ESCI upf.
School of International Studies

MEMÒRIA DEL TREBALL DE FI DE GRAU DEL GRAU (ESCI-UPF)

Patch-Based and Single-Cell-Based Analysis of Peripheral Blood and Bone Marrow Histology for CHIP Detection

AUTOR/A: Laia Barcenilla Mañá	**NIA: 107694**
GRAU: Bachelor's Degree in Bioinformatics	

CURS ACADÈMIC: 2024-2025

DATA: 17 de juny de 2025

TUTOR/S: Dr Rao Muhammad Umer, Dr Carsten Marr

FULL DE RESUM DEL TREBALL DE FI DE GRAU DEL BDBI (ESCI-UPF)

TÍTOL DEL PROJECTE: Patch-Based and Single-Cell-Based Analysis of Peripheral Blood and Bone Marrow Histology for CHIP Detection	
AUTOR/A: Laia Barcenilla Mañá	**NIA: 107694**
CURS ACADÈMIC: 2024-2025	
DATA: 17 de juny de 2025	
TUTOR/S: Dr Rao Muhammad Umer, Dr Carsten Marr	

PARAULES CLAU (mínim 3)

- Català: aprenentatge profund, hematopoesi clonal de potencial indeterminat (CHIP), segmentació.
- Castellà: aprendizaje profundo, hematopoyesis clonal de potencial indeterminado (CHIP), segmentación.
- Anglès: deep learning, clonal hematopoiesis of indeterminate potential (CHIP), segmentation.

RESUM DEL PROJECTE (extensió màxima: 100 paraules per llengua)

- Català: L'estudi avalua l'ús de models d'aprenentatge profund per a la classificació de CHIP utilitzant imatges digitals de làmina completa de frotis de sang perifèrica i medul·la òssia. S'han implementat dues estratègies: una basada en parxes aplicada a les imatges de sang i medul·la òssia, i un enfocament basat en cèl·lules individuals (single-cell). El model es va desenvolupar sota un paradigma d'aprenentatge de múltiples instàncies, centrat en dos mètodes d'agregació: un basat en pesos d'atenció i un altre basat en una arquitectura de transformer. L'estudi conclou que els mètodes actuals no poden discriminar entre mostres CHIP positives i negatives.

- Castellà: El estudio evalúa modelos de aprendizaje profundo para clasificar CHIP usando imágenes digitales de lámina completa de frotis de sangre periférica y médula ósea. Se aplicaron dos metodologías: una basada en parches en imágenes de sangre y médula, y otra en células individuales (single-cell) solo en sangre periférica. El modelo sigue un paradigma de aprendizaje por múltiples instancias, con dos estrategias de agregación: pesos de atención y arquitectura transformer. Los resultados

indican que los métodos actuales no logran discriminar entre muestras CHIP positivas y negativas.

- Anglès: The study evaluates the use of deep learning models for the classification of CHIP using peripheral blood and bone marrow smear whole-slide images. Two methodologies were implemented: a patch-based approach applied to both blood and bone marrow images, and a single-cell-based approach applied exclusively to peripheral blood images. The model was developed under a multiple instance learning paradigm, focusing on two aggregation strategies: one based on attention scores and another based on a transformer architecture. The study concludes that current methods cannot discriminate between CHIP-positive and CHIP-negative samples.

Bachelor's Degree in Bioinformatics
(UPF–UPC–UB–UAB)
Final Degree Project

Patch-Based and Single-Cell-Based Analysis of Peripheral Blood and Bone Marrow Histology for CHIP Detection

Laia Barcenilla Mañá

Scientific Directors: Dr.Umer Rao Muhammad[1], Dr.Carsten Marr[1]

[1] Helmholtz Munich, Computational Health Center, Germany

Abstract

CHIP is a frequent genetic alteration associated with a higher risk of cardiovascular and hematologic malignancies. Currently, CHIP detection requires complex genetic analysis, which limits its routine clinical implementation. In this context, PB and BM smear WSIs represent a potentially accessible and economical source for automated CHIP diagnosis through the use of deep learning techniques.

This study evaluates the capacity of models to distinguish between CHIP-positive and CHIP-negative samples through the use of PB and BM WSIs. The analysis was conducted using visual encoders and aggregation mechanisms. The objective was to determine if CHIP status can be predicted from cell morphology of PB and BM smears using classification models from histopathology and hematology.

The results show that models based on PB patches, especially those including a quality control step, outperform PB single-cell analysis and BM approaches. In particular, the combination of DinoBloom and AB-MIL reached the best performance with an AUC of 0.59. In comparison, BM images showed lower performance and higher variability, probably due to their higher structural complexity.

These results suggest that PB patches provide a more favorable environment for the extraction of morphological characteristics relevant to CHIP classification. However, the relatively low values of the metrics and the high variability between partitions reveal that, with the current data and used architectures, models cannot capture sufficiently discriminative signals for a reliable diagnosis.

Overall, the study suggests that automated CHIP diagnosis through WSIs requires larger, better-balanced datasets and more optimized architectures to reach clinical application. The development of this methodology could significantly reduce costs and time for CHIP detection.

