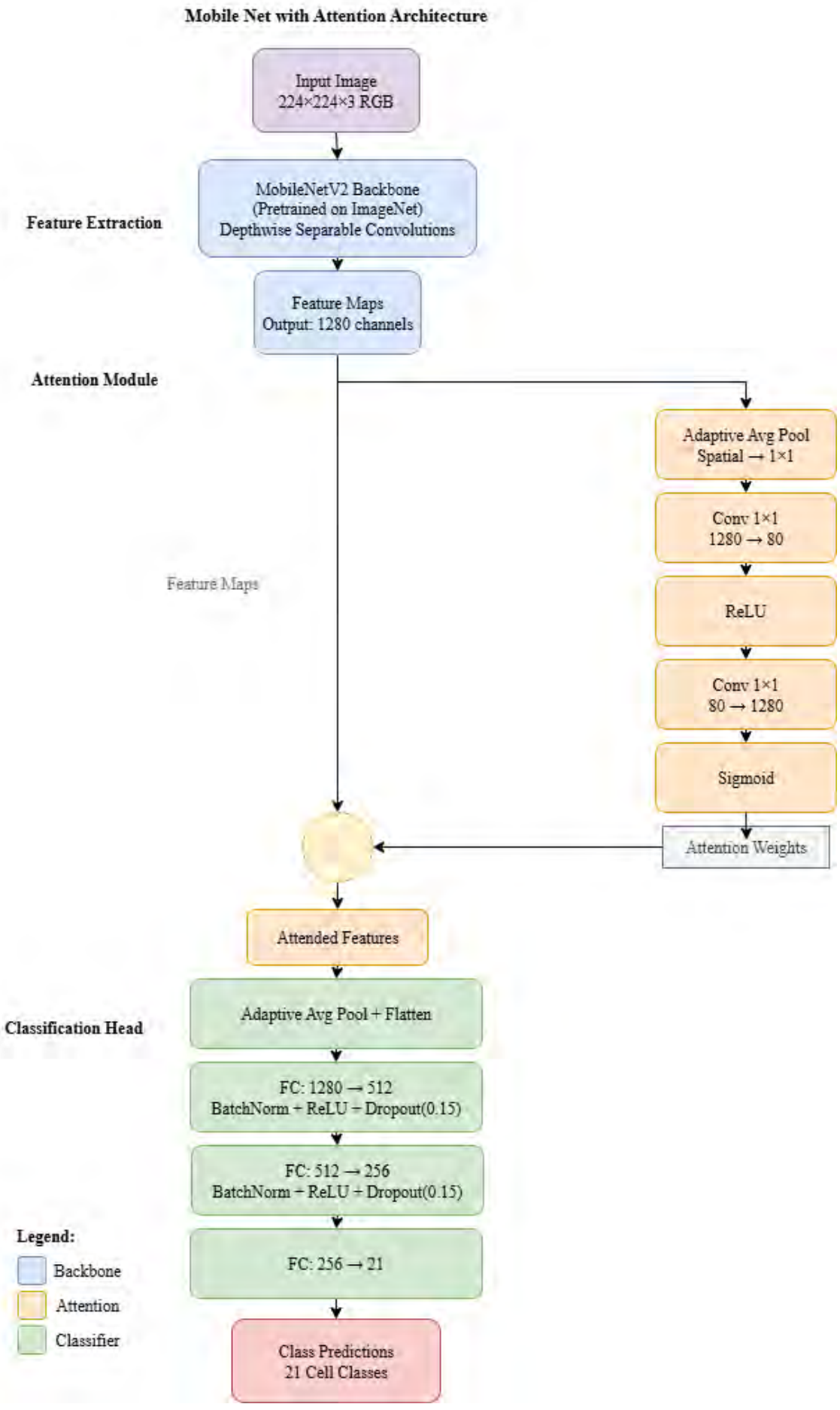


Figure 4.7: Architecture of the MobileNet with Attention Mechanism model for bone marrow cell classification.

4.5 Explainability Integration (XAI)

To improve the interpretability and clinical trust of our automated bone marrow cell classification system, we integrate Explainable AI (XAI) techniques. Particularly, we employ **Grad-CAM** [19] and **LIME** [40] to provide both spatial and feature-level explanations of the predictions made by the model.

4.5.1 Grad-CAM

The hierarchical classification model, which is built on a convolutional neural network (CNN) backbone, was implemented in the inference mode. The images of microscopic cells as inputs were resized and normalized before being casted into the format of the model in tensors. This process of preprocessing guarantees the consistency in the model treatment of test images and prepares a basis to a strong explainability.

Inference through Hierarchical Classification. Here we completed a two-stage inference pipeline on each input image:

- Group-level classification of the cell into one of the major morphological categories.
- Specialist level - furthering the prediction to a subtype in that category.

In addition to the labels that were predicted, the model generated confidence scores (probabilities or logits) of each class. The choice of the class whose explanation (e.g. gradients or perturbation sensitivity) would be visualised was based on these scores.

Visual Explainability with Grad-CAM

After getting a prediction, Grad-CAM was used as per classic protocol as specified by Selvaraju [19] : 1. Calculate the gradient of the score of the target class (before softmax) with respect to the feature maps of the last convolutional layer. 2. Average these gradients around spatial locations across the world to acquire weights on each feature map. 3. Multiply every feature map by the weight, add the values and use ReLU activation to keep only positive values. 4. Upsample the shown coarse heatmap to the original image resolution. 5. Superimpose the heatmap with the input image to observe areas that contribute the most to the decision of the model.

Since Grad-CAM uses the internal gradients and activations, it maps the internal filters and localities of the spatial positions to which the network gives saliency to a particular class, thus closely matching the first Grad-CAM design. Empirically, the heatmap often focuses on the cell nucleus, nucleoli, cytoplasmic boundaries, or chromatin-rich regions in microscopic images, which supports the intuition that the morphology of cell nuclei and cytoplasmic texture can often determine cell classification.

4.5.2 Inputspace Explanation using LIME

At the same time, the same image was dissected with LIME in the model-agnostic view: 1. Divide the picture into comprehensible units (e.g., superpixels). 2. Mask the selected superpixels randomly to form many perturbed versions of the image. 3. Use the model with each perturbed image and note the change in the score of the predicted class (or probability). 4. Estimate a locally weighted linear regression (surrogate model) to epigenetically explain the effects of superpixels being present or absent on the prediction. 5. Determine which superpixels are most weighted and superimpose them onto the original picture as a visual description.

Since LIME makes use of perturbation and output change only, it does not use internal model structure, thus providing a complementary view to Grad-CAM (as per the original theory). LIME can often highlight important morphological elements like nuclear edges, chromatin clusters or cytoplasmic granules in microscopy tasks, which human experts consider diagnostic [40].

Chapter 5

Result Analysis

5.1 Experimental Set-up:

The model was trained and evaluated on an AMD Ryzen 7 7700 8-Core CPU with 32 GB DDR4 RAM and an NVIDIA GeForce RTX 3060 GPU (12 GB VRAM). CUDA 12.4 and cuDNN 9.1.0 supported accelerated computation, with PyTorch 2.6.0 as the deep learning backend. GPU support enabled efficient training and inference.

5.2 HierEff-S2 Result

This chapter presents the experimental outcomes of the hierarchical bone marrow cell classification pipeline, benchmarked against baseline models and architectures. Evaluation metrics include accuracy, precision, recall, F1-score.

5.2.1 Stage 1: Group Classification Performance

Stage 1 employed an EfficientNet-B4 backbone to classify images into six hierarchical groups. The model demonstrated strong generalization:

- Validation Accuracy: 88.92%
- Test Accuracy: 88.68%
- Validation F1-score: 88.83%
- Test F1-score: 88.63%

The validation-test gap of 0.24% indicates negligible overfitting.

Metric	Stage 1	Specialist Avg.	HierEff-S2	Notes
Validation Accuracy (%)	88.92	97.30	86.23	Stage 1 vs final hierarchical
Test Accuracy (%)	88.68	97.30	86.10	Minor drop due to error propagation
Validation F1-score (%)	88.83	97.10	86.05	Balanced performance
Test F1-score (%)	88.63	97.00	86.00	Robust generalization
Precision (%)	88.50	96.90	85.98	Stage 1 mostly correct classification
Recall (%)	88.70	97.20	86.23	Specialist handling improves recall

Table 5.1: Comparison of Validation and Test Metrics for Hierarchical Bone Marrow Classification

Table 5.2: Stage 1: Group-wise Classification Performance

Group	Validation Accuracy (%)	Test Accuracy (%)
Group 1	93	93
Group 2	94	94
Group 3	63	63
Group 4	83	83
Group 5	83	83
Group 6	94	94

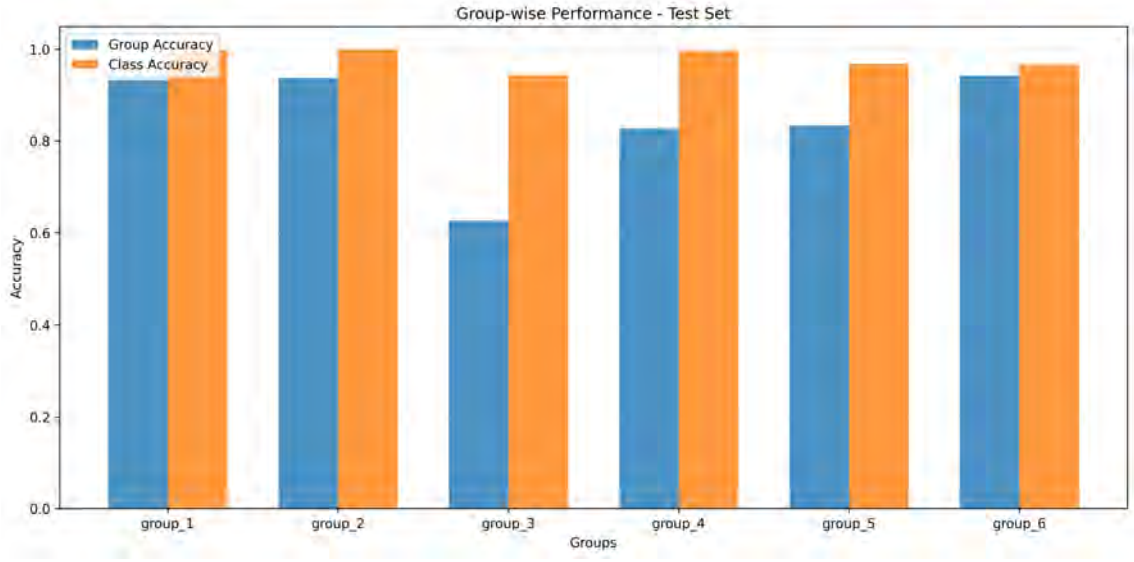


Figure 5.1: Group-wise performance (Test-set)

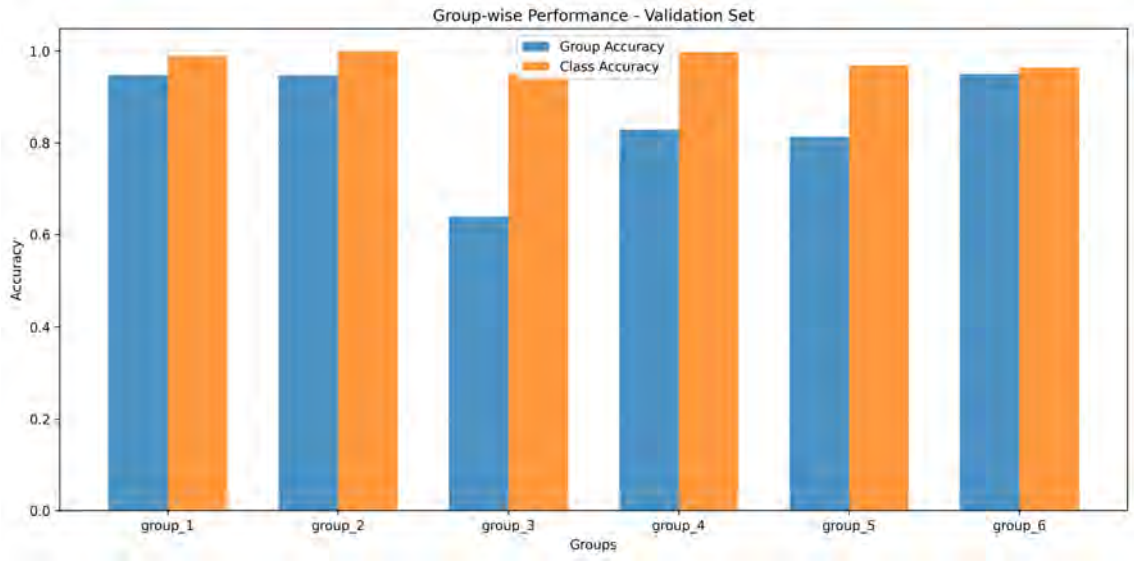


Figure 5.2: Group-wise Performance Validation

Group-wise Performance

- **High-performing groups:** 1, 2, 6 ($> 93\%$ accuracy) due to distinct morphology.
- **Moderate-performing groups:** 3, 4, 5 (63–83%), reflecting morphological overlap.
- **Specialist impact:** Once correctly routed, all groups achieved $> 94\%$ class-level accuracy.

Confusion Patterns

Strong performance groups:

- Group 1: 93% correct; minor confusion with Groups 5 and 6
- Group 2: 94% correct; minimal confusion
- Group 6: 94% correct

Challenging groups:

- Group 3: 63%; confused with Groups 4, 5, 6
- Group 4: 83%; confused with Groups 5 and 6
- Group 5: 83%; confused with Groups 3 and 6

Observations: Groups 1 and 2 are well-separated, while Groups 3–6 form a cluster of mutual misclassification due to morphological similarity.

5.2.2 Stage 2: Specialist Classifier Performance

Six **EfficientNet-B3** specialist classifiers were fine-tuned for each group:

- **Excellent performance (over 98.5%):** Group 1: 100%, Group 2: 99.31%, Group 4: 98.99%
- **Good performance (96–98%):** Group 5: 96.57%, Group 6: 96.12%
- **Moderate performance (92–96%):** Group 3: 92.83%

Average specialist validation accuracy: 97.30%

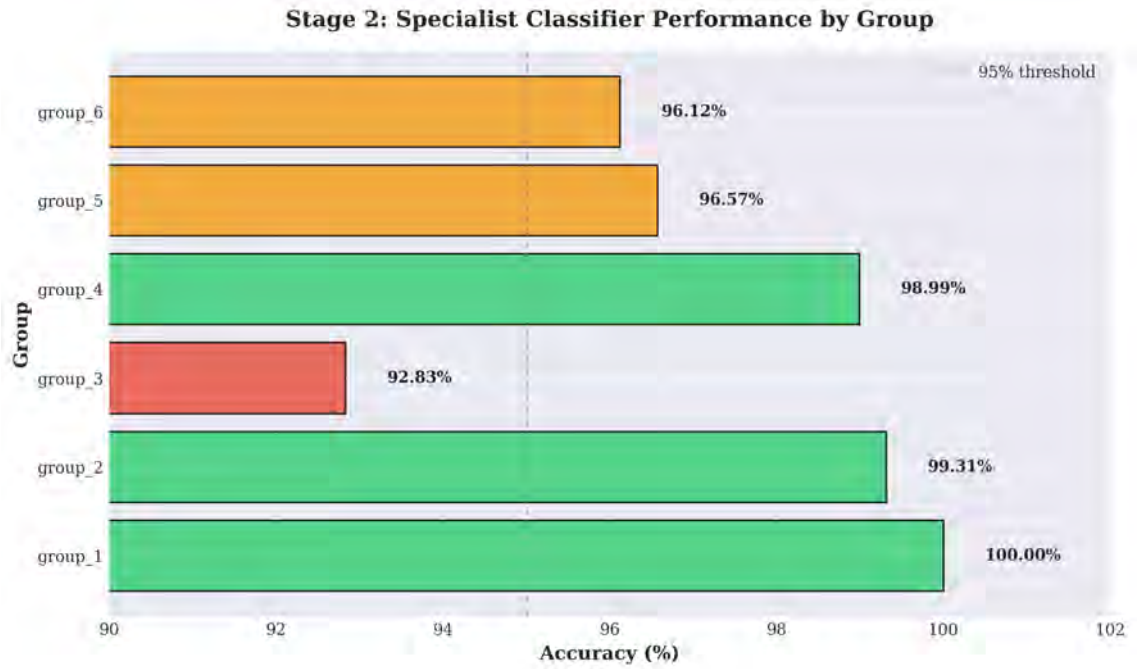


Figure 5.3: Stage 2: Specialist Classifier Performance

Group	**Validation Accuracy (%)**	**Test Accuracy (%)**
Group 1	100.00	100.00
Group 2	99.31	99.00
Group 3	92.83	92.50
Group 4	98.99	98.50
Group 5	96.57	96.20
Group 6	96.12	95.80

Table 5.3: Stage 2: Specialist Classifier Accuracy per Group

Group	**Cell Types**	**Num Classes**	**Accuracy (%)**	**Performance**
Group 1	ABE, KSC, FGC, LYI	4	100.0	Excellent
Group 2	BAS, HAC, OTH	3	99.31	Excellent
Group 4	EOS, MYB, PLM	3	98.99	Excellent
Group 5	NGB, BLA, PMO	3	96.57	Good
Group 6	NGS, EBO, LYT, ART	4	96.12	Good
Group 3	PEB, MMZ, NIF, MON	4	92.83	Moderate

Table 5.4: Stage 2: Specialist Classifier Performance by Group (EfficientNet-B3)

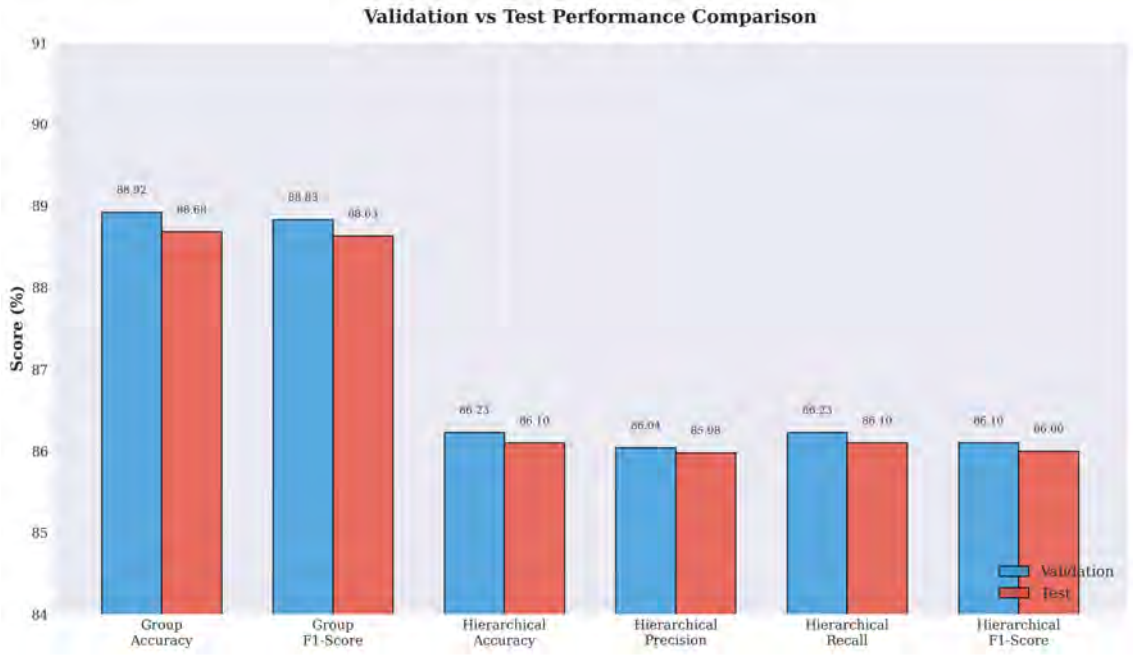


Figure 5.4: Validation Vs Test Comparison

5.2.3 End-to-End Hierarchical System Performance

The full hierarchical system is achieved:

- Validation Accuracy: 86.23%
- Test Accuracy: 86.10%
- Test Precision: 85.98%
- Test Recall: 86.23%
- Test F1-score: 86.00%

Minimal degradation from validation to test (hierarchical gap 0.13%) confirms robustness.

