Ki-67-Derived Supervision for Proliferation Classification from H&E Histopathology Images

Dominika Petríková, Patrik Ištvanko, Ivan Cimrák
University of Žilina
Žilina, Slovakia
dominika.petrikova, istvanko, ivan.cimrak@fri.uniza.sk

***Abstract*—Predicting tumor proliferative activity directly from routine hematoxylin and eosin (H&E) histopathology images offers a promising alternative to immunohistochemical (IHC) staining. However, the formulation and quality of automated supervisory signals derived from Ki-67 whole-slide images heavily dictate model performance. In this study, we perform a comparative evaluation of a calibrated Ki-67 annotation procedure against an original automated annotation baseline for three-class proliferation classification (low, intermediate, high) on H&E images. Using 66 registered H&E/Ki-67 whole-slide image pairs, we train and evaluate two deep learning architectures (ResNet18 and Data-efficient Image Transformer, DeiT). To rigorously assess how annotation choice impacts predictive capability, models are evaluated under both a mixed patch-level protocol and a strict per-slide protocol across three independent splits. Our results demonstrate that the calibrated annotations substantially improve patch-level macro F1-scores (reaching up to 67.36% for DeiT) while shifting class distributions. However, under strict per-slide evaluation on unseen source slides, performance decreases across all models (macro F1 of 31–34%), indicating that while calibrated annotations provide superior local supervisory signals, cross-slide generalization remains bottlenecked by slide-level tissue variability. This work demonstrates the impact of Ki-67 annotation calibration and establishes the necessity of slide-level evaluation when validating supervisory annotation strategies in computational pathology.**

I. Introduction

The Ki-67 proliferation index is an important biomarker used to characterize proliferative activity in tumor tissue. Its assessment typically relies on immunohistochemically stained tissue sections, in which Ki-67-positive nuclei are identified and their proportion within the evaluated cell population is determined. Although digital image analysis can improve the reproducibility and efficiency of this process, Ki-67 assessment still requires an additional immunohistochemical staining procedure and a corresponding tissue section [1]–[3].

Hematoxylin and eosin (H&E) staining, in contrast, is routinely used in histopathological examination and provides detailed information about tissue morphology and cellular organization. The increasing availability of digitized H&E slides, together with advances in deep learning, has motivated the development of computational methods for predicting molecular and immunohistochemical characteristics directly from tissue morphology [4], [5]. In the context of Ki-67, previous studies have demonstrated that proliferation-related information can be inferred from H&E images using labels obtained from corresponding immunohistochemically stained tissue [4], [5].

A major challenge in developing such models is the acquisition of sufficiently large and reliable labeled datasets. Manual annotation of histopathological images is labor-intensive and requires expert knowledge, while the availability of sufficiently large labeled datasets remains an important limitation for deep-learning-based biomarker prediction [5]. Registered H&E and Ki-67 image pairs provide an alternative source of supervision, since proliferation information obtained from the Ki-67 image can be transferred to the spatially corresponding H&E region [4], [5]. The quality of these automatically derived labels, however, depends on the reliability of the underlying Ki-67 quantification procedure.

Another important consideration is the generalization of deep learning models beyond the images represented during training. Histopathological image data may vary in tissue appearance, staining characteristics, image acquisition, and other dataset-specific properties, potentially affecting model performance on independent data [6]. Moreover, inappropriate separation of histopathological image data can introduce information leakage when related images or patches are distributed across training and evaluation subsets, resulting in an unrepresentative estimate of model generalization [7]. Consequently, performance obtained using random patch-level data splitting should be distinguished from performance obtained on patches originating from previously unseen source slides.

The primary scientific contribution of this study is a comparative evaluation of the calibrated Ki-67 annotation procedure designed to generate improved supervisory signals for H&E-based proliferation prediction. We systematically compare this calibrated annotation approach against an original uncalibrated baseline across 66 registered H&E/Ki-67 whole-slide image pairs using both a convolutional neural network (ResNet18) and a vision transformer (DeiT).