Abbreviations: Clonal Hematopoiesis of Indeterminate Potential (CHIP), Peripheral Blood (PB), Bone Marrow (BM), Whole Slide Image (WSI), Attention-Based Multiple Instance Learning (AB-MIL)

Supplementary information: Supplementary data are available at the GitHub link: https://github.com/laiaabm/final-degree-project

1 Introduction

Clonal Hematopoiesis of Indeterminate Potential (CHIP) is a premalignant state characterized by the clonal expansion of hematopoietic stem cells that have acquired one or more somatic mutations. DNMT3A, TET2, and ASXL1, all epigenetic regulators, are the most frequently mutated genes in CHIP [1, 2]. Affecting approximately 10% of individuals aged around 70 years [3], these alterations have been demonstrated to be age-associated, with prevalence increasing progressively with age [1]. Clinically, CHIP is linked to an elevated risk of hematologic malignancies, as well as a higher risk of cardiovascular disease, thromboembolism, and other fatal disorders [2, 3]. Numerous studies have corroborated these findings, underscoring the clinical relevance of CHIP [4] and highlighting the vital importance of early detection [5].

In the realm of peripheral blood (PB) and bone marrow (BM) analysis, microscopic evaluation by trained experts is still essential for the diagnosis of severe hematologic disorders, including acute myeloid leukemia and myelodysplastic syndromes [6]. Pathology, which is crucial for accurate diagnosis and treatment [7], relies on microscopic examination of morphological changes.

Recent advances in whole-slide scanning microscopy have enabled the digitalization of PB and BM smears into whole slide images (WSIs) [8]. This digitization has facilitated computer-aided diagnostics, especially through the use of artificial intelligence [9]. In particular, deep learning has proven highly effective in WSI analysis [10], thereby improving the precision of diagnostics [9, 11].

Multiple instance learning (MIL) has been applied as a key methodology in computational pathology. In this procedure, labels are assigned at the bag level (e.g., WSI) while learning from individual instances (e.g., patches) within each bag [12]. MIL is particularly well-suited for histopathology, where detailed annotations are challenging to obtain [13]. Several MIL variants such as architectures based on attention mechanisms and transformers [12], have been recently developed. These methods enable the capture of complex contextual and spatial relationships between instances with enhanced precision [14, 15].

Recent studies have successfully applied classification models to hematology images [16, 17]. Beginning with a large, high-resolution PB or BM smear WSI that is segmented into smaller patches, each of which undergoes quality control [18]. The machine learning workflow initiates with each individual patch being subjected to feature extraction by a vision encoder, including UNI [19] or DinoBloom [20]. Subsequently, the most relevant features of each patch are extracted and stored as d-dimensional vectors for further analysis. As a final step, the obtained feature vectors are aggregated using two distinct methodologies, attention-based deep MIL [21] and transformer-based correlated MIL [22], for final classification, ultimately determining whether the WSI comes from a CHIP-positive or CHIP-negative patient.

Currently, the majority of computational pathology approaches focus on learning features at the patch or region level from WSIs [23–26]. Although these techniques have improved machine learning's diagnostic capacity, they often overlook certain valuable insights that could be gained through single-cell analysis of white blood cells [27]. Recent publications have demonstrated the effectiveness of neural cellular automata models for image detection and segmentation, especially in hematology applications [28–31]. Adopting this approach, each white blood cell was isolated from every PB patch, enabling single-cell-level analysis. Following this methodology, our goal was to identify more complex cellular characteristics that might enhance diagnostic precision.

CHIP detection is currently dependent on gene expression profiling. Despite their accuracy, these approaches are costly, time-consuming, and remain inconsistently applied in clinical settings [32]. The development of a computational pathology model capable of identifying CHIP mutations directly from WSIs could substantially increase diagnostic accessibility.

On the basis of these developments, this study aims to investigate whether CHIP mutation status can be directly predicted from WSIs of PB and BM smears using computational pathology methods. To this end, two machine learning approaches have been implemented: a patch-based approach applied to both PB and BM smears, and a single-cell-based approach focused exclusively on PB images.

2 Methods

2.1 Data Collection

The dataset employed for this study was provided by Dr. Judith Hecker (Technical University

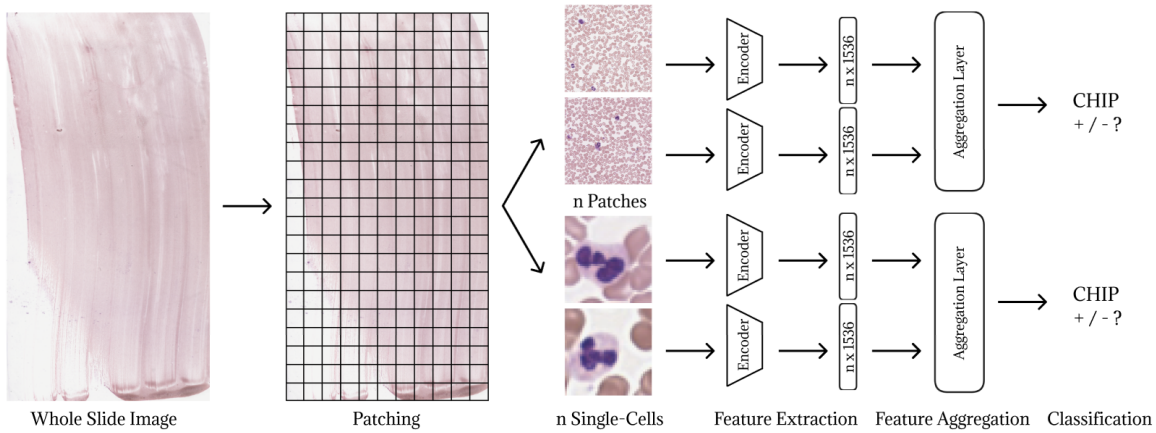


Figure 1: Overview of the analytical pipeline. Beginning with BM or PB WSI, the process initiates with patching. Subsequently, two alternative paths are undertaken: direct utilization of PB or BM patches in the patch-based approach or segmentation of PB patches for the single-cell analysis. Finally, feature extraction is performed using UNI and DinoBloom encoderss, followed by feature aggregation through AB-MIL and TransMIL deep learning models for the final classification of CHIP.