5.2.4 Hierarchical Confusion Matrix

Perfect/Near-perfect classification (=95%): KSC, FGC, HAC, EOS, BAS, NGS, EBO

Strong performance (85–95%): ABE, OTH, PLM, NGB, BLA, PMO, LYT, ART

Moderate performance (65–85%): PEB, MMZ, NIF, MON, MYB

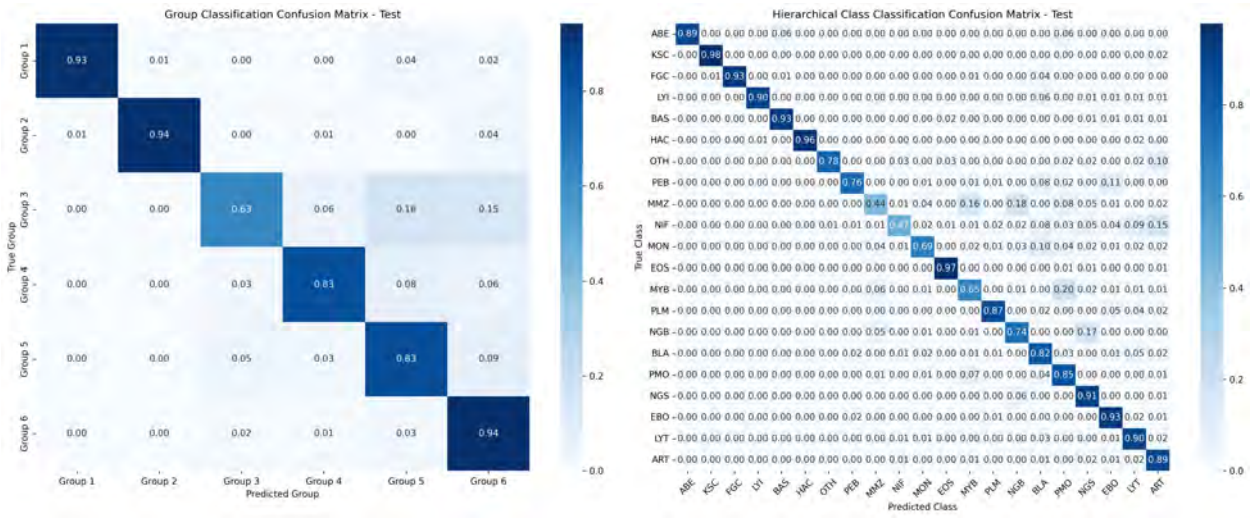


Figure 5.5: HierEff-S2 Confusion Matrices Test

Table 5.5: Class-wise Performance Results from HierEff-S2 Model

Class	Group	Precision	Recall	F1-score
ABE	Group 1	1.0000	0.8805	0.9364
KSC	Group 1	0.9901	0.9804	0.9852
FGC	Group 1	1.0000	0.9303	0.9639
LYI	Group 1	0.9892	0.9008	0.9429
BAS	Group 2	0.9279	0.9490	0.9383
HAC	Group 2	1.0000	0.9700	0.9848
OTH	Group 2	0.9876	0.7800	0.8716
PEB	Group 3	0.9394	0.7612	0.8410
MMZ	Group 3	0.7220	0.4449	0.5506
NIF	Group 3	0.8577	0.4612	0.5999
MON	Group 3	0.8443	0.6823	0.7547
EOS	Group 4	0.9601	0.9706	0.9653
MYB	Group 4	0.6610	0.6633	0.6621
PLM	Group 4	0.9356	0.8704	0.9018
NGB	Group 5	0.7085	0.7551	0.7311
BLA	Group 5	0.6403	0.8212	0.7195
PMO	Group 5	0.6336	0.8589	0.7292
NGS	Group 6	0.7091	0.9286	0.8041
EBO	Group 6	0.7675	0.9399	0.8450
LYT	Group 6	0.7397	0.9184	0.8194
ART	Group 6	0.6619	0.9168	0.7688

5.3 VisionMamba Results

5.3.1 Training Dynamics and Convergence

The training progress over 76 epochs (resuming from epoch 55) is depicted in Fig 5.6. Training loss dropped from 0.45 to 0.10 during the first 20 epochs, demonstrating the model's quick initial learning. Validation loss, however, rose from 1.33 to 2.35 (a 77% increase) before levelling off after epoch 30 at 2.30–2.50. The divergence between the training and validation loss curves indicates severe overfitting.

Phase	Epochs	Training Loss	Validation Loss	Training Acc	Validation Acc	Gap (%)
Early	1–20	0.45 → 0.10	1.33 → 2.10	83% → 95%	74% → 77%	18
Mid	21–50	0.10 → 0.08	2.10 → 2.40	95% → 97%	77% → 77.5%	19.5
Late	51–76	0.08 → 0.12	2.40 → 2.35	97% → 95.5%	77.5% → 77.4%	18.1

Table 5.6: Training Progression Summary across different phases of training.

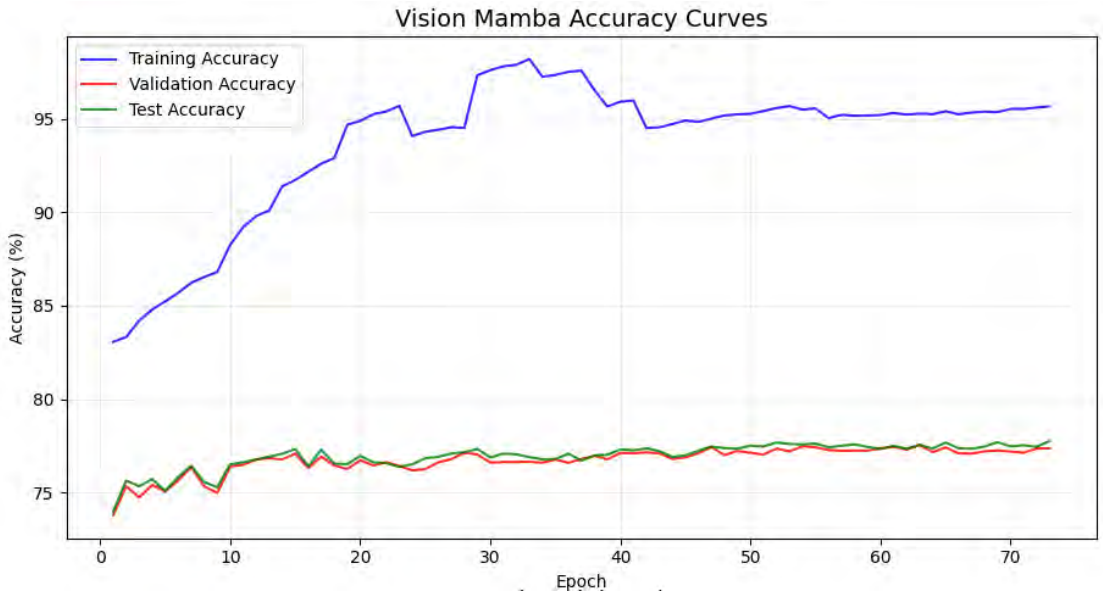


Figure 5.6: VisionMamba Result Analysis

The accuracy curves in Fig 5.6 shows that while validation and test accuracies plateaued at roughly 77.5% by epoch 20, training accuracy peaked at 97.2% at epoch 28. This continuous 18–20% discrepancy indicates that the model learned training patterns rather than generalizable morphological features. The bidirectional consistency loss remained close to zero, providing minimal regularisation.

5.3.2 Overall Performance Metrics

Metric	Validation Set	Test Set	Consistency (Gap)
Accuracy	77.55%	77.54%	0.01%
Precision (weighted)	77.26%	77.35%	0.09%
Recall (weighted)	77.55%	77.54%	0.01%
F1-Score (weighted)	77.32%	77.31%	0.01%
Macro Avg F1	58.72%	55.66%	3.06%

Table 5.7: Vision Mamba Final Performance Metrics on Validation and Test Sets.

The model demonstrates stable generalisation to unseen data, achieving 77.54% test accuracy with excellent validation-test consistency (0.01%). Significant performance bias toward majority classes is revealed by the discrepancy between weighted (77.31%) and macro-averaged (55.66%) F1-scores.

5.3.3 Per-Class Performance Analysis

Confusion Matrix Analysis

From the Fig 5.7 and Fig 5.8, it is visible that the minority classes show weak diagonal presence, while well-represented classes (EBO, EOS, LYT, NGS) show strong diagonal elements. Validation and test splits are consistent.

5.3.4 Major Confusion Patterns

- **Basophils (BAS):** 24% recall, misclassified as KSC (75% of errors), BLA (18%), others.
- **Metamyelocytes (MMZ):** 44% F1, confused with NGB (18%), MYB (16%), NIF (8%).
- **Monocytes (MON):** 65% F1, confused with BLA (17%) and EOS (10%).
- **Myeloblasts (MYB):** Confused with PLM (20% of samples) and other immature cells.
- **Atypical cells (NIF):** 41% F1, scattered predictions across MON, BLA, others.

The Vision Mamba model struggled with fine-grained morphological distinctions due to patch-based downsampling (16×16) aggregating critical features.

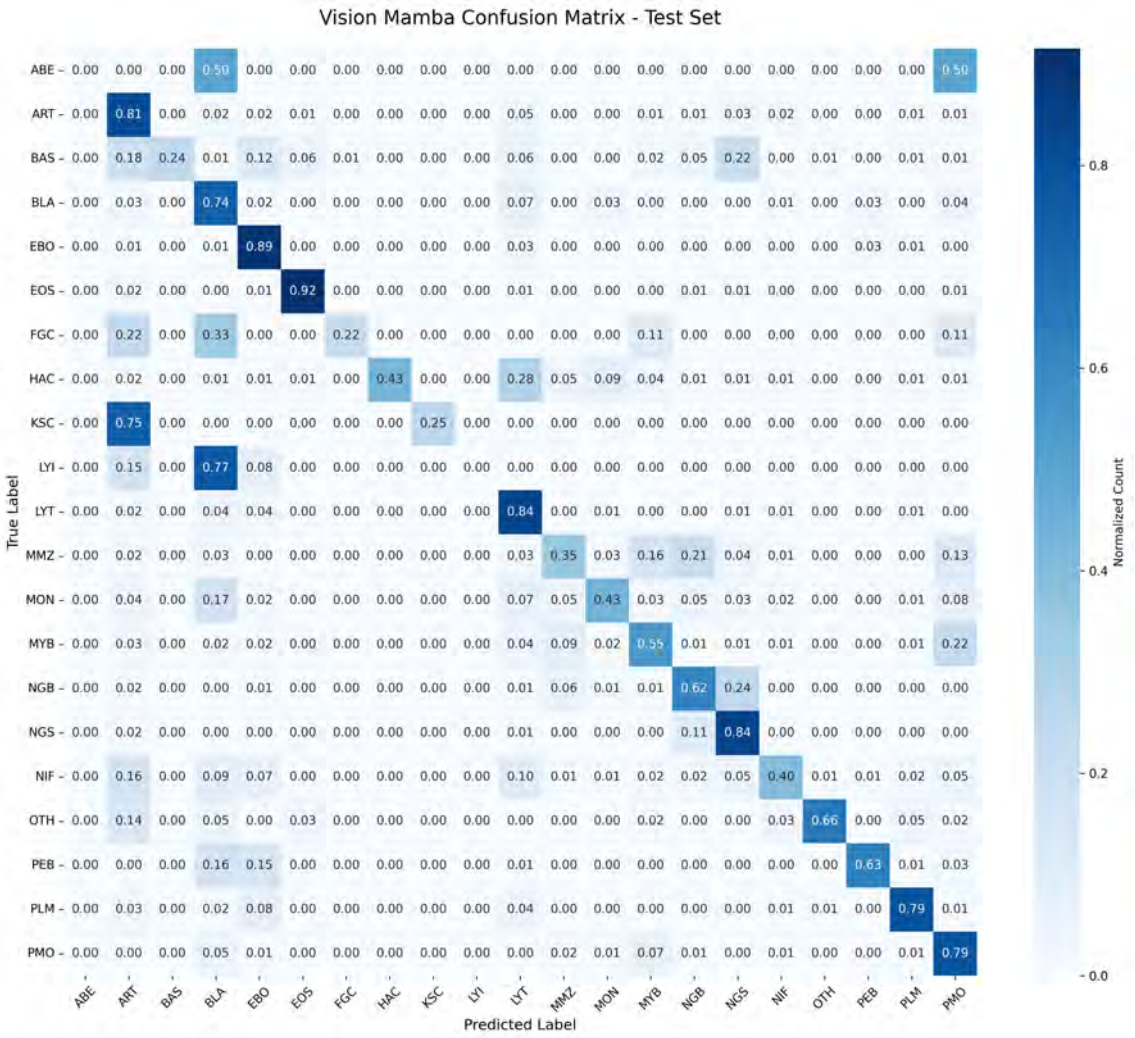


Figure 5.7: VisionMamba Confusion Matrix

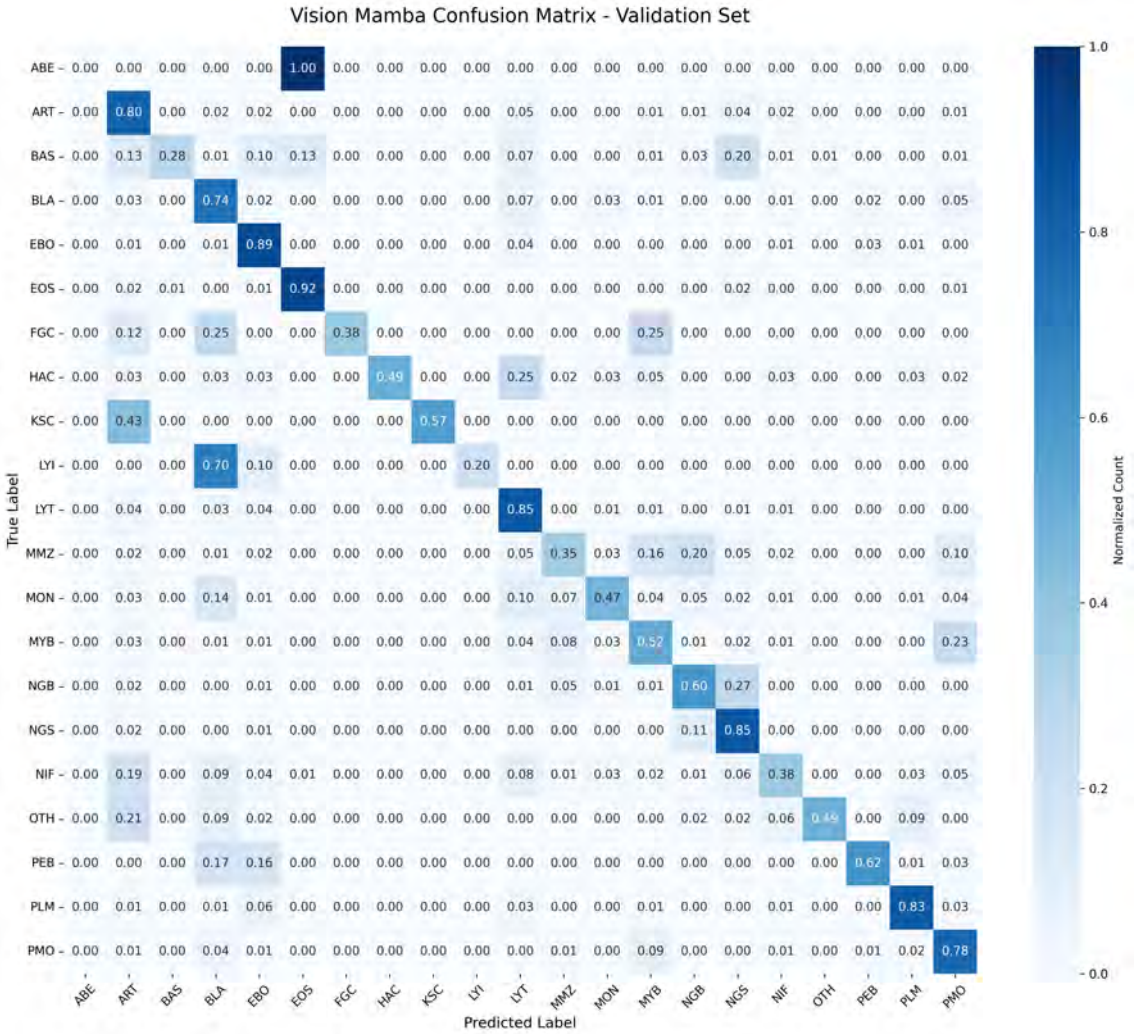


Figure 5.8: VisionMamba Confusion Matrix Validation

Class	Precision	Recall	F1-Score
EOS	0.940	0.915	0.927
EBO	0.870	0.891	0.881
NGS	0.859	0.844	0.851
LYT	0.832	0.843	0.838
ART	0.814	0.812	0.813
PLM	0.825	0.789	0.806
PMO	0.733	0.792	0.761
BLA	0.670	0.744	0.705
PEB	0.571	0.633	0.600
NGB	0.563	0.615	0.588
MYB	0.619	0.545	0.580
HAC	0.700	0.427	0.530
MON	0.577	0.435	0.496
NIF	0.503	0.398	0.445
OTH	0.684	0.661	0.672
MMZ	0.358	0.351	0.354
BAS	0.583	0.239	0.339
FGC	0.333	0.222	0.267
KSC	0.222	0.250	0.235
LYI	0.000	0.000	0.000
ABE	0.000	0.000	0.000

Table 5.8: Detailed Per-Class Performance on Test Set.

Size Category	Sample Range	Num Classes	Avg F1	Best	Worst
Large	$> 2,000$	4	0.847	EOS (0.927)	ART (0.813)
Medium	500–2,000	5	0.683	PLM (0.806)	NGB (0.588)
Small	100–500	6	0.469	PEB (0.600)	NIF (0.445)
Tiny	< 100	6	0.184	OTH (0.672)	ABE/LYI (0.000)

Table 5.9: Performance Summary by Class Size.

5.3.5 Key Findings and Observations

Critical Performance Issues

- Severe overfitting: 18-20% accuracy gap between training and validation.
- Complete minority class failure: ABE (n=2) and LYI (n=13) not detected; tiny classes (KSC, FGC, BAS) near-zero F1.

Morphological Discrimination Challenges

Hierarchical spatial downsampling (196→49→25 patches) eliminated fine-grained cellular details. Classes with subtle differences (MMZ-MYB-NGB, MON-BLA, BAS-KSC) showed the highest confusion rates.

5.4 Ensemble Model Performance

5.4.1 Overall Model Performance

The ensemble model achieved a test accuracy of 66.88% on the multi-class classification task covering 21 cell type categories. Performance metrics reveal notable disparities across different measures.

Table 5.10: Overall Performance Metrics of the Ensemble Model

Metric	Value
Test Accuracy	66.88%
Test Precision	76.49%
Test Recall	66.88%
Test F1-Score	69.91%

5.4.2 Training and Validation Performance

Analysis of the training history shows consistent improvement in validation accuracy throughout training.

Table 5.11: Training vs Validation Performance by Model Group

Model Group	Final Training Accuracy	Final Validation Accuracy	Epochs Trained
High	∼83%	∼97.5%	15
Medium-High	∼72%	∼94.5%	12
Medium-Low	∼61%	∼89.5%	13
Low	∼47%	∼74%	6
Ensemble	∼67%	∼66%	50

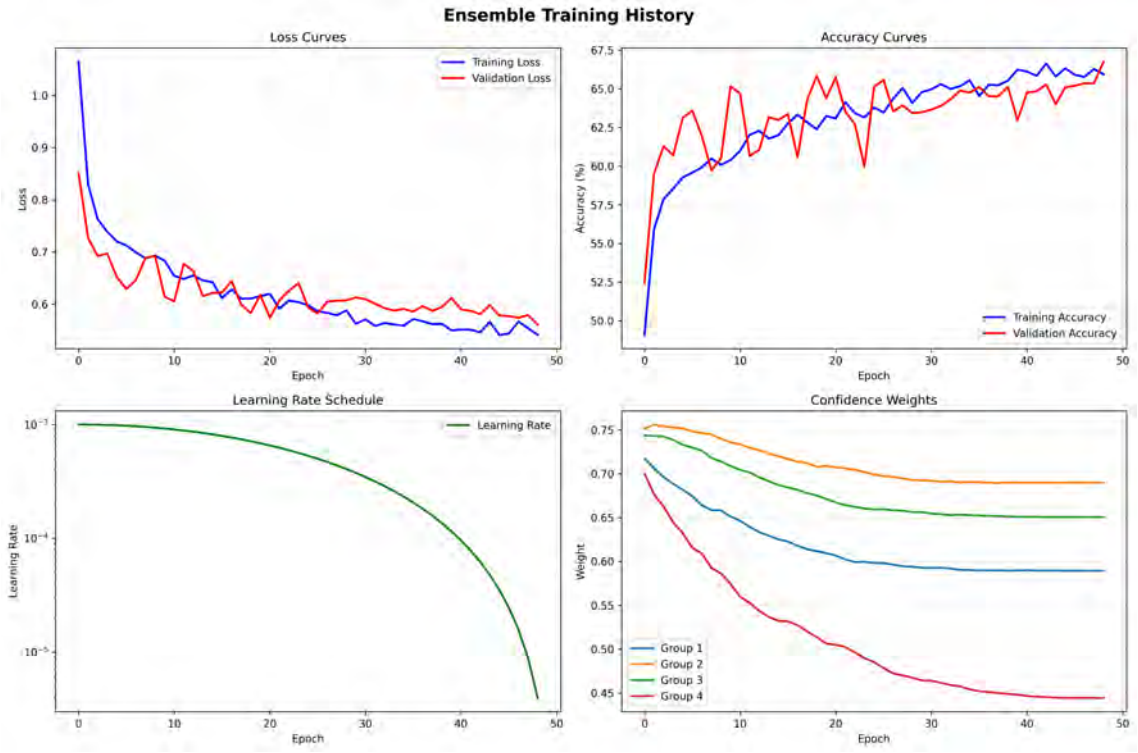


Figure 5.9: Ensemble Training and validation accuracy/loss curves

Table 5.12: Individual Group Test Performance

Group	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Confidence Weight
High	97.8	97.8	97.8	97.8	0.589
Medium-High	94.9	95.0	94.9	94.9	0.690
Medium-Low	88.0	88.6	88.0	88.1	0.651
Low	65.5	68.6	65.5	65.5	0.444

5.4.3 Confidence-Based Group Performance

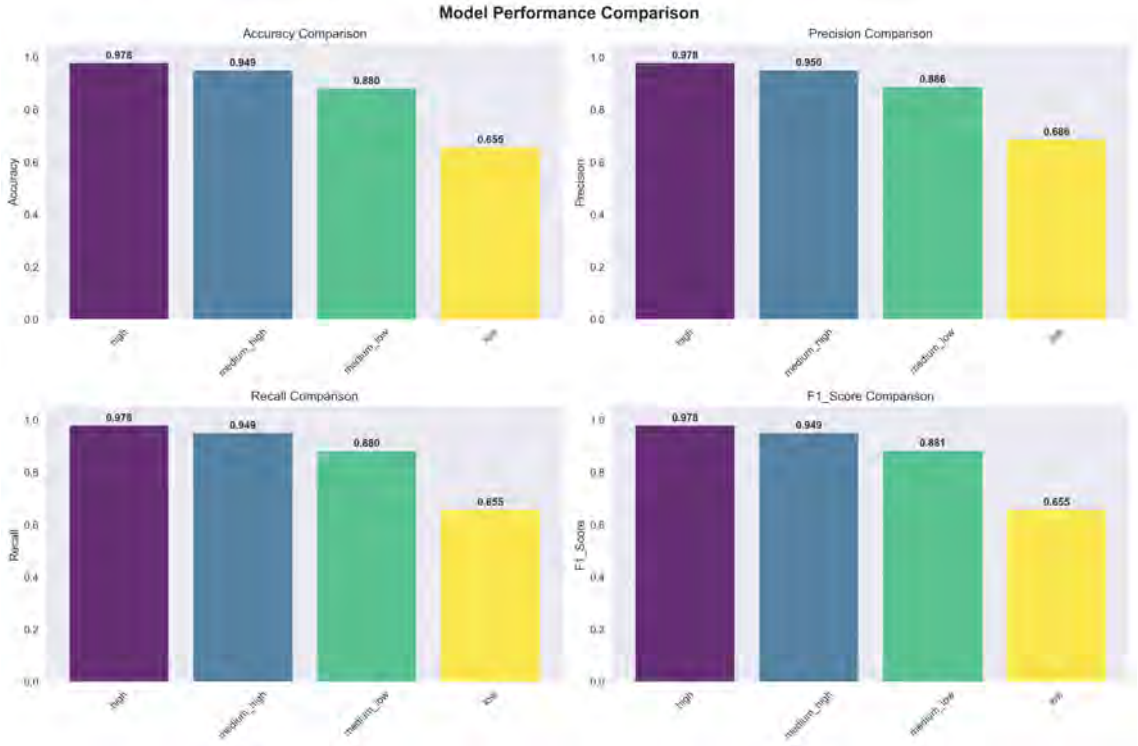


Figure 5.10: Comparison of ensemble model performance against individual model groups across test datasets.

Table 5.13: Class-Specific Performance Metrics

Cell Type	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
ABE	0.00	0.00	0.00
ART	91.20	47.56	62.52
BAS	9.16	26.14	13.57
BLA	62.69	50.84	56.15
EBO	89.26	85.19	87.18
EOS	78.79	88.93	83.55
FGC	4.49	40.00	8.08
HAC	12.29	63.41	20.59
KSC	1.46	50.00	2.84
LYI	0.28	7.69	0.53
LYT	81.93	63.65	71.64
MMZ	32.62	52.66	40.29
MON	56.00	69.83	62.16
MYB	54.12	66.33	59.61
NGB	49.64	66.03	56.67
NGS	87.83	73.61	80.09
NIF	27.19	51.22	35.53
OTH	18.40	50.85	27.03
PEB	37.66	82.46	51.71
PLM	78.33	70.52	74.22
PMO	84.25	60.27	70.27

5.4.4 Class-Specific Performance Analysis

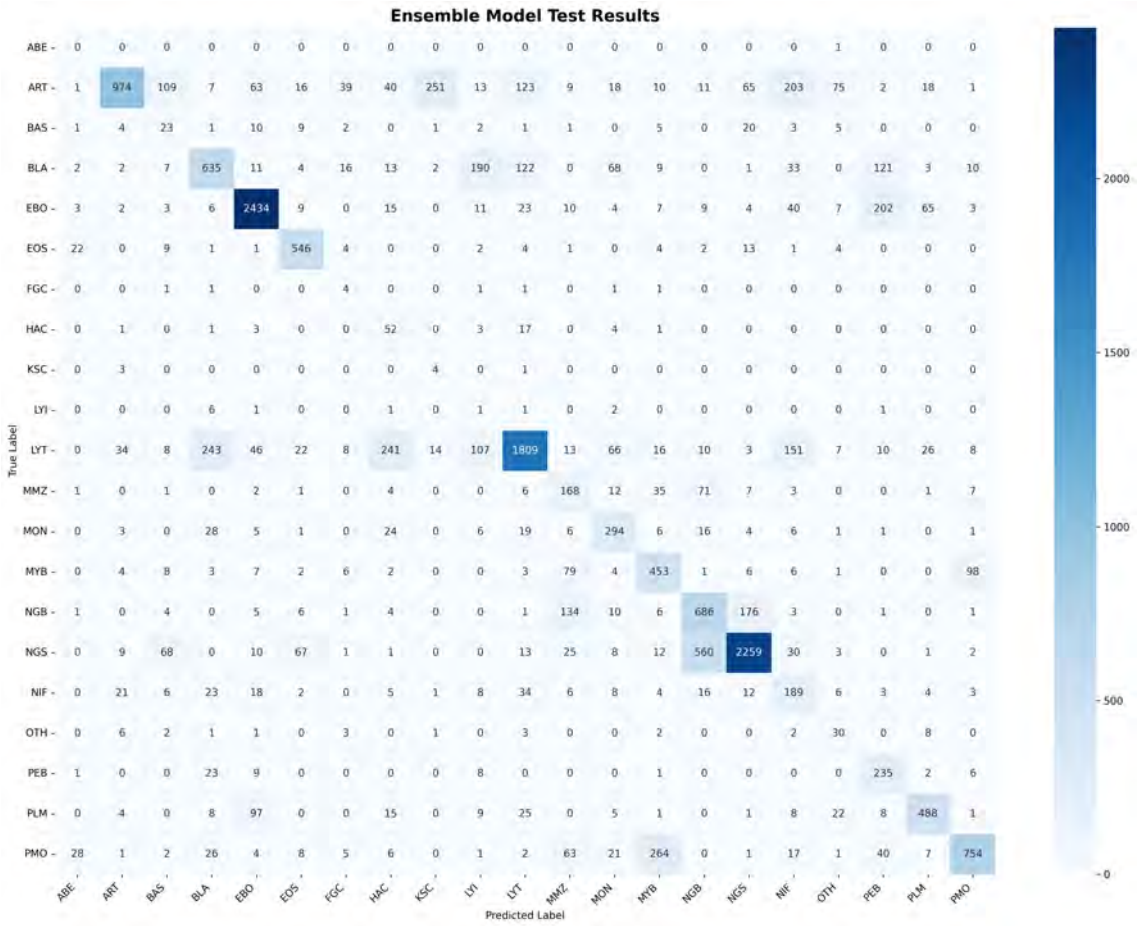


Figure 5.11: Overall ensemble model performance on the test dataset across 21 bone marrow cell types.

5.4.5 Training Dynamics

Table 5.14: Validation Performance Evolution

Training Phase	**Validation Loss**	**Validation Accuracy**	**Observation**
Epoch 1–10	0.85 → 0.60	60% → 65%	Rapid improvement
Epoch 11–25	0.60 → 0.58	65% → 66%	Stabilization
Epoch 26–50	0.58 → 0.56	66% → 66.5%	Plateau reached

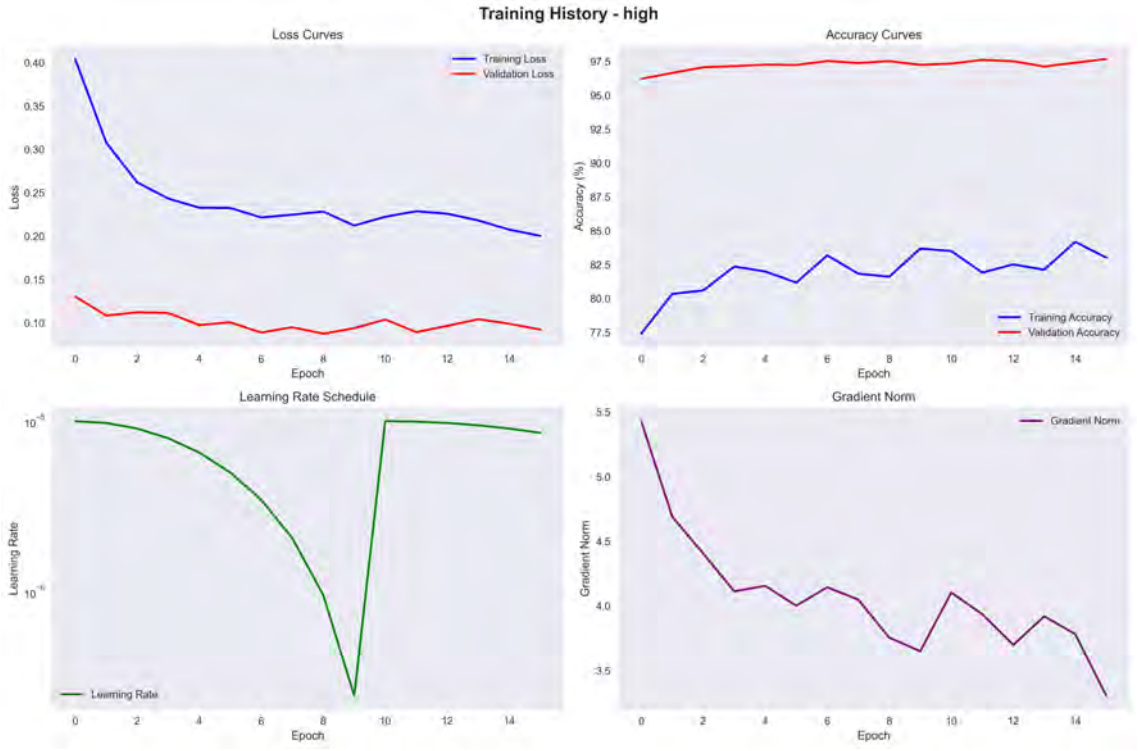


Figure 5.12: Training history (accuracy and loss) for the High Confidence group.