of Munich) and originates from Carsten Marr's project at Helmholtz Munich. It comprises high-resolution WSIs of hematoxylin and eosin (H&E)-stained BM and PB smears, scanned at 40× magnification.

The samples were obtained from femoral head donations of 19 CHIP-positive and 19 CHIP-negative patients. Each donation includes four stained BM smears and two stained PB smears. The CHIP-positive samples comprise 66 images from the BM and 34 from the PB, while the CHIP-negative WSIs consist of 38 from the BM and 23 from the PB.

2.2 WSI Patching

WSIs from BM and PB smears were divided into 224×224 non-overlapping patches. Approximately 3,200 patches were extracted from every BM WSI and around 40,000 patches from each PB WSI. This patch-based segmentation facilitates efficient handling of large-scale image data by enabling focused analysis on specific regions within each slide. On average, a BM WSI weighs between 3,088,384 KB and 2,365,009,920 KB, while a PB WSI ranges from 1,077,022,720 KB to 7,782,899,712 KB. In comparison, each individual patch image has an approximate size of 80 KB. To summarize the analysis process, Figure 1 provides a general overview of the workflow conducted.

2.3 Quality Control

A quality control (QC) based on the Haemorasis approach [18] was performed on each PB patch. The objective was to discriminate clinically relevant tiles from those that did not provide reliable information or could adversely affect subsequent analyses. The QC model was based on DenseNet121, a pretrained convolutional neural network fine-tuned on a PB dataset. During the analysis, each patch was evaluated by the QC model, which discarded tiles with either too low cell concentration (very few or no cells), too high cell concentration (very dense or overlapping cells), or blurry images. This process ensured that the retained tiles corresponded to the recommended analysis area for hematologists: the monolayer.

2.4 Segmentation

To implement the single-cell-based approach, individual white blood cells were extracted from each PB patch. This segmentation was performed using a neural cellular automata (NCA) approach. Inspired by classical cellular automata and extended through the integration of neural networks, the model was designed to learn patterns directly from images. The NCA model was based on the premise that each cell (pixel) evolved iteratively based on the states of its neighboring cells.

Each NCA cell was represented by a 6-channel feature vector. The first three corresponded to the input image, and the remaining three were latent channels. At the beginning, the initial state was defined as the image, and zeros were assigned to the additional channels. Regarding the local perception of the neighborhood, a depthwise convolution with reflecting padding was employed. The output of this local perception was concatenated with the current cell state, forming an extended representation that incorporated the cell's current state and its environment. This representation underwent a transformation through a two-layer fully connected neural network, resulting in a vector that signified the proposed change for the cell.

To introduce stochasticity and prevent the update of cells in a synchronized and rigid manner, controlled noise was incorporated. Each cell was assigned a probability (fire rate, e.g., 0.5) of updating its state at each stage. If a cell was not activated, its state remained unchanged. The model evolved its state over a predetermined number of iterations (e.g., 32). During the process, the original content of the image was preserved, but the underlying latent representations were updated.

The resulting output comprised a continuous activation map, with values ranging from 0 to 1, which was then processed to specify the locations of the cells. The postprocessor identified local peaks, which generally corresponded to the centers of the cells. To avoid false positives, minimum activation filtering was applied, and a minimum distance between peaks was enforced to prevent the detection of multiple peaks as distinct cells.

In the final stage, detected cells were segmented into 40×40 patches, preserving surrounding red blood cells to retain local spatial context.

2.5 Feature Extraction

To align with the pretrained model specifications, preprocessing was performed on the data. First, it was verified that the images were in RGB format; if patches were in CMYK or RGBA, they were converted accordingly. Subsequently, they were resized to a fixed size (e.g., 224×224 pixels). Finally, the images, stored as PIL arrays, underwent a series of transformations, including normalization and tensorization.

Following the preprocessing stage, two pretrained vision encoders were employed: UNI2 [19] and DinoBloom-G [20]. Each model processed the images independently, generating a numerical embedding of 1,536 dimensions that encoded the visual representation of each WSI.

2.6 Feature Aggregation

The dataset used in this study was partitioned using a case-wise k-fold cross-validation strategy. To prevent data leakage and overfitting, all images belonging to the same case were assigned to the same subset. The data was split into training, validation, and test sets. The training set was shuffled to introduce slight variability, while the validation and test sets remained unshuffled to ensure consistent evaluation.

The model was trained under a weakly supervised learning setting, as only slide-level labels were available. A MIL architecture was employed, using two variants: attention-based deep MIL (AB-MIL) [21] and transformer-based correlated MIL (TransMIL) [22]. These models aggregated features from individual instances to construct a comprehensive representation of each WSI, enabling classification into CHIP-positive or CHIP-negative categories.