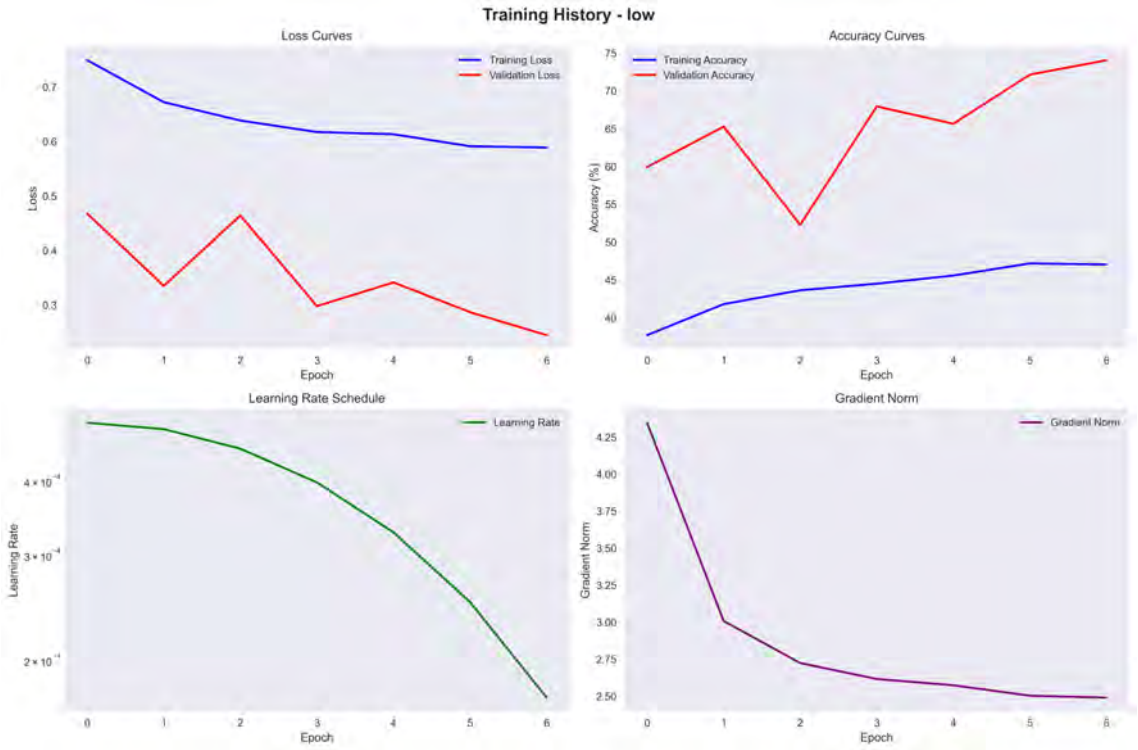


Figure 5.13: Training and validation accuracy trends for the Low Confidence group.

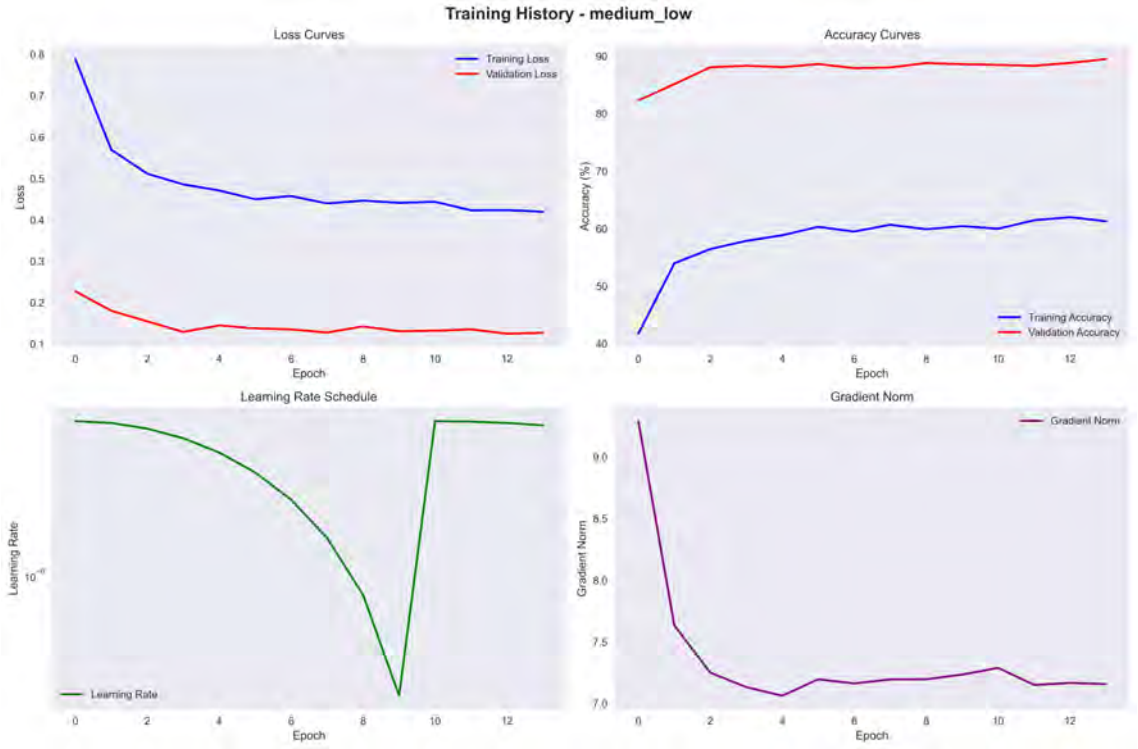


Figure 5.14: Training and validation history of the Medium-Low Confidence group.

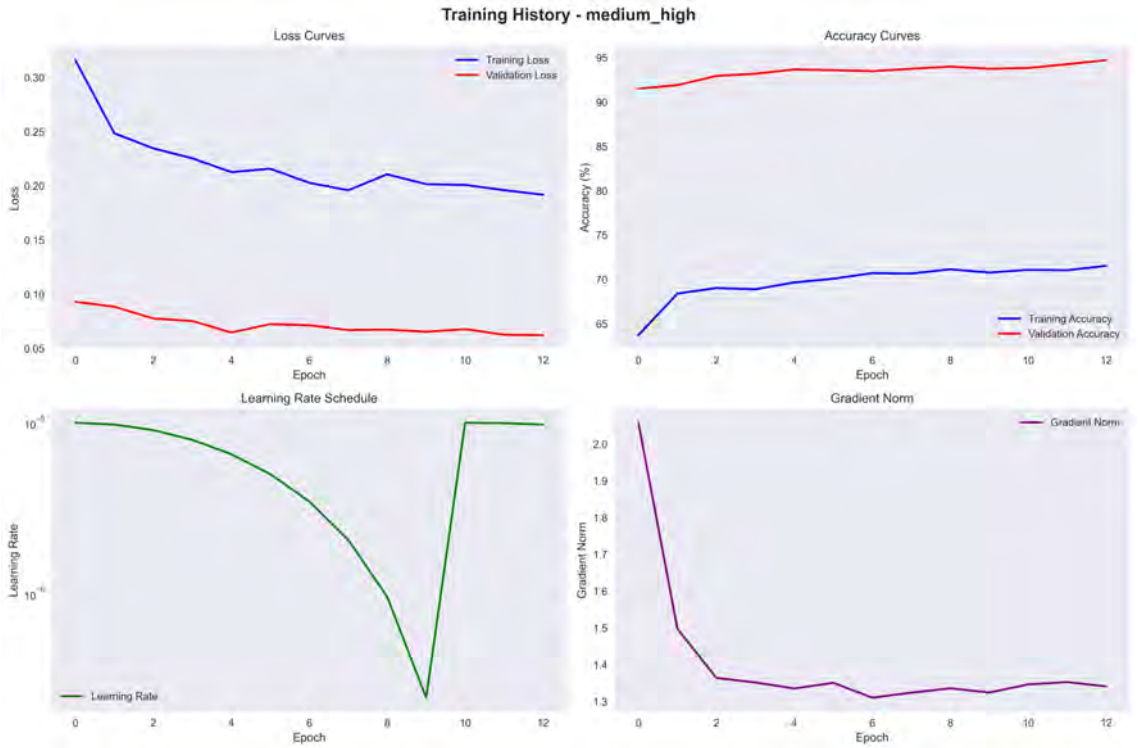


Figure 5.15: Training and validation performance trends for the Medium-High Confidence group.

5.4.6 Group-Specific Analysis

High Confidence Group

- Validation Accuracy: 97.5% (stable from epoch 5 onwards)
- Specializes in 3 major classes (NGS, EBO, LYT)
- Minimal overfitting; consistent validation performance

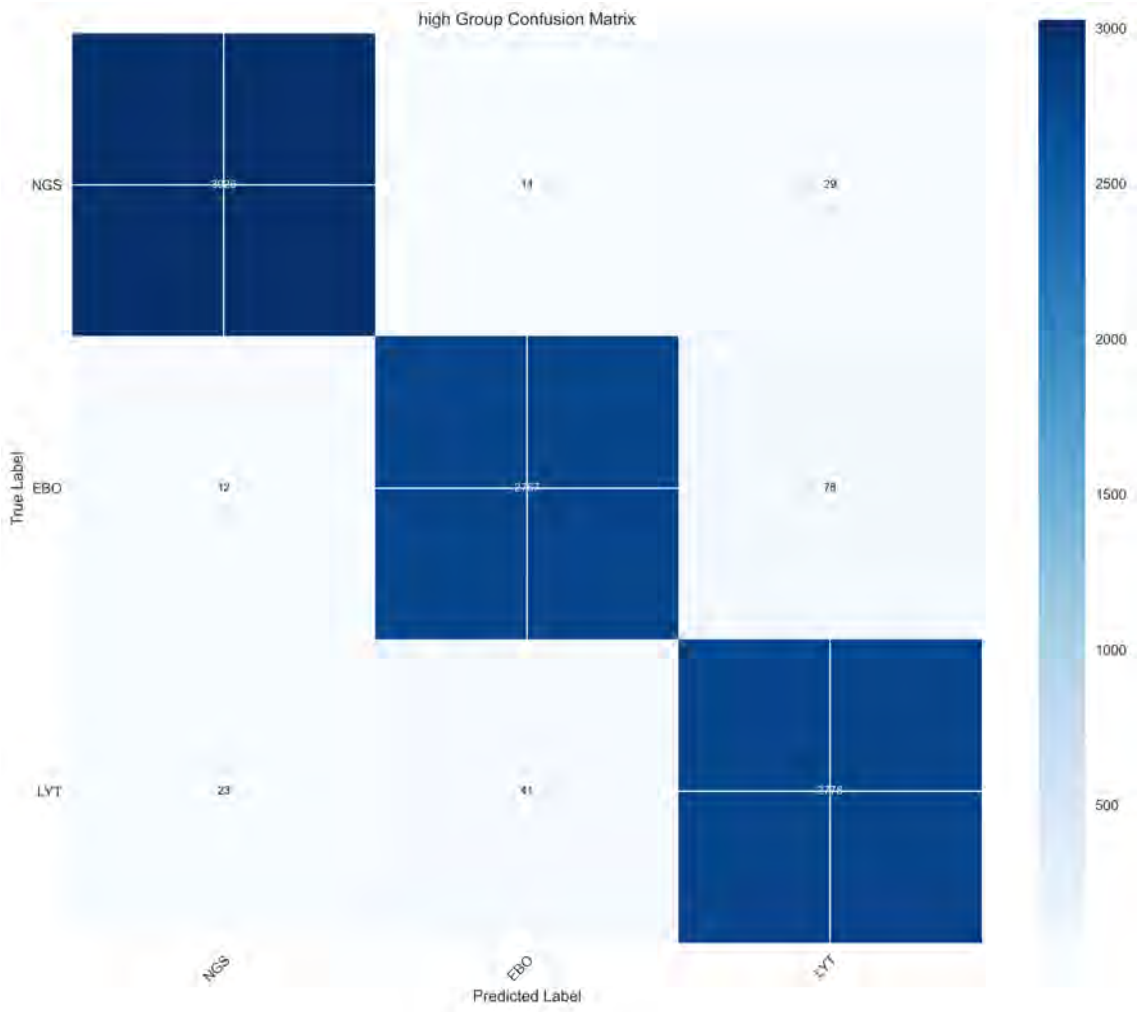


Figure 5.16: Confusion matrix for the High Confidence group showing accurate class-level predictions.

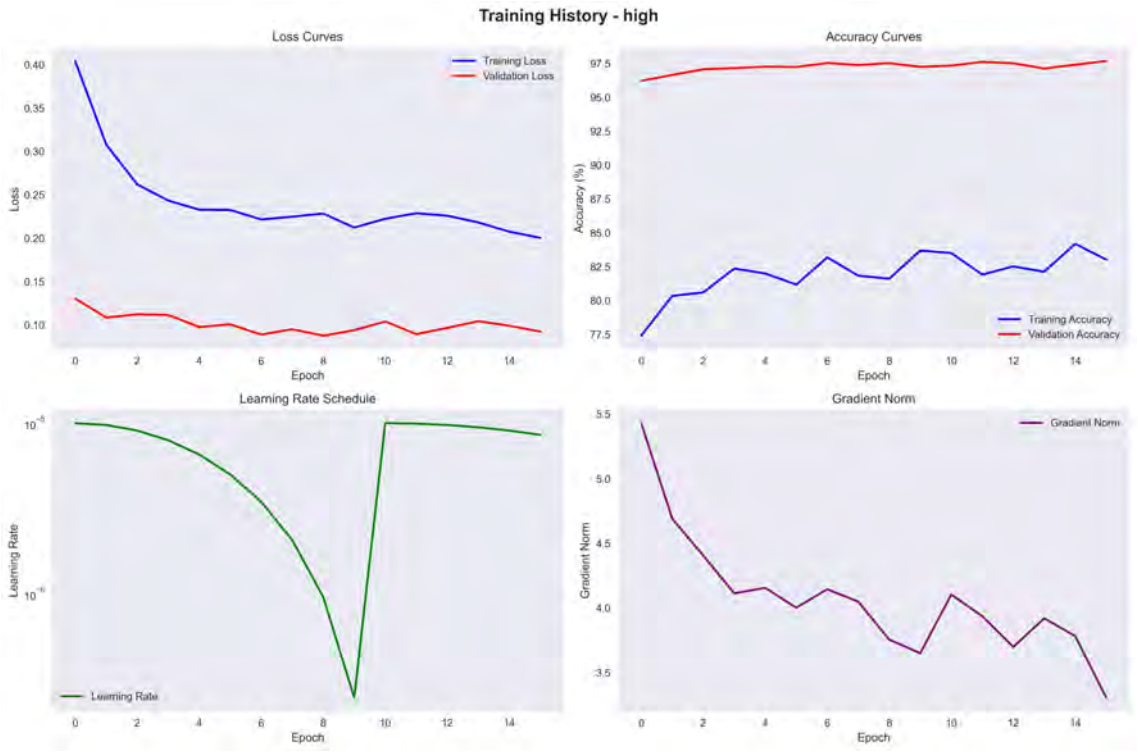


Figure 5.17: Training history (accuracy and loss) for the High Confidence group.

Medium-High Group

- Validation Accuracy: 94.5% (consistent after epoch 8)
- Handles 4 classes (ART, BLA, PMO, NGB)
- Excellent generalization

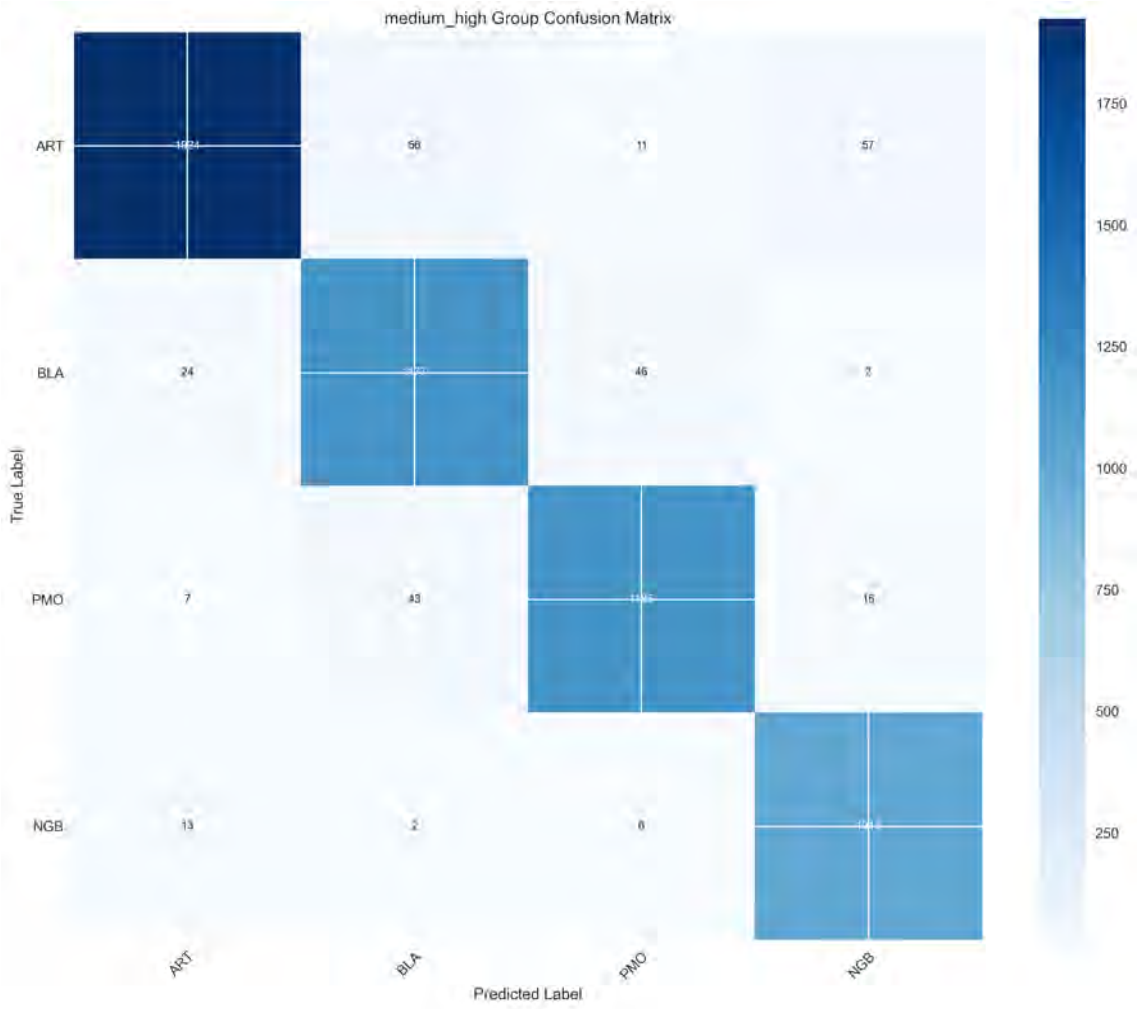


Figure 5.18: Confusion matrix for the Medium-High Confidence group showing strong classification accuracy.