The model was trained for 50 epochs using a supervised learning approach. Each iteration consisted of a forward pass to generate predictions, loss computation using cross-entropy loss by comparing predictions with ground truth labels, and backpropagation using the AdamW optimizer, which combines weight decay with adaptive learning rate adjustments.

To facilitate convergence, a learning rate schedule with a warm-up of 10 epochs was employed, linearly increasing the learning rate from 0 to 0.0001. Afterward, a cosine scheduler reduced the learning rate following a cosine decay function, lowering it to a minimum of 1e-6. In parallel, the weight decay, initially set to 0.04, increased gradually up to 0.4. Moreover, a momentum of 0.9 was incorporated to accelerate convergence during optimization.

During training, metrics such as AUC, binary F1-score, and accuracy were computed, along with a confusion matrix. Additionally, checkpoints were saved at the end of each epoch, retaining both the best model (based on validation performance) and the latest checkpoint.

For model evaluation, the trained checkpoint was loaded using the same architecture as during training. During inference, predictions for each

sample in the test set were generated, and the same metrics used during validation were computed.

2.7 Code Availability

All the code used for the analysis is available online at: https://github.com/laiaabm/final-degree-project. The repository also includes pretrained model weights for QC and segmentation. The pretrained networks for feature extraction can be downloaded from UNI at https://github.com/mahmoodlab/UNI and from DinoBloom at https://github.com/marrlab/DinoBloom.

3 Results and Discussion

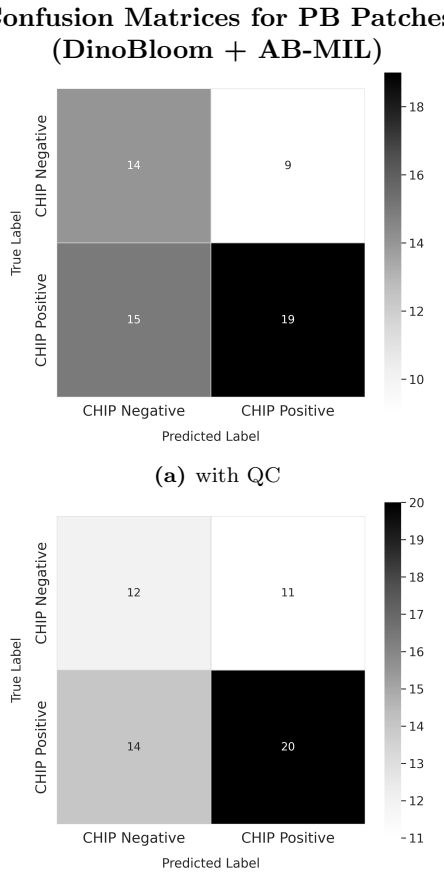
Several distinct analyses were conducted to compare the performance of each approach. The models were implemented using two types of encoders (UNI [19] and DinoBloom [20]), combined with two aggregators (AB-MIL [21] and TransMIL [22]). The performance of the classification models was evaluated using AUC, F1-score, and accuracy (ACC).

3.1 PB Patches

The performance of the classification model on PB patches with QC is summarized in Table 1. To assess whether the QC step was discarding relevant features, the same analysis was repeated on PB patches without any QC, excluding only those patches where more than 50% of the area was blank. The results of this second analysis are also summarized in Table 1.

The results of the PB model with QC confirm that the highest performance was achieved using the DinoBloom encoder combined with the AB-MIL aggregation method, reaching an AUC of 0.59, an F1-score of 0.60, and an ACC of 0.58. This configuration demonstrates a stronger discriminative capacity for identifying CHIP-positive samples.

In comparison, the results obtained on PB patches without QC display an overall decrease in performance across all models. For instance, the same configuration (DinoBloom and AB-MIL) reached an AUC of 0.52, an F1-score of 0.59, and an ACC of 0.56. These results suggest that the QC step contributes to improved model performance by removing irrelevant patches, such as artifacts or noisy patches. Nevertheless, there remains the possibility that some subtle features are discarded, affecting the model's sensitivity.



(a) with QC

(b) without QC

Figure 2: Comparison of confusion matrices for PB patch classification with and without QC using DinoBloom encoder and AB-MIL aggregator.

Figure 2 presents the confusion matrices obtained for DinoBloom and AB-MIL, the most effective configuration. Figure 2a shows the results on PB patches with QC, which result in slightly fewer false positives. Conversely, Figure 2b displays the confusion matrix for patches without QC. Although this second case presents fewer false negatives, the matrix with QC exhibits a more pronounced diagonal, indicating a clearer separation between classes and better class discrimination. Overall, the obtained results confirm the substantial benefit of the QC step in improving the model's performance. Confusion matrices for other combinations can be found in the supple-

Table 1: Summary of the metrics computed from the analysis of PB patches with and without QC in the test datasets. Performance is reported in terms of AUC, F1-score, and ACC.