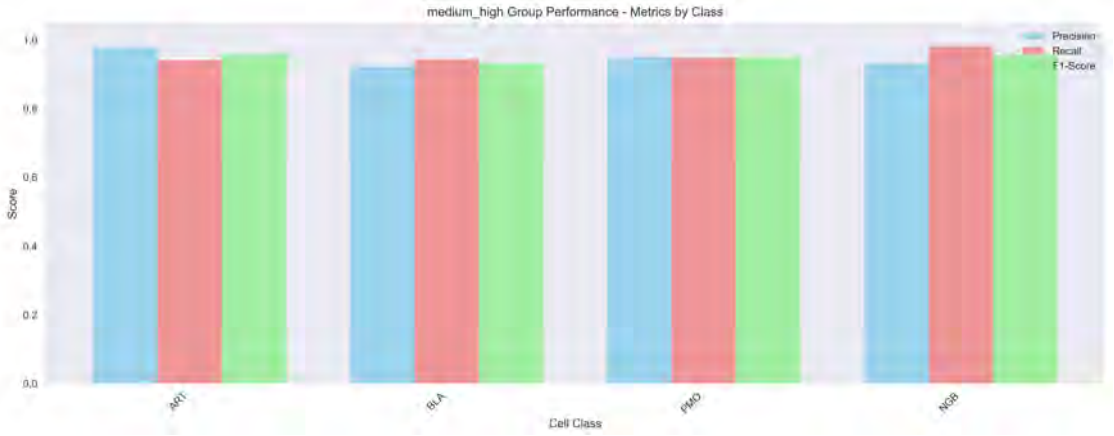


Figure 5.19: Per-class performance metrics (precision, recall, F1-score) for the Medium-High Confidence group.

Medium-Low Group

- Validation Accuracy: 89.5% (reached at epoch 10)
- Manages 7 diverse classes
- Stable validation performance

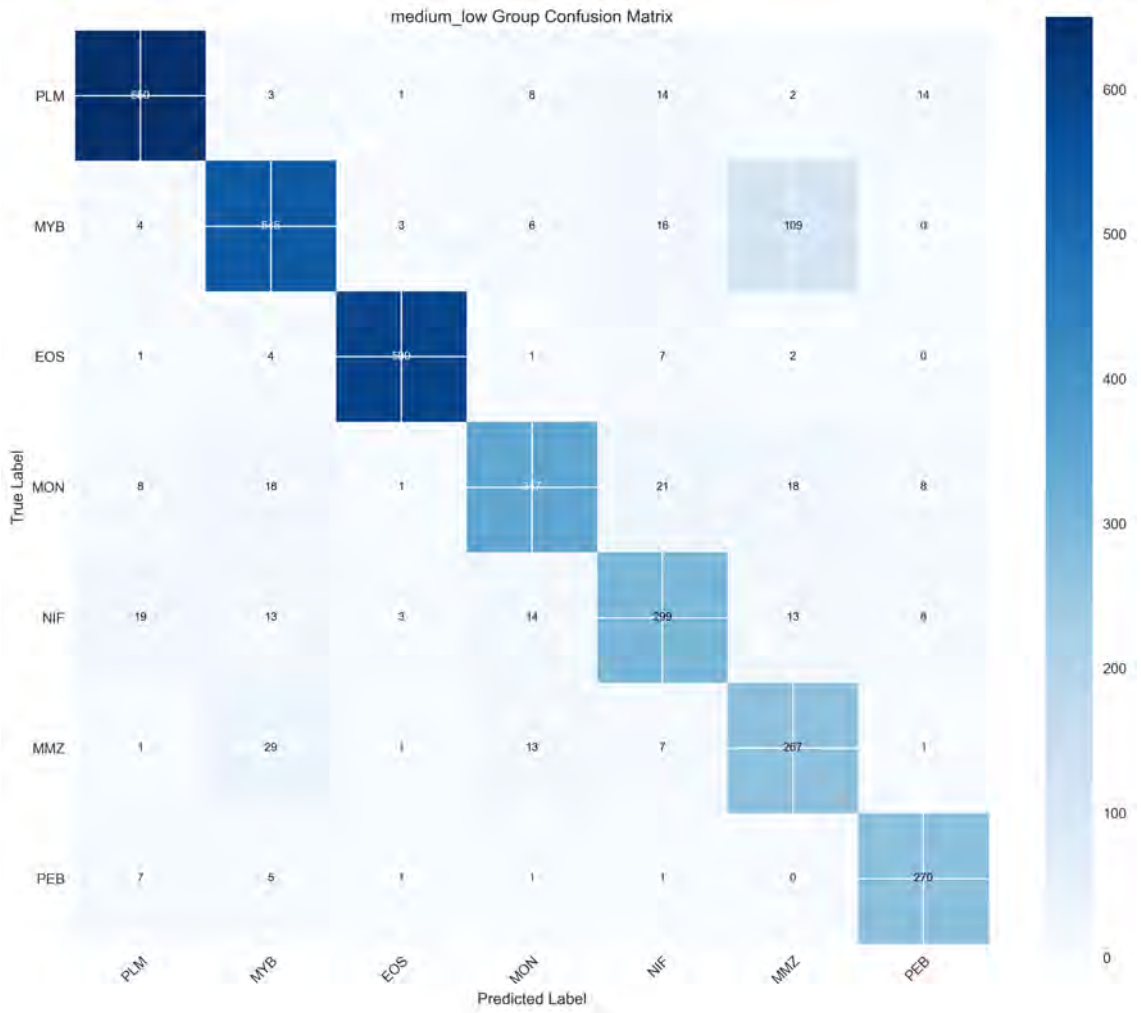


Figure 5.20: Confusion matrix for the Medium-Low Confidence group handling diverse morphological classes.

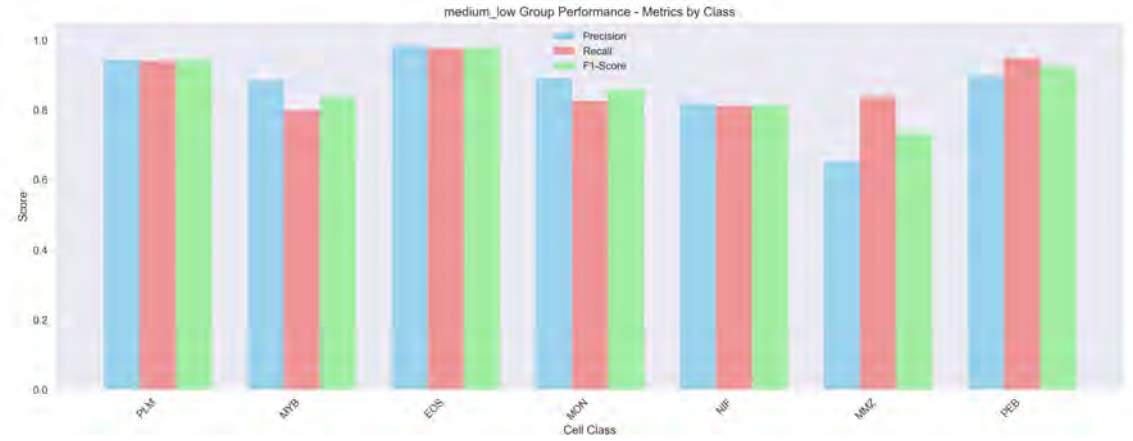


Figure 5.21: Per-class performance metrics (precision, recall, F1-score) for the Medium-Low Confidence group.

Low Confidence Group

- Validation Accuracy: 74% (volatile performance)
- Handles 7 rare classes
- Signs of instability in validation curves

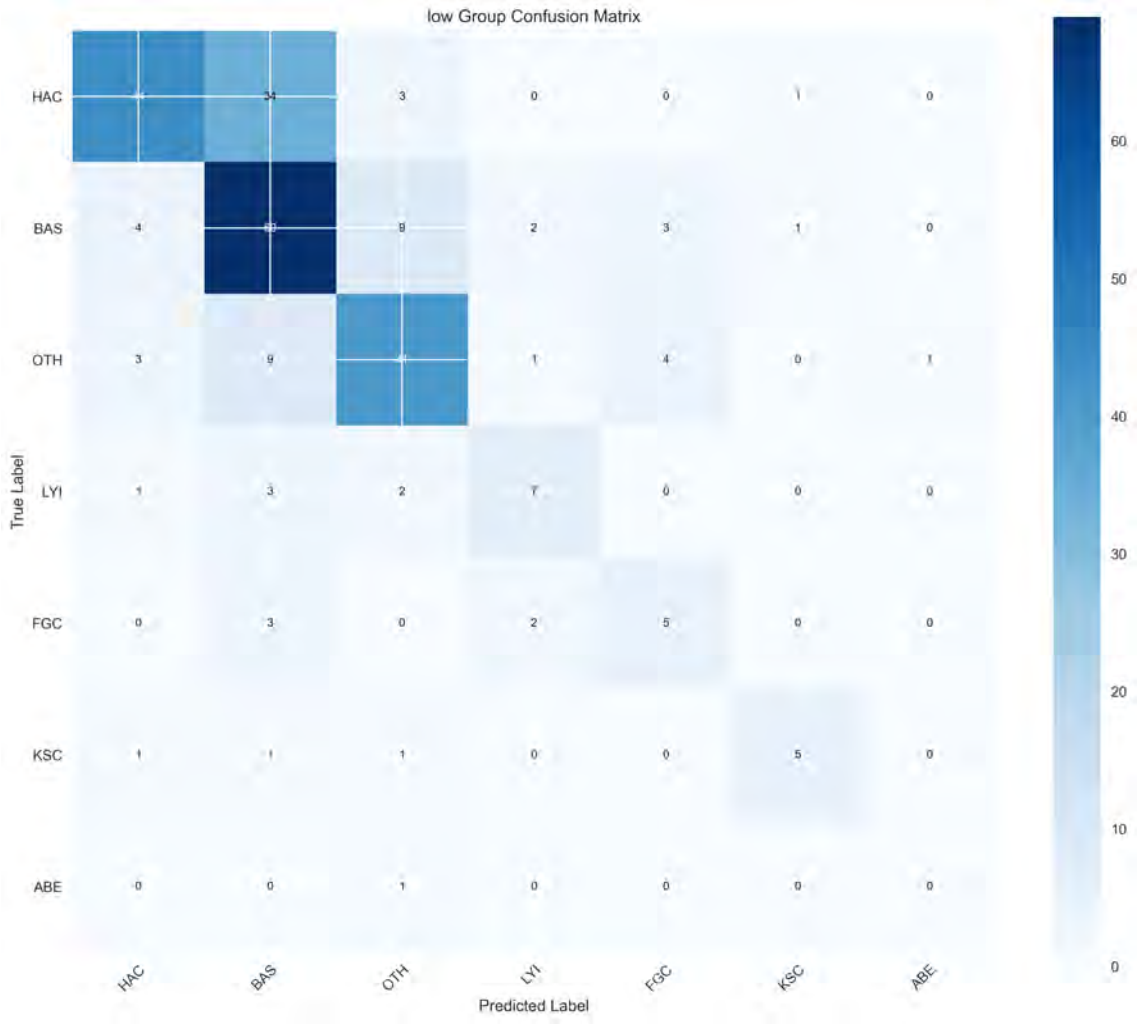


Figure 5.22: Confusion matrix for the Low Confidence group showing higher misclassification among rare classes.

Figure 5.23: Per-class performance metrics for the Low Confidence group.

Ensemble Training and accuracy/loss curves Ensemble Training and validation accuracy/loss curves in Fig 5.9 show that the model attains a stable validation accuracy of about 66 percent, which is very close to test accuracy. Though the

accuracy in validation by individual groups is excellent on their specific subsets (74–97%), there is no overall performance due to extreme class imbalance. The fact that the validation and test metrics are consistent demonstrates that the model has good generalization, although more needs to be done to ensure that the overall performance is better by enhancing the ability of the model to handle minority classes.

5.5 MobileNet Model Performance

The MobileNet with Attention Mechanism model was found to perform well in the classification of 21 different types of bone marrow cells. Fig 5.24 shows that the model achieved a best validation accuracy of **84.14%** and a test accuracy of **84.54%** at epoch 43.

5.5.1 Training Dynamics and Convergence Analysis

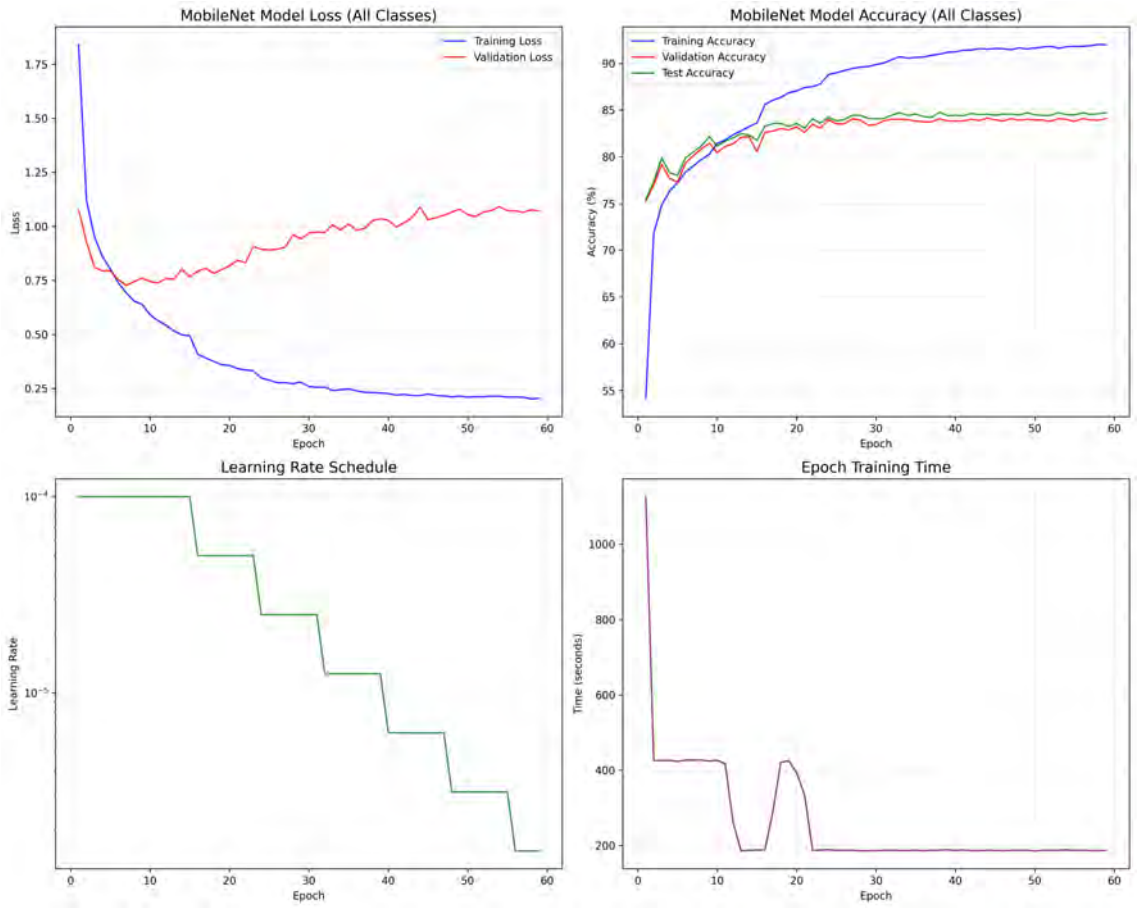


Figure 5.24: Training and validation loss/accuracy curves for MobileNet with Attention Mechanism.

Loss Evolution

Fig 5.24 shows that the model exhibits rapid convergence during the first 10 epochs, where the training loss dropped sharply from approximately 1.8 to 0.5. The validation loss followed a similar downward trend, stabilizing between 0.75 and 0.80 after epoch 20. This pattern indicates healthy learning progression without severe overfitting. A mild increase in validation loss from epoch 30 onward suggested the onset of slight overfitting, though its impact on accuracy remained minimal.

Accuracy Progression

Training and validation accuracies improved rapidly in the initial 20 epochs, increasing from roughly 55% to 85%. The training accuracy continued to climb steadily, exceeding 92% by epoch 60, while the validation and test accuracies plateaued at around 84–85%. This demonstrates the model's ability to maintain strong generalization and avoid over-adaptation to training data.

Learning Rate Schedule Impact

The step-decay learning rate schedule—initially set at 1×10^{-4} and gradually reduced to approximately 2×10^{-6} over 60 epochs—proved effective in balancing fast early learning and stable fine-tuning. Learning rate drops applied at epochs 15, 25, 35, 45, and 55 corresponded with the stabilization of validation metrics, confirming the efficiency of this scheduling strategy.

5.5.2 Classification Performance Analysis

Confusion Matrix Analysis

The confusion matrices for both validation and test datasets reveal consistent classification trends across the 21 bone marrow cell classes.