Encoder	Aggregator	With QC			Without QC		
		AUC	F1-score	ACC	AUC	F1-score	ACC
UNI	AB-MIL	0.41 ±0.24	0.43 ±0.14	0.40 ±0.19	0.37 ±0.31	0.36 ±0.13	0.36 ±0.19
UNI	transMIL	0.40 ±0.22	0.50 ±0.19	0.42 ±0.23	0.44 ±0.18	0.53 ±0.13	0.48 ±0.16
DinoBloom	AB-MIL	0.59 ±0.28	0.60 ±0.17	0.58 ±0.15	0.52 ±0.30	0.59 ±0.25	0.56 ±0.25
DinoBloom	transMIL	0.43 ±0.16	0.52 ±0.15	0.46 ±0.15	0.46 ±0.14	0.47 ±0.13	0.41 ±0.15

mentary material, at Figures S1-S4.

In both approaches, models employing the DinoBloom encoder outperform those using UNI, indicating that DinoBloom, pretrained on PB images, may be better suited for extracting relevant morphological features for CHIP classification. In contrast, UNI was pretrained on histology images. This suggests that the model may not adapt effectively to PB smear data owing to differences in tissue architecture and cellular composition.

Focusing on the aggregators, AB-MIL tends to achieve better results than TransMIL. This higher performance reveals that attention-based pooling mechanisms are more effective at aggregating and identifying relevant morphological features for cell patches. This capacity is especially advantageous for PB smears, where the spatial disposition of the cells is less structured and more variable. In contrast, TransMIL, which is based on modeling spatial relationships, can be limited with this type of data, where such relations are not consistent or informative.

Additionally, some folds revealed signs of overfitting and underfitting. For example, Table S1 shows the results for the UNI encoder combined with the AB-MIL aggregator. For both folds 2 and 4, a considerably high performance on the validation set can be observed; however, the performance significantly decreases on the test set, suggesting overfitting. Conversely, fold 3 exhibits higher performance on the test set compared to validation, indicating underfitting. Other tables showing values for all folds can be found in the supplementary material, Tables S2-S8.

These observations indicate that the model is not consistently capturing true morphological features, thereby reducing the predictor's reliability. The strong variability observed across folds underscores the necessity for enhanced tuning strategies, such as regularization, data augmentation, or architectural modifications.

Despite these results, the highest metric values do not exceed an AUC of 0.59, suggesting that classification based on PB patches remains a significant challenge and that the models still have considerable scope for improvement.

In spite of the efforts, it is important to recognize the limitations of CHIP classification uniquely through the use of PB patches based on morphological characteristics. CHIP-related cellular features may be extremely subtle, which reduces the detectable signal for the model. Moreover, the high cellular heterogeneity and variability in PB smear preparation may hinder the identification of consistent patterns.

3.2 PB Segmentation

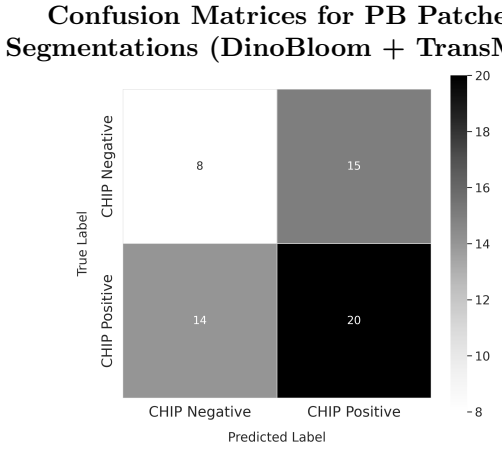
The white blood cell segmentation of PB smears was evaluated using two image dimensions: 40×40 and 224×224 resized pixels. Resizing was performed to conform to the input dimensions required by the majority of models.

Contrary to our initial hypothesis, increasing the patch size from 40×40 to 224×224 did not result in improved model performance. Table 2 presents the testing results for this analysis. Some models exhibited slight benefits from resizing, such as DinoBloom combined with TransMIL, which achieved a higher AUC (0.51 ± 0.31), F1-score (0.58 ± 0.17), and ACC (0.53 ± 0.13). Conversely, others, including UNI with TransMIL, demonstrated decreased performance when the image size was increased. Tables showing the values for each fold are presented in supplementary material, Tables S9-S16.

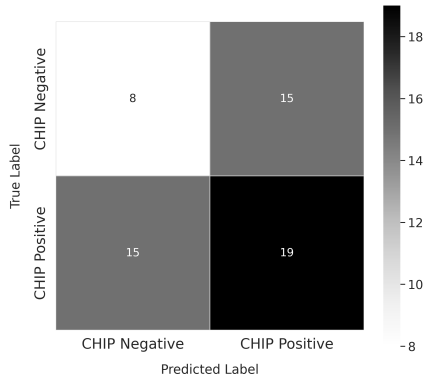
These results suggest that increasing the size of the images doesn't necessarily improve the model's performance. A plausible explanation is that resizing to 224×224 might introduce artifacts or dilute relevant morphological characteristics, specifically when dealing with small nuclei. Moreover,

Table 2: Summary of performance metrics for white blood cell segmentation on PB patches with original size 40×40 and resized to 224×224, evaluated on the test dataset. Metrics include AUC, F1-score, and ACC.

Encoder	Aggregator	40×40			224×224		
		AUC	F1-score	ACC	AUC	F1-score	ACC
UNI	AB-MIL	0.42 ±0.28	0.55 ±0.08	0.50 ±0.12	0.47 ±0.17	0.55 ±0.08	0.51 ±0.11
UNI	transMIL	0.40 ±0.18	0.60 ±0.14	0.52 ±0.28	0.41 ±0.24	0.43 ±0.14	0.40 ±0.19
DinoBloom	AB-MIL	0.40 ±0.18	0.50 ±0.21	0.47 ±0.20	0.47 ±0.16	0.45 ±0.20	0.48 ±0.26
DinoBloom	transMIL	0.39 ±0.16	0.57 ±0.14	0.49 ±0.16	0.51 ±0.31	0.58 ±0.17	0.53 ±0.13



(a) 40 × 40



(b) 224 × 224

Figure 3: Comparison of confusion matrices for PB patch segmentations with and without resizing using DinoBloom encoder and TransMIL aggregator.

working with larger images increases the complexity of the model and may exacerbate overfitting problems, especially when working with a limited dataset.

The comparison of the confusion matrices displayed in Figure 3 indicates that both resized and non-resized images produce similar results. Both models exhibit a considerable number of false positives and false negatives, suggesting difficulty in accurately identifying CHIP-negative samples. Supplementary material Figures S5-S8 show the confusion matrices for the remaining configurations.

The DinoBloom vision encoder outperformed UNI. However, their overall performances were quite similar. In contrast, when comparing aggregators, TransMIL consistently achieved better results than AB-MIL. Despite PB smears exhibiting unstructured cell distributions, the analyzed patches are focused on a single white blood cell surrounded by red blood cells. This relatively consistent organization enables TransMIL mechanisms to effectively capture spatial relationships between cells, leading to enhanced discriminative capacity relative to more general attention-based methods such as AB-MIL.

In comparison to the results obtained with PB patches, the white blood cell segmentation models show a more limited performance. While the model for the patches with QC reached an AUC of 0.59 (DinoBloom and AB-MIL), the results for the single-cell analysis barely reached an AUC of 0.51 (DinoBloom and TransMIL) in resized images. Furthermore, for both approaches there are difficulties in distinguishing correctly between CHIP-positive and CHIP-negative samples, but the patches analysis shows a better discriminative capacity, especially when using QC. Nonetheless, the segmentation approach doesn't even present consistent enhancements even when increasing the image size, suggesting that classification based on patches is a more effective strategy.

3.3 BM Patches

For the BM analysis, only the patches were evaluated, results are shown in Table 3. Among all the possible configurations, UNI combined with TransMIL achieved the highest performance in

Table 3: Summary of the metrics computed from the analysis of BM patches in the validation and test datasets. Performance is reported in terms of AUC, F1-score, and ACC.

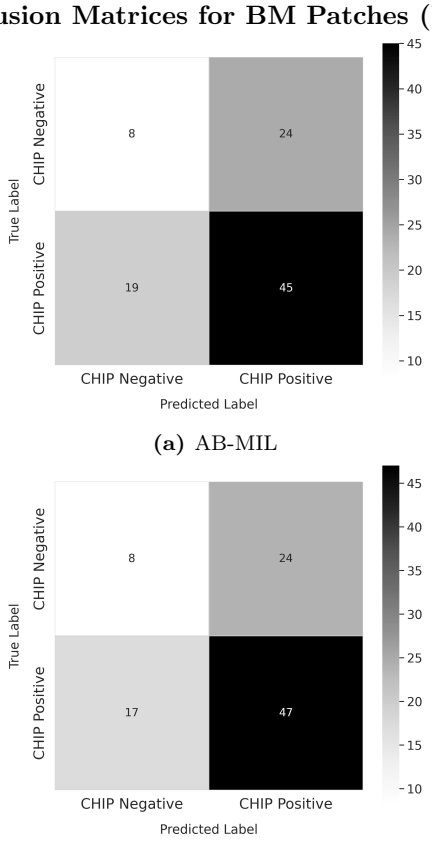
Encoder	Aggregator	Validation			Test		
		AUC	F1-score	ACC	AUC	F1-score	ACC
UNI	AB-MIL	0.38 ±0.29	0.59 ±0.35	0.51 ±0.27	0.51 ±0.34	0.69 ±0.19	0.57 ±0.24
UNI	transMIL	0.41 ±0.28	0.70 ±0.17	0.58 ±0.20	0.53 ±0.34	0.71 ±0.17	0.60 ±0.22
DinoBloom	AB-MIL	0.44 ±0.18	0.59 ±0.26	0.53 ±0.15	0.50 ±0.35	0.57 ±0.30	0.50 ±0.26
DinoBloom	transMIL	0.40 ±0.24	0.58 ±0.33	0.49 ±0.24	0.45 ±0.32	0.67 ±0.08	0.54 ±0.11

both datasets, validation and testing, obtaining an AUC of 0.53 ± 0.34, an F1-score of 0.71 ± 0.17, and an ACC of 0.60 ± 0.22 in the test set. In contrast, the DinoBloom encoder shows a lower performance and greater variability in the metrics, especially when combined with AB-MIL, showing an AUC of 0.50 ± 0.35, an F1-score of 0.57 ± 0.30, and an ACC of 0.50 ± 0.11. These results may suggest that the features extracted by DinoBloom are less informative for the BM patches, probably due to higher noise levels. Tables S17-S20 show the metrics AUC, F1-score, and ACC per fold.