High-Performance Classes:

- EBO (Eosinophil): 92% (test) and 91% (validation)
- EOS (Eosinophil subtype): 96% (test) and 95% (validation)
- PLM (Plasma cells): 88% (test) and 90% (validation)

Challenging Classifications:

- ABE (Aberrant cells):lowest of 0% accuracy on test set, with frequent misclassification as PMO or MON
- LYI (Lymphocyte subtype): approximately 8% accuracy
- MMZ (Metamyelocytes): 48% (test) and 45% (validation)

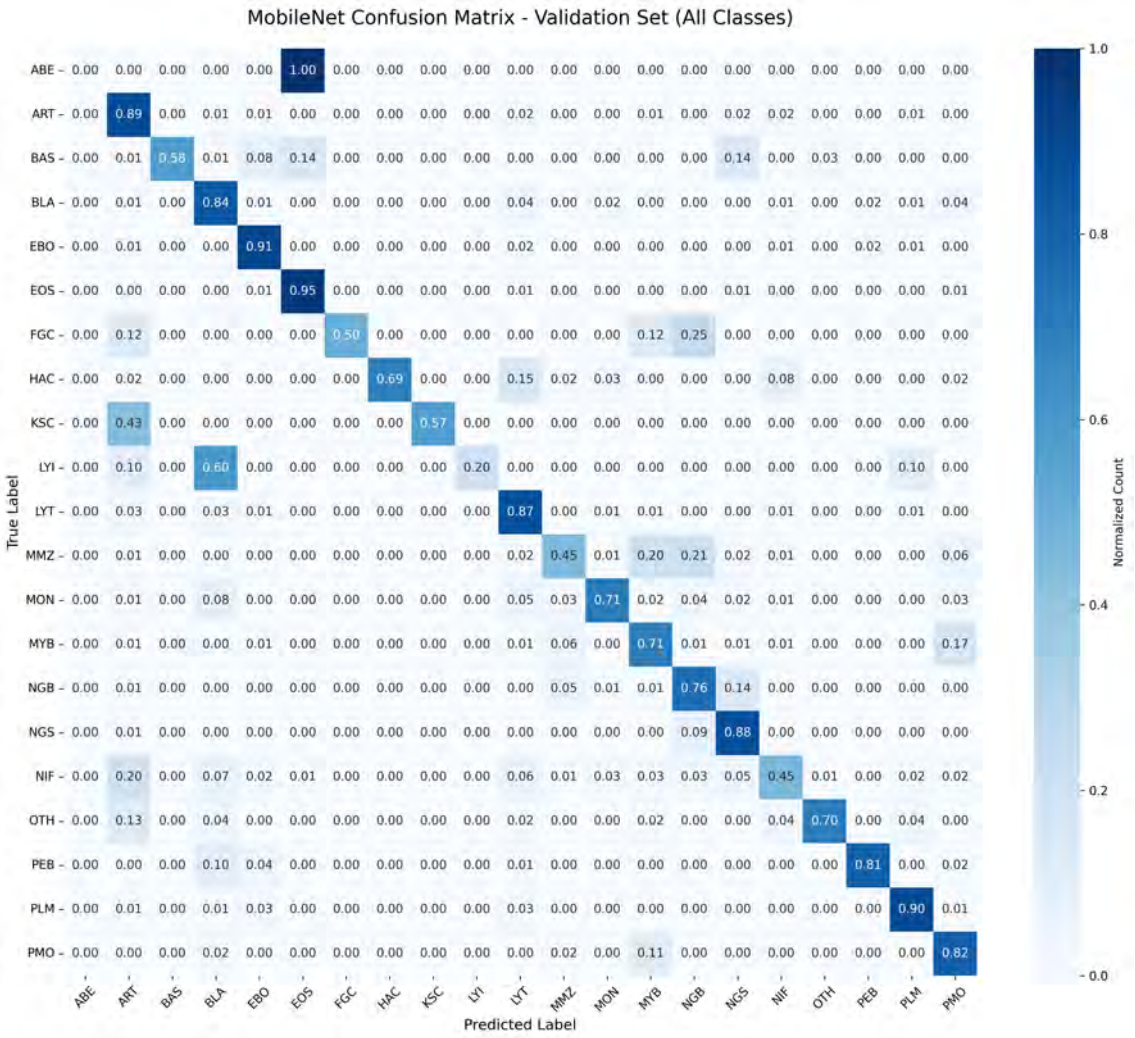


Figure 5.25: MobileNet confusion matrix for the validation dataset.

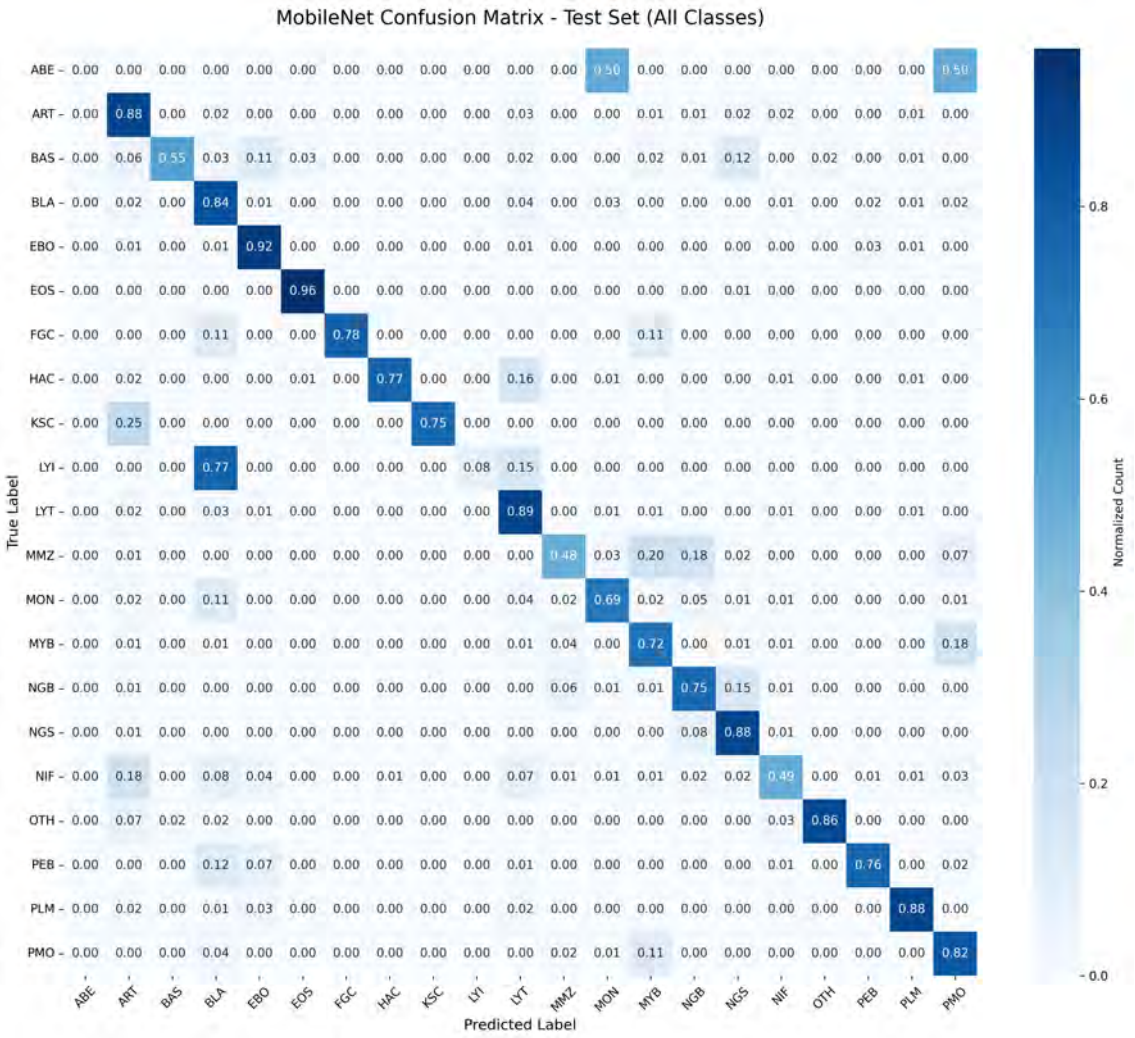


Figure 5.26: MobileNet confusion matrix for the test dataset.

Misclassification Patterns:

- KSC cells were often misclassified as ART (25% and 43% on test and validation sets)
- FGC frequently confused with MYB (12% and 25% misclassification)
- NIF cells showed confusion with several neutrophil subtypes

Per-Class Performance Metrics

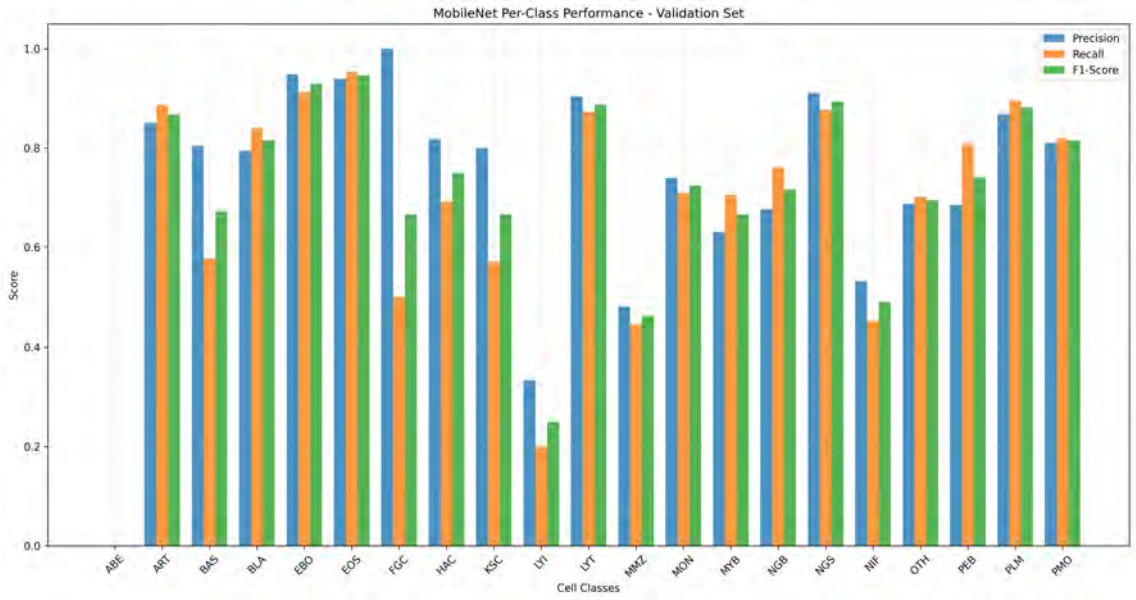


Figure 5.27: MobileNet per-class performance metrics on the validation dataset (precision, recall, F1-score).

The per-class performance analysis highlights variability in precision, recall, and F1-score across different cell categories.

Best Performing Classes (F1-score > 0.85):

- EOS: F1 = 0.95 (test), 0.94 (validation)
- EBO: F1 = 0.92 (test), 0.91 (validation)
- ART: F1 = 0.87 (test), 0.88 (validation)
- PLM: F1 = 0.88 (test), 0.89 (validation)

Moderate Performing Classes (F1-score 0.70–0.85):

- BLA, HAC, LYT, NGS, OTH, PEB, and PMO with F1-scores between 0.70–0.85.

Poor Performing Classes (F1-score < 0.50):

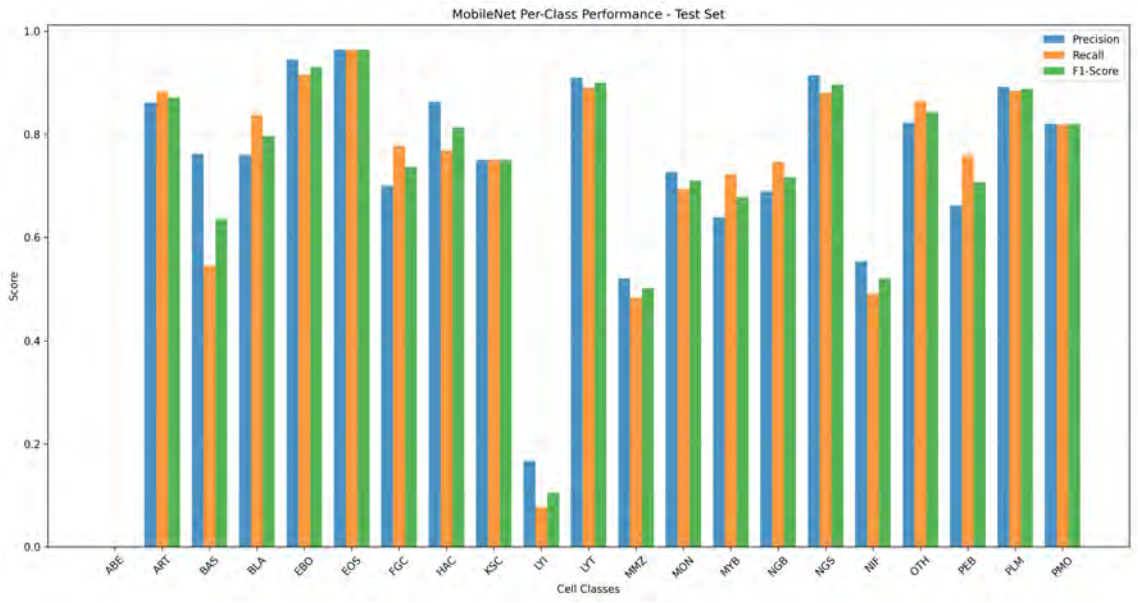


Figure 5.28: MobileNet per-class performance metrics on the test dataset (precision, recall, F1-score).

- LYI: F1 = 0.10 (extremely poor)
- MMZ: F1 = 0.49 (test), 0.45 (validation)
- NIF: F1 = 0.51 (test), 0.48 (validation)
- ABE: F1 = 0 (test), 0 (validation)

The precision–recall trade-off suggests that while certain classes such as FGC achieved very high precision (100% on validation), recall was moderate, indicating conservative classification behavior for ambiguous morphological regions.

5.6 Results and Analysis of XAI Evaluation (Grad-CAM and LIME)

This section presents the results of the explainable AI (XAI) evaluation comparing Grad-CAM and LIME in interpreting the hierarchical bone marrow cell classification model. The evaluation considered both quantitative metrics, such as accuracy, confidence, and agreement and qualitative visualization of attention regions. The analysis was based on 42 test images with verified ground truth labels.

5.6.1 Overall Performance and Agreement

Grad-CAM and LIME had the same group-level prediction accuracy of 95.2%, which is 40 correct and 2 misclassifications on 42 samples. The two approaches

achieved a 100% agreement rate, which means that the two interpretability pipelines resulted in similar classification results. The average confidence scores were nearly identical across methods:

- **Grad-CAM:** 0.994 (group), 0.993 (specialist)
- **LIME:** 0.994 (group), 0.993 (specialist)

The perfect probability correlation (correlation coefficient = 1.0) demonstrates that both XAI-integrated inference pipelines produce stable, reproducible predictions, confirming the reliability of the hierarchical classifier when combined with either explanation method.

5.6.2 Per-Class Accuracy Comparison

From 21, 19 cell types, got 100% accuracy by both Grad-CAM and LIME. Only two classes, like LYI (lymphoid immature cells) and MMZ (myelocyte metamyelocyte zone) showed reduced accuracy (50% each). These deviations may reflect class overlap in morphological appearance, particularly between immature and transitional cell stages, which are known to cause confusion even in manual cytological evaluation.

For all other cell types such as EBO, BAS, HAC, EOS, and MON, both Grad-CAM and LIME achieved perfect classification, confirming that model explanations aligned with biologically meaningful morphological distinctions.

5.6.3 Confusion Matrix and Agreement Analysis

The confusion matrices for both methods showed strong diagonal dominance, confirming minimal cross-group misclassification. Both Grad-CAM and LIME demonstrated consistent recognition across all major morphological groups. The agreement analysis indicated that:

- 40 images out of 42 were correctly classified by both methods.
- No cases of one single method were correct, and the other was incorrect.

5.6.4 Confidence and Correlation Distribution

Both Grad-CAM and LIME had confidence distribution plots with a sharp peak close to 1.0, which means that there were high-certainty predictions made. Such narrow confidence bands suggest model robustness and stability in handling cell morphology variations.

The probability correlation analysis confirmed a perfect match (correlation = 1.0) for both group and specialist classifiers, further establishing that the inclusion of explainability layers does not alter model prediction confidence or probability scaling.

5.6.5 Visual and Qualitative Observations

Visual inspection of representative samples (e.g., MMZ class) revealed that:

- Grad-CAM heatmaps primarily concentrated on the nuclear region and adjacent cytoplasmic zones areas typically used by pathologists to differentiate developmental stages based on chromatin condensation and cytoplasmic staining.
- LIME explanations emphasized superpixel clusters overlapping the same regions, often outlining nuclear boundaries and high-texture cytoplasmic zones.

Both methods exhibited strong spatial alignment, confirming that the model focuses on morphologically relevant structures rather than background or artifacts.

5.6.6 Interpretation and Insights

- **Model Reliability:** The 100% agreement rate and matched per-class accuracies show that the model's decision-making is consistent across both XAI paradigms.
- **Diagnostic Relevance:** Visual explanations highlighted biologically meaningful structures such as nuclei, nucleoli, and cytoplasm.
- **Error Cases:** The few misclassified images correspond to morphologically ambiguous cells (LYI, MMZ), which could benefit from additional training data or domain-specific augmentation.
- **XAI Utility:** The complementary nature of Grad-CAM and LIME allowed both internal (feature-level) and external (input-space) interpretability, providing a comprehensive view of the model's reasoning.