The confusion matrices for the UNI encoder can be seen in Figure 4. In general, both matrices show similar patterns, with almost the same number of true positives and true negatives. TransMIL slightly outperformed AB-MIL, achieving two more true positives than the latter model. From the confusion matrices, it can be inferred that the models fail to discriminate between CHIP-positive and CHIP-negative samples, primarily due to the absence of correctly predicted CHIP-negative samples. This may be attributed to the class imbalance in the dataset, which comprises more CHIP-positive samples than CHIP-negative, causing the model to not learn features of CHIP-negative samples properly. The confusion matrices for DinoBloom are shown in Figure S9 from the supplementary material.

Furthermore, it is notable that the F1-score for BM is relatively high, especially with UNI and TransMIL (0.71 in the test set), despite low values for AUC and ACC. This discrepancy may indicate a classification imbalance or higher sensitivity of the model toward a particular class, resulting in a high F1-score but limited discriminative capacity. Examination of the confusion matrices confirms this, as the model predicts many more CHIP-positive samples than CHIP-negative ones.

Comparing the results obtained for the BM and PB patches (with QC), it is determined



(a) AB-MIL

(b) TransMIL

Figure 4: Confusion matrices for BM patches using UNI encoder and both AB-MIL and TransMIL aggregators.

that the best configuration for PB patches is DinoBloom with AB-MIL, reaching an AUC of 0.59 ± 0.28 on the test set. In contrast, for BM patches, the best performance is achieved with UNI and TransMIL (AUC 0.53 ± 0.34). In general, BM results exhibit lower values and higher variability than PB, suggesting that BM patches represent a greater challenge for CHIP classifica-

tion.

From a biological point of view, BM presents a higher complexity and a denser environment, which can hinder the identification of relevant morphological features associated with CHIP. In contrast, PB smears contain more dispersed cells, which may facilitate the extraction of relevant features by the models. Additionally, the preparation of BM samples often introduces artifacts or cellular superposition, which may add noise and reduce the quality of the patches used for model training.

Overall, obtaining an AUC close to 0.5 indicates that the model shows limited performance in differentiating CHIP-positive from CHIP-negative samples. Additionally, the high variability underscores the necessity for improved model stability, which could be gained by incorporating additional data or making architectural adjustments.

4 Conclusion

The study has evaluated the use of deep learning models for the classification of CHIP through PB and BM smears using several approaches. Even if some configurations showed relatively superior performance, overall results indicate that the task remains a significant challenge. In particular, a limited discriminative capacity and high variability across partitions were observed, suggesting that the models for PB patches, PB segmentations, and BM patches do not effectively capture the relevant morphological characteristics for CHIP diagnosis.

Although QC techniques and distinct patch sizes were incorporated for PB segmentations, the results reflect limitations in the discriminative capacity of the models, with performance close to randomness and high variability between folds.

Considering the complexity of the task and the morphological variability among samples, the limited dataset hinders the training of complex models and reduces data representativeness. Consequently, future work should focus on building a larger and more balanced dataset to allow better class characterization.

Moreover, the results indicate the necessity for further optimization of the used architectures to develop more accurate models for the task. If successful, the methodology used could substantially reduce the time and costs of diagnosis, revolutionizing CHIP detection.

In conclusion, based on current PB and BM methods, we cannot reliably discriminate between CHIP-positive and CHIP-negative samples.

5 Acknowledgements

I would like to express my appreciation to my supervisors, Dr. Umer Rao Muhammad and Dr. Carsten Marr, for their support and for giving me the opportunity to be part of the group. I am also grateful to the whole Helmholtz Munich community for the amazing experience.

References

[1] Tharani Krishnan, Joao Paulo Solar Vasconcelos, Emma Titmuss, Robert J. Vanner, David F. Schaeffer, Aly Karsan, Howard Lim, Cheryl Ho, Sharlene Gill, Stephen Yip, Stephen K. Chia, Hagen F. Kennecke, Derek J. Jonker, Eric X. Chen, Daniel J. Renouf, Chris J. O'Callaghan, and Jonathan M. Loree. Clonal hematopoiesis of indeterminate potential and its association with treatment outcomes and adverse events in patients with solid tumors. *Cancer research communications*, 5:66–73, 1 2025.

[2] Adrian Schmidt. Clonal hematopoiesis of indeterminate potential: New insights from recent studies, 2021.

[3] Siddhartha Jaiswal, Pierre Fontanillas, Jason Flannick, Alisa Manning, Peter V. Grauman, Brenton G. Mar, R. Coleman Lindsley, Craig H. Mermel, Noel Burtt, Alejandro Chavez, John M. Higgins, Vladislav Moltchanov, Frank C. Kuo, Michael J. Kluk, Brian Henderson, Leena Kinnunen, Heikki A. Koistinen, Claes Ladenvall, Gad Getz, Adolfo Correa, Benjamin F. Banahan, Stacey Gabriel, Sekar Kathiresan, Heather M. Stringham, Mark I. McCarthy, Michael Boehnke, Jaakko Tuomilehto, Christopher Haiman, Leif Groop, Gil Atzmon, James G. Wilson, Donna Neuberg, David Altshuler, and Benjamin L. Ebert. Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine*, 371, 2014.

[4] Jasmine Singh, Nancy Li, Elham Ashrafi, Le Thi Phuong Thao, David J. Curtis, Erica M. Wood, and Zoe K. McQuilten. Clonal hematopoiesis of indeterminate potential as a prognostic factor: a systematic review and meta-analysis, 7 2024.

[5] David P. Steensma. Clinical consequences of clonal hematopoiesis of indeterminate potential, 2018. Accessed: 2025-05-22.