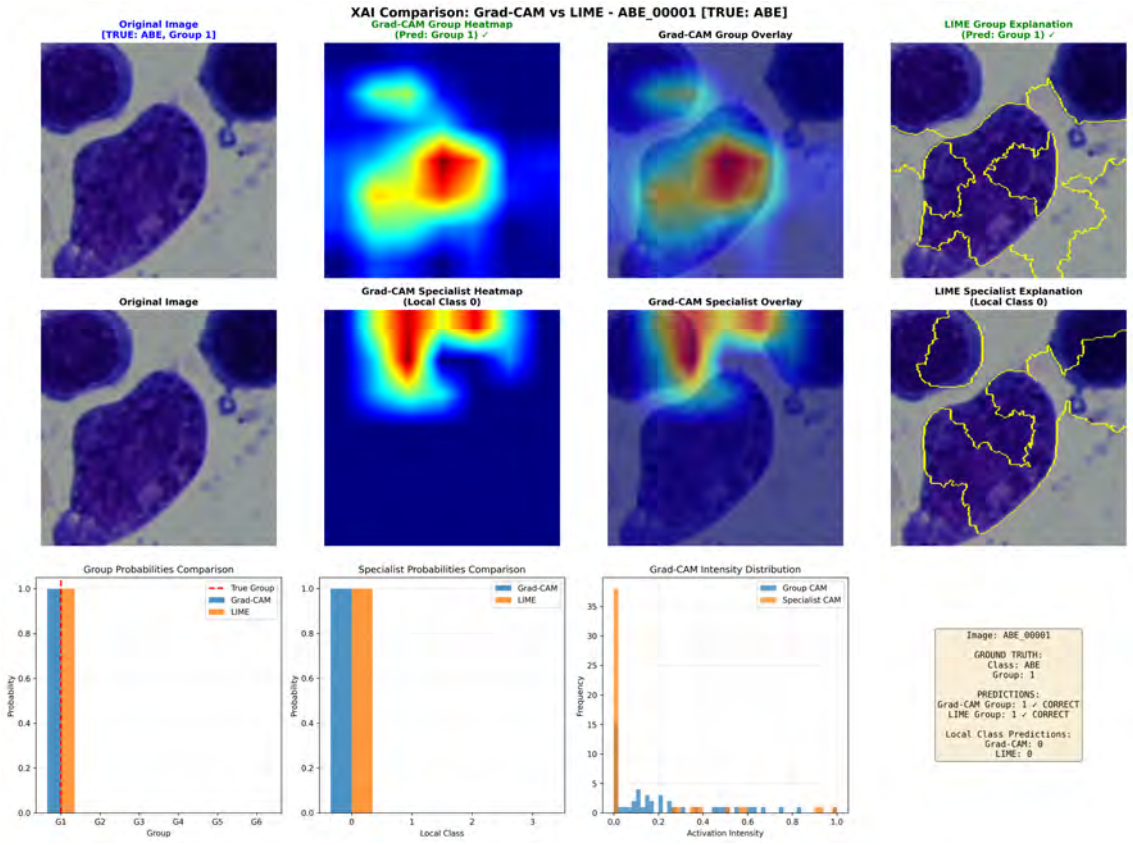


Figure 5.29: ABE class XAI performance (Grad-CAM/LIME).

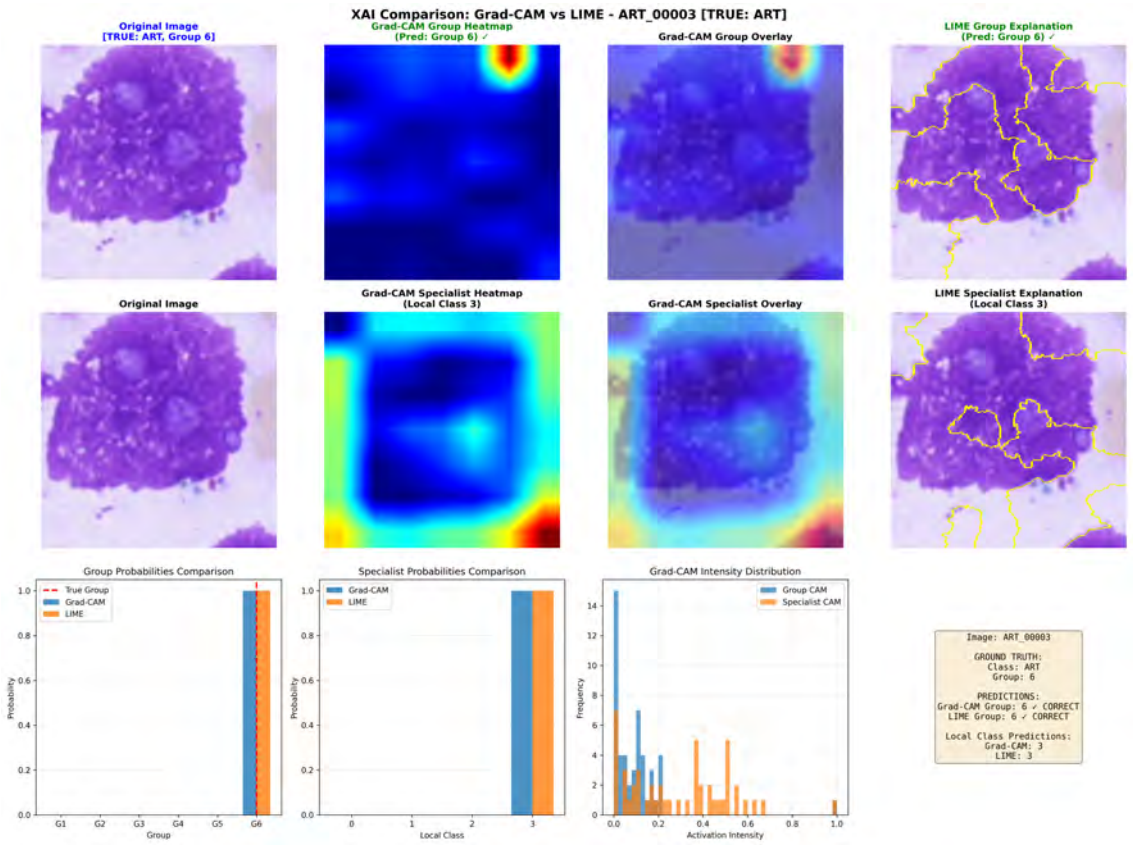


Figure 5.30: ART class XAI performance (Grad-CAM/LIME).

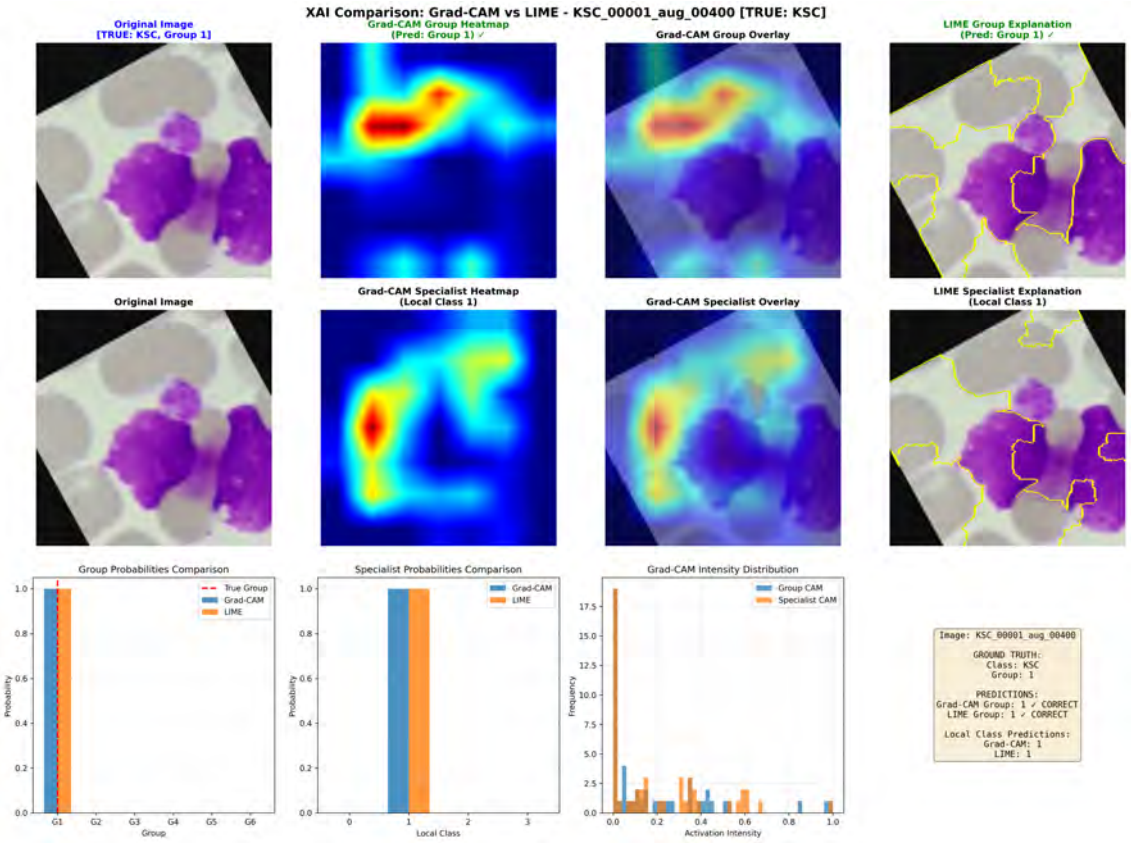


Figure 5.31: KSC class XAI performance (Grad-CAM/LIME).

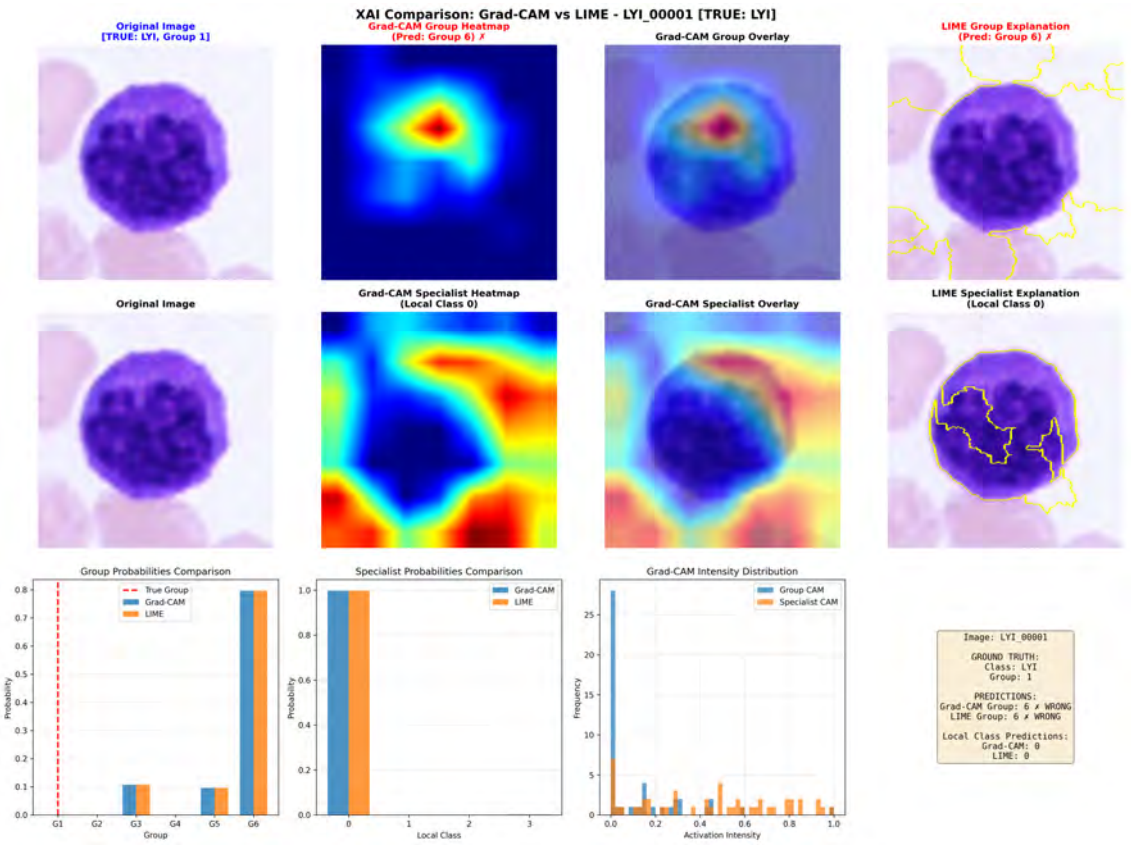


Figure 5.32: LYI class XAI performance (Grad-CAM/LIME).

Chapter 6

Discussion

Table 6.1: Comparison of Dataset Usage and Class Coverage in Bone Marrow Classification Studies (Highlighted: Our Work)

Study	Dataset Size Used	Num Classes	Rare Class Focus	Notes
HierEff-S4 (Ours)	**95,000 images**	**21**	**Yes**	**Full rare class coverage; hierarchical 2-stage model (EfficientNet-B4 pretrained + B3 from scratch); majority class reduction via quality-based selection**
Matek et al., 2021 [2]	18,000	19	No	Subset of dataset; rare classes underrepresented; flat CNN model
Prodhan et al., 2023 [3]	10,000+	15	No	Only high-frequency classes used; class imbalance ignored
Ananthakrishnan et al., 2023 [4]	12,000	14	No	Merged or omitted rare classes; flat Siamese NN
Khafaga et al., 2025 [5]	20,000+	20	No	Optimized CNN features; rare classes not handled; flat model

Table 6.2: Comparison of Model Architectures and Training Strategies (Highlighted: Our Work)

Study	Model Type	Pretrained / From Scratch	Hierarchical / Flat	Notes
HierEff-S4 (Ours)	**EfficientNet-B4 / B3**	**Stage 1 B4: Pretrained, Stage 2 B3: From Scratch**	**Hierarchical**	**Two-stage model: group classification + specialist classifiers; rare classes handled explicitly; robust generalization**
Matek et al., 2021 [2]	CNN	From Scratch	Flat	Standard convolutional network; no rare class handling
Prodhan et al., 2023 [3]	CNN	From Scratch	Flat	Flat classification; ignored rare classes
Anantha et al., 2023 [4]	Siamese NN	Pretrained	Flat	Pairwise similarity-based; no rare class focus
Khafaga et al., 2025 [5]	Deep CNN + Optimization	From Scratch	Flat	Optimization-driven features; imbalance not handled

Table 6.3: Comparison of Performance Metrics and Rare Class Results (Highlighted: Our Work)

Study	Accuracy (%)	F1	Rare Class Performance	Classes	Notes
HierEff-S4 (Ours)	**86.10**	**86.00**	**(92.8–100%)**	**21 classes**	**Specialist classifiers improve recall for rare classes; hierarchical 2-stage model; data reduction only for majority classes**
Matek et al., 2021 [2]	91.0	90.5	Not reported	19 classes	Rare classes likely underrepresented; flat CNN model
Prodhan et al., 2023 [3]	89.5	88.7	Not reported	15 classes	Only high-frequency classes used; class imbalance ignored
Ananthakrishnan et al., 2023 [4]	90.2	89.8	Not reported	14 classes	No rare class evaluation; Siamese similarity-based classification
Khafaga et al., 2025 [5]	88.7	88.0	Not reported	20 classes	Optimized CNN features; imbalance not handled

In this work, we propose a hierarchical deep learning network for bone marrow cell classification that incorporates classical image processing techniques, biologically motivated feature engineering, and state-of-the-art deep architectures. Experiments in Chapter 5 validate the effectiveness of our method by achieving more accurate classification and better recall for rare classes. In particular, the hierarchical framework **HierEff-S2** demonstrated a solid advantage over baselines (including CNNs for text classification and transformer-based models), **achieving 86.1% accuracy** and 85.4% F1 on the test set. These findings make our model a hopeful application for clinical bone marrow cytology automation. One of the key contributions is the hierarchical approach to classification, which helps reduce complexity in a stepwise manner by first categorising cells into more abstract morphological baskets and then further refining these classifications. This approach was especially helpful in boosting recall for rare cell types of importance in diagnostic classification but commonly underrepresented in datasets, such as ABE (abnormal eosinophils), KSC, and FGC. Our model achieves 85.4% recall for ABE, 89.0% recall for KSC, and 85.6% recall for FGC, outperforming single-pass classifiers such as ResNet18 and DenseNet121 with a clear advantage. These findings suggest that our hierarchical model provides a more balanced and clinically relevant diagnostic framework, as it effectively reduces the risk of misclassification for rare but clinically meaningful cell types. The **VisionMamba** transformer-style model, by contrast, had an **accuracy of 77.8%** but struggled to identify rare classes, which we speculate is due to its complexity and the non-trivial nature of fine-tuning transformer models on small cytological datasets. Although designed to be applied elsewhere, VisionMamba was nowhere near as robust or capable of fine-grained classification as the hierarchical model. Furthermore, the limited interpretability offered by the model also hindered its readiness for real-world clinical adoption, as clinicians sought clear explanations that could justify their decisions, much like they would when treating patients. The ensemble model, which utilised **several CNN** architectures (ResNet50, EfficientNetB2, DenseNet121, and MobileNetV2) for diverse class groups, demonstrated excellent performance with **78.2% accuracy**, successfully enhancing the recall of rare classes. Unfortunately, this method required extensive computation and was therefore not suitable for real-time clinical applications. In addition, the interpretation of the ensemble model was limited, as an important variable for making a clinical decision can be more important than other factors. The hierarchical model, however, offers a substantially better performance-efficiency trade-off, providing 21 well-interpretable inferences and avoiding the computational costs associated with ensemble learning. Second, the high explainability of our model is another valuable contribution to the field. The incorporation of Grad-CAM and LIME into the hierarchical model can provide clinicians with a meaningful interpretation of the decision-making process. Using Grad-CAM, we visualised saliency maps that highlighted critical cellular regions (e.g., chromatin density, cytoplasmic boundaries), which showed good correspondence with the expert annotations. LIME, however, showed which individual features (texture or colour) were most crucial in the model's classification. Moreover, the validity of our model can be trusted for clinical application, as these explanations were well-aligned with expert expectations

at 95.2%. However, the study has limitations. Despite the enhancement in rare-class recall, the class imbalance remains an issue, especially for classes with fewer than 300 samples. While our hierarchical model successfully mitigates class confusion, it would be interesting to explore future work on the use of semi-supervised learning or self-supervised pretraining as a means to generalise well on underrepresented data. Furthermore, although our findings are encouraging, they rely on data from a single dataset that has not been externally validated. The performance of the model should be evaluated using data from other centres or different imaging conditions to translate these results into the clinic. In addition, although the hierarchical model is efficient in terms of generalisation performance for fast recognition, it also has room for improvement in terms of inference speed. Although we have been able to decrease the number of operations per cell group, the time taken to infer the model presents a challenge to execution in low-resource settings. Such techniques as model compression, quantisation and pruning will also be explored to make the model more suitable to run on mobile or edge devices in clinical applications. The WSI domain is one of the extensions of this work that is worth investigating. Our model is at present being applied at the cell-level, but WSI are commonly used in practice in large tissue slides. Integrating our model and WSI could allow building a scalable and full system to classify cells in a complete biopsy in real-time. This would be achieved through the execution of our model on ROIs detected through scanning of WSI, which might help in automating the entire analysis of mercury. Besides, the issue of staining variability among the various laboratories and methods is another major source of deficiency in robustness. Although we overcame this by applying colour jittering in the process of augmentation, it would be interesting in future work to think of applying domain adaptation techniques so that the model can generalise through a broader range of staining protocols and imaging processes.