[6] Robert P. Hasserjian, Ulrich Germing, and Luca Malcovati. Diagnosis and classification of myelodysplastic syndromes. *Blood*, 142(26):2247–2257, December 2023.

[7] A. Abayomi, A. Adesina, D. Berney, C. Carillo, R. D'Angelo, I. Diomande, S. Duale, R. Emodi, A. Eslan, A. Field, J. Flanigan, K. Fleming, M. Hale, A. Howat, Y. Ilyasu, S. Lucas, D. Milner, A. Nelson, L. Ngendehayo, and R. Zarbo. Quality pathology and laboratory diagnostic services are key to improving global health outcomes: Improving global health outcomes is not possible without accurate disease diagnosis. *American Journal of Clinical Pathology*, 143(3):325–328, 2015.

[8] Josef Riedl. Digital imaging/morphology is the next chapter in hematology. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 56(5):690–697, 2018.

[9] Chen Li, Xintong Li, Md Mamunur Rahaman, Xiaoyan Li, Hongzan Sun, Hong Zhang, Yong Zhang, Xiaoqi Li, Jian Wu, Yudong Yao, and Marcin Grzegorzek. A comprehensive review of computer-aided whole-slide image analysis: From datasets to feature extraction, segmentation, classification and detection approaches. *Artificial Intelligence Review*, 55(6):4809–4878, 2021.

[10] Liron Pantanowitz, Navid Farahani, and Anil Parwani. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International*, page 23, 6 2015.

[11] D. Kwon. How artificial intelligence is transforming pathology. *Nature*, 2025.

[12] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19839, 2023.

[13] Miao Cui and David Y. Zhang. Artificial intelligence and computational pathology, 4 2021.

[14] Philip Chikontwe, Meejeong Kim, Soo Jeong Nam, Heounjeong Go, and Sang Hyun Park. Multiple instance learning with center embeddings for histopathology classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12345 of *Lecture Notes in Computer Science*, pages 519–528. Springer, 2020.

[15] Qichen Sun, Zhengrui Guo, Rui Peng, Hao Chen, and Jinzhuo Wang. Any-to-any learning in computational pathology via triplet multimodal pretraining. *arXiv preprint arXiv:2505.12711*, 2025.

[16] Rakhmonalieva Farangis Oybek Kizi, Tagne Poupi Theodore Armand, and Hee-Cheol Kim. A review of deep learning techniques for leukemia cancer classification based on blood smear images. *Applied Biosciences*, 4(1):9, 2025.

[17] Rabia Asghar, Sanjay Kumar, Arslan Shaukat, and Paul Hynds. Classification of white blood cells (leucocytes) from blood smear imagery using machine and deep learning models: A global scoping review. *Plos one*, 19(6):e0292026, 2024.

[18] José Guilherme de Almeida, Emma Gudgin, Martin Besser, William G Dunn, Jonathan Cooper, Torsten Haferlach, George S Vassiliou, and Moritz Gerstung. Computational analysis of peripheral blood smears detects disease-associated cytomorphologies. *Nature Communications*, 14(1):4378, 2023.

[19] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

[20] Valentin Koch, Sophia J Wagner, Salome Kazeminia, Ece Sancar, Matthias Hehr, Julia A Schnabel, Tingying Peng, and Carsten Marr. Dinobloom: a foundation model for generalizable cell embeddings in hematology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 520–530. Springer, 2024.

[21] Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

[22] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

[23] Neofytos Dimitriou, Ognjen Arandjelović, and Peter D Caie. Deep learning for whole slide image analysis: an overview. *Frontiers in Medicine*, 6:264, 2019.

[24] Ozan Ciga, Tony Xu, Sharon Nofech-Mozes, Shawna Noy, Fang-I Lu, and Anne L Martel. Overcoming the limitations of patch-based learning to detect cancer in whole slide images. *Scientific Reports*, 11(1):8894, 2021.

[25] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[26] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

[27] Qin Ma and Dong Xu. Deep learning shapes single-cell data analysis. *Nature Reviews Molecular Cell Biology*, 23(5):303–304, 2022.

[28] Michael Deutges, Ario Sadafi, Nassir Navab, and Carsten Marr. Neural cellular automata for lightweight, robust and explainable classification of white blood cell images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 693–702. Springer, 2024.

[29] John Kalkhof, Camila González, and Anirban Mukhopadhyay. Med-nca: Robust and lightweight segmentation with neural cellular automata. In *International Conference on Information Processing in Medical Imaging*, pages 705–716. Springer, 2023.

[30] John Kalkhof and Anirban Mukhopadhyay. M3d-nca: Robust 3d segmentation with built-in quality control. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 169–178. Springer, 2023.

[31] Avni Mittal, John Kalkhof, Anirban Mukhopadhyay, and Arnav Bhavsar. Medsegdiffnca: Diffusion models with neural cellular automata for skin lesion segmentation. *arXiv preprint arXiv:2501.02447*, 2025.

[32] Taralynn Mack, Caitlyn Vlasschaert, Kelly von Beck, Alexander J Silver, J Brett Heimlich, Hannah Poisner, Henry R Condon, Jessica Ulloa, Andrew L Sochacki, Travis P Spaulding, et al. Cost-effective and scalable clonal hematopoiesis assay provides insight into clonal dynamics. *The Journal of Molecular Diagnostics*, 26(7):563–573, 2024.