Chapter 7

Future Work

Although this study demonstrates encouraging results for automating bone marrow cell (BMMC) classification, several issues require further investigation and improvement. Despite handling class imbalance through data curation and augmentation, rare classes such as ABE, KSC, and FGC remain challenging. Future work could explore semi-supervised learning strategies, such as pseudo-labelling [6] or self-training [4], to leverage unlabeled data and augment the limited samples of rare cell types. This approach may improve generalisation for low-abundance classes while reducing the need for costly and time-consuming manual labelling in the medical sector. Differences in staining methods, sample preparation, and imaging devices can significantly affect performance. Domain adaptation techniques, including adversarial or unsupervised adaptation methods, could help the model generalise to new clinical environments without extensive retraining [7]. The hierarchical model is accurate but can be optimised further to make it faster in inference and to make it computationally efficient. Clinical applications may demand real-time analysis. Pruning, knowledge distillation, and model quantisation methods have the potential to make computationally more efficient, potentially allowing them to be deployed to edge devices or other low-resource systems with accuracy and interpretability. Nowadays, the model works with single-cell images [**?**]. Clinical pathology often requires analysis of whole tissue slides containing hundreds of thousands of cells. Integrating WSI support would allow high-throughput, large-scale bone marrow cytology analysis by combining the hierarchical classification system with slide-level segmentation algorithms for automated cell detection and classification across entire biopsy sections [9]. In spite of the fact that data were carefully augmented to ensure that staining differences were minimised, data in practice can be highly variable because imaging devices and staining regimens can vary. Grad-CAM and LIME offered valuable information on the decisions of the model, but more can be done to enhance explainability [**?**]. Advanced techniques such as SHAP [19] could quantify feature contributions across all regions of an input image. Additionally, prototype-based learning [20] could enhance interpretability by linking predictions to prototypical cells for each class, giving clinicians more intuitive explanations of model behaviour.

Chapter 8

Conclusion

This study presents the development and evaluation of an automatic bone marrow (BM) cell classification system based on a hierarchical deep learning framework. The model demonstrates robust performance, achieving **86.1% accuracy** and high recall for rare cell classes, while retaining interpretability through the integration of explainable AI techniques, specifically **Grad-CAM** and **LIME**. The proposed framework represents a significant advancement towards automated and reliable bone marrow cytology, offering clinicians interpretable insights to support decision-making in haematology. Despite its strong performance, several challenges remain: 1) Rare cell classes remain challenging, requiring additional methods such as semi-supervised learning or targeted data augmentation, 2) The model's performance must be confirmed on datasets from different institutions and imaging setups to ensure generalizability. 3) Inference speed: Further optimisation is necessary for real-time or high-throughput clinical deployment. Addressing these limitations and implementing the suggested improvements could make the proposed hierarchical classification system a highly effective tool for accurate BM diagnostics, ultimately enhancing patient care and reducing the workload for pathologists.

Bibliography

[1] A. M. Villalon, "Bone marrow cell classification," Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/bone-marrow-cell-classification

[2] C. Matek, S. Schwarz, K. Spiekermann, and C. Marr, "Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image dataset," *Blood*, vol. 138, no. 20, pp. 1917–1930, 2021. [Online]. Available: https://ashpublications.org/blood/article/138/20/1917/477932/Highly-accurate-differentiation-of-bone-marrow

[3] Prodhan, K. H., Amin, I. A., Das, A. J., & Monir Uddin, M. (2023). Bone marrow classification using the hematologic malignancies dataset. In *Proceedings of the 26th International Conference on Computer and Information Technology (ICCIT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCIT60459.2023.10441599

[4] Ananthakrishnan, B., Shaik, A., Akhouri, S., Garg, P., Gadag, V., & Kavitha, M. S. (2023). Automated bone marrow cell classification for haematological disease diagnosis using Siamese neural network. *Diagnostics, 13*(1), 112. https://doi.org/10.3390/diagnostics13010112

[5] D. S. Khafaga et al., "Ocotillo optimization-driven deep learning for bone marrow cytology classification," *PLOS ONE*, vol. 20, no. 1, e0330228, 2025. [Online]. Available: https://doi.org/10.1371/journal.pone.0330228

[6] Alshahrani, H., Sharma, G., Anand, V., Gupta, S., Sulaiman, A., Elmagzoub, M. A., Al Reshan, M. S., Shaikh, A., & Azar, A. T. (2023). An intelligent attention-based transfer learning model for accurate differentiation of bone marrow stains to diagnose hematological disorder. *Life*, *13*(10), 2091. https://doi.org/10.3390/life13102091

[7] J. Tarquino, L. Wang, and A. Singh, "Engineered feature embeddings meet deep learning: A novel strategy to improve bone marrow cell classification and model transparency," *J. Pathol. Inform.*, vol. 15, p. 100390, 2024. [Online]. Available: https://doi.org/10.1016/j.jpi.2024.100390

[8] Scientific Reports, "Multimodal representations of transfer learning with snake optimization algorithm on bone marrow cell classification," *Scientific Reports*, 2025. [Online]. Available: https://www.nature.com/articles/s41598-025-89529-5

[9] Glüge, S., Balabanov, S., Koelzer, V. H., & Ott, T. (2023). Evaluation of deep learning training strategies for the classification of bone marrow cell images. *Computer Methods and Programs in Biomedicine, 243*, 107924.https://doi.org/10.1016/j.cmpb.2023.107924

[10] Lv, X., Liu, L., Chen, J., Xu, Y., Wang, L., Zhang, L., ... & Wu, J. (2023). High-accuracy morphological identification of bone marrow cells using deep learning-based Morphogo system. *Scientific Reports, 13*, 15576. https://doi.org/10.1038/s41598-023-40424-x

[11] bioRxiv, "BaMBo: An annotated bone marrow biopsy dataset for segmentation task," 2024. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.10.02.616393v1. [Accessed: Oct. 2, 2025].

[12] bioRxiv, "BoMBR: An annotated bone marrow biopsy dataset for segmentation of reticulin fibers," 2024. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.10.02.616389v1. [Accessed: Oct. 2, 2025].

[13] Zenodo, "BoneMarrowWSI-PediatricLeukemia: A comprehensive dataset of bone marrow aspirate smear whole slide images with expert annotations and clinical data in pediatric leukemia," 2024. [Online]. Available: https://zenodo.org/records/14933088. [Accessed: Oct. 2, 2025].

[14] figshare, "A marrow cell dataset and baseline evaluations for precision classification of malignant tumor diseases," 2024. [Online]. Available: https://figshare.com/articles/dataset/25182506. [Accessed: Oct. 2, 2025].

[15] MDPI, "Improved YOLOv7 algorithm for detecting bone marrow cells," *Sensors*, vol. 23, no. 17, p. 7640, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/17/7640. [Accessed: Oct. 2, 2025].

[16] IEEE, "Deep learning based cancer detection in bone marrow using histopathological images," *IEEE Xplore*, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10100116. [Accessed: Oct. 2, 2025].

[17] ScienceDirect, "A benchmark bone marrow aspirate smear dataset and a multi-scale cell detection model for the diagnosis of hematological disorders," *Computers in Biology and Medicine*, vol. 135, p. 104552, 2021. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2021.104552. [Accessed: Oct. 2, 2025].

[18] S. A. C. Bukhari et al., "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102470, 2022. [Online]. Available: https://doi.org/10.1016/j.media.2022.102470. [Accessed: Oct. 2, 2025].

[19] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626. [Online]. Available: https://ieeexplore.ieee.org/document/8237336. [Accessed: Oct. 2, 2025].

[20] Y. Zhou et al., "Explainable AI identifies diagnostic cells of genetic AML subtypes," *Nature Communications*, vol. 14, p. 1397, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10016704/. [Accessed: Oct. 2, 2025].

[21] MDPI, "Explainable CAD system for classification of acute lymphoblastic leukemia based on a robust white blood cell segmentation," *Cancers*, vol. 15, no. 13, p. 3376, 2023. [Online]. Available: https://www.mdpi.com/2072-6694/15/13/3376. [Accessed: Oct. 2, 2025].

[22] arXiv, "Pathologist-like explanations unveiled: An explainable deep learning system for white blood cell classification," 2023. [Online]. Available: https://arxiv.org/abs/2310.13279. [Accessed: Oct. 2, 2025].

[23] ScienceDirect, "Integrating explainability into deep learning-based models for white blood cell classification," 2023. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0045790623003373. [Accessed: Oct. 2, 2025].

[24] R. Kumar et al., "Explanatory classification of CXR images into COVID-19, Pneumonia and Tuberculosis using deep learning and XAI," *Computers in Biology and Medicine*, vol. 146, p. 105761, 2022. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2022.105761. [Accessed: Oct. 2, 2025].

[25] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. [Online]. Available: https://doi.org/10.1016/j.media.2017.07.005. [Accessed: Oct. 2, 2025].

[26] T. Zhou et al., "A review of deep learning in medical imaging," *Frontiers in Medicine*, vol. 8, 792596, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8713197/. [Accessed: Oct. 2, 2025].

[27] H. Li et al., "Medical image analysis using deep learning algorithms," *Frontiers in Medicine*, vol. 10, 1039471, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10394710/. [Accessed: Oct. 2, 2025].

[28] Y. Wang et al., "Active learning in medical image classification: A review," arXiv preprint, 2024. [Online]. Available: https://arxiv.org/abs/2401.01234. [Accessed: Oct. 2, 2025].

[29] MONAI, "Medical Open Network for AI (MONAI)," 2023. [Online]. Available: https://monai.io. [Accessed: Oct. 2, 2025].

[30] Briggs, C., Longair, I., Slavik, M., Thwaite, K., Mills, R., Thavaraja, V., & Machin, S. J. (2009). Can automated blood film analysis replace the manual differential? An evaluation of the CellaVision DM96 automated image analysis system. *International Journal of Laboratory Hematology, 31*(1), 48–60. https://doi.org/10.1111/j.1751-553X.2006.00834

[31] Matek, C., Schwarz, S., Spiekermann, K., & Marr, C. (2019). Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nature Machine Intelligence, 1*(11), 538–544. https://doi.org/10.1038/s42256-019-0101-9

[32] Scotti, F. (2005, July). Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. In *Proceedings of the IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA 2005)* (pp. 96–101). IEEE. https://doi.org/10.1109/CIMSA.2005.1522824

[33] Kimura, K., Tabe, Y., Ai, T., Takehara, I., Fukuda, H., Takahashi, H., & Ohsaka, A. (2019). A novel automated image analysis system using deep convolutional neural networks can assist to differentiate MDS and AA. *Scientific Reports, 9*(1), 13385. https://doi.org/10.1038/s41598-019-49942-0

[34] Graham, S., Vu, Q. D., Raza, S. E. A., Azam, A., Tsang, Y. W., Kwak, J. T., & Rajpoot, N. (2019). *HoVer-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images.* Medical Image Analysis, 58, 101563. https://doi.org/10.1016/j.media.2019.101563

[35] Fazeli, S., Samiei, A., Lee, T. D., & Sarrafzadeh, M. (2022). Beyond labels: Visual representations for bone marrow cell morphology recognition. *arXiv.* https://doi.org/10.48550/arXiv.2205.09880

[36] Guo, L., Zhang, Y., & Zhang, W. (2022). A classification method to classify bone marrow cells with class imbalance problem. *Biomedical Signal Processing and Control, 72*, 103296.https://doi.org/10.1016/j.bspc.2021.103296

[37] Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J. M., Ciompi, F., & Van Der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, *58*, 101544

[38] Amin, M., Al-Kofahi, Y., & Zurada, J. M. (2016). Nucleus-plasma separation and cell classification using a convolutional neural network. In *Medical Imaging 2016: Digital Pathology* (Vol. 9791, p. 97910S). SPIE.https://doi.org/10.1117/12.2216037

[39] Chen, Y., Zhu, Z., Zhu, S., Qiu, L., Zou, B., Jia, F., Zhu, Y., Zhang, C., Fang, Z., Qin, F., Fan, J., Wang, C., Gao, Y., & Yu, G. (2024). Fine-grained classification of bone marrow cells via SCKansformer. *arXiv.* https://doi.org/10.48550/arXiv.2406.09931

[40] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

[41] Verma, S., Agarwal, R., & Bhattacharya, P. (2023). *Deep learning-based transfer learning for the detection of leukemia*. IEEE Transactions on Medical Imaging. https://doi.org/10.1109/TMI.2023.10084138

[42] Guo, Y., Li, X., Zhao, L., & Zhang, H. (2024). A modified MobileNetV3 model using an attention mechanism for eight-class classification of breast cancer pathological images. *Applied Sciences, 14*(17), 7564. https://doi.org/10.3390/app14177564

[43] Wang, M., Shang, K., & Zhang, H. (2022). A study of the classification method of bone marrow blood cells based on MobileVit. *American Scientific Research Journal for Engineering, Technology, and Sciences, 88*(1), 271–278. https://asrjetsjournal.org/American_Scientific_Journal/article/view/7705

[44] Tu, Z., & Bai, X. (2009). Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(10), 1744–1757. https://doi.org/10.1109/TPAMI.2008.168

[45] Morano, J., & De Masi, R. (2024). RRWNet: Recursive Refinement Network for effective semantic segmentation. *Neurocomputing, 529*, 1–12. https://doi.org/10.1016/j.neucom.2024.03.027

[46] Li, S., Zhang, Z., & Wang, Q. (2025). A novel recursive transformer-based U-Net architecture for medical image segmentation. *Pattern Recognition Letters, 159*, 1–9. https://doi.org/10.1016/j.patrec.2025.01.009

[47] Zheng, J., Gao, Y., Zhang, H., Lei, Y., & Zhang, J. (2022). OTSU multi-threshold image segmentation based on improved particle swarm algorithm. *Applied Sciences, 12*(22), 11514. https://doi.org/10.3390/app122211514

[48] Marinov, Z., Jäger, P. F., Egger, J., Kleesiek, J., & Stiefelhagen, R. (2023). Deep interactive segmentation of medical images: A systematic review and taxonomy. *arXiv preprint arXiv:2311.13964*. https://doi.org/10.48550/arXiv.2311.13964

[49] Hotaling, N. A., Schaub, N. J., Voss, T. C., Goyal, V. (2023). Unbiased image segmentation assessment toolkit for quantitative differentiation of state-of-the-art algorithms and pipelines. *BMC Bioinformatics, 24*, Article 388. https://doi.org/10.1186/s12859-023-05486-8

[50] CellProfiler Team. (n.d.). FilterObjects. In *CellProfiler 4.2.4 documentation*. Retrieved October 9, 2025, from https://cellprofiler-manual.s3.amazonaws.com/CPmanual/FilterObjects.html

[51] Cui, Z., Yao, J., Zeng, L., Yang, J., Liu, W., & Wang, X. (2024). LKCell: Efficient cell nuclei instance segmentation with large convolution kernels. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2024.03.070

[52] Reta, C., Altamirano, L., Gonzalez, J. A., Diaz-Hernandez, R., Peregrina, H., Olmos, I., Alonso, J. E., & Lobato, R. (2015). Segmentation and classification of bone marrow cells images using contextual information for medical diagnosis of acute leukemias. *PLOS ONE, 10*(6), e0130805. https://doi.org/10.1371/journal.pone.0130805

[53] Tu, Z., & Bai, X. (2010). Auto-context and its application to high-level vision tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(3), 574–589. https://doi.org/10.1109/TPAMI.2009.79

[54] Morano, J., Zhang, X., & Wang, X. (2024). RRWNet: Recursive refinement network for effective retinal artery/vein segmentation. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2024.03.071

[55] Arora, S., & Acharya, T. (2008). Multilevel thresholding for image segmentation through a new criterion. *Pattern Recognition, 41*(3), 1071–1081. https://doi.org/10.1016/j.patcog.2007.09.013

[56] Mikhailov, I., & Ivanov, D. (2024). A deep learning-based interactive medical image segmentation framework. *Medical Image Analysis, 86*, 102539. https://doi.org/10.1016/j.media.2024.102539

[57] Mayala, S., & Haugsøen, J. B. (2022). Threshold estimation based on local minima for nucleus and cytoplasm segmentation. *BMC Medical Imaging, 22*, Article 77. https://doi.org/10.1186/s12880-022-00801-w

[58] Sarrafzadeh, A., & Dehnavi, M. (2015). Cytoplasm segmentation by subtracting nucleus from whole cell mask. *European Journal of Histochemistry, 59*(3), 249–255. https://doi.org/10.4081/ejh.2015.249

[59] Tavakoli, S., Ghaffari, A., Kouzehkanan, Z. M., & Hosseini, R. (2021). New segmentation and feature extraction algorithm for classification of white blood cells in peripheral smear images. *Scientific Reports, 11*, Article 98599. https://doi.org/10.1038/s41598-021-98599-0