To thoroughly analyze how annotation quality influences model learning, we test both annotation strategies under two distinct evaluation protocols: a mixed patch-level setting and a strict slide-level setting. This dual-evaluation approach enables us to assess whether performance gains from annotation calibration hold under optimistic local conditions as well as across independent clinical slides. Our findings show that while calibrated annotations enhance supervisory quality and patch-level macro F1-scores, achieving robust cross-slide generalization

requires addressing inter-slide tissue and staining variations beyond label refinement alone.

II. Related Work

Deep learning has increasingly been investigated as a means of extracting molecular and immunohistochemical information directly from routinely acquired H&E-stained histopathological images. In the context of Ki-67, Liu et al. proposed a convolutional neural network for predicting Ki-67-positive and Ki-67-negative cells directly from H&E images. Training samples were obtained using corresponding immunohistochemically stained images as a source of Ki-67 expression information. Their approach achieved an average classification accuracy of 93.71% for positive, negative, and background cell images and demonstrated a correlation between predicted and immunohistochemistry-derived Ki-67 quantification [4]. This study demonstrated that morphological information contained in H&E images can be used to infer proliferation-related information conventionally obtained from IHC staining.

More recently, Akbarnejad et al. constructed a large dataset of registered H&E and corresponding IHC images for Ki-67, estrogen receptor, progesterone receptor, and HER2. Image analysis of the registered IHC images was used to derive labels for the corresponding H&E samples, and a Vision Transformer-based pipeline was subsequently trained to predict these biomarkers from H&E morphology [5]. Their work further demonstrates the potential of registered H&E–IHC image pairs for generating supervision without requiring manual annotation of every H&E image.

Recent work has also moved from categorical prediction toward direct estimation of the Ki-67 proliferation index. Transformer-based regression has been investigated for estimating Ki-67 expression from H&E whole-slide images using automatically generated labels from corresponding IHC images [8]. Such approaches further highlight the potential of transferring information from IHC to H&E, while also emphasizing the importance of reliable reference label generation and evaluation across heterogeneous image data.

In parallel with developments in Ki-67 prediction, Vision Transformers have become increasingly relevant to computational histopathology. Unlike conventional CNNs, which construct image representations primarily through local convolutional operations, transformer architectures use self-attention to model relationships between image regions. Baroni et al. investigated Vision Transformers for breast cancer histopathology classification and demonstrated that pretraining, normalization, and image preprocessing can substantially affect model performance. Their evaluation across multiple histopathological datasets also showed that performance obtained on development data does not necessarily transfer unchanged to independent image collections [6]. Data-efficient transformer architectures such as DeiT therefore provide an interesting alternative to established CNN models for learning from domain-specific histopathological datasets.

Despite these advances, two aspects remain particularly relevant for H&E-based proliferation prediction. First, the performance of supervised models depends on the availability and quality of the proliferation information transferred from the corresponding IHC images. Automatically generated labels enable larger training datasets, but inaccuracies introduced during IHC quantification can affect the supervision provided to the downstream H&E model [5]. Second, evaluation based on randomly separated image patches may not adequately represent generalization to previously unseen source slides. In digital pathology, related patches or slides distributed across training and evaluation subsets can introduce information leakage and result in an unrepresentative estimate of model generalization, separation at the patient or source level is therefore important when constructing independent evaluation sets [7].

The present study addresses these aspects by comparing proliferation labels derived using two independently developed Ki-67 annotation-generation procedures and evaluating their use for H&E-based three-class proliferation classification. A convolutional ResNet architecture is compared with the transformer-based DeiT architecture under a common experimental protocol. In addition to conventional patch-level evaluation, slide-level splitting is considered to examine model generalization to source slides that are entirely excluded from model development. The study therefore investigates both the influence of the Ki-67 annotation-generation procedure and the effect of the evaluation protocol on downstream classification performance.

III. Methods

The proposed workflow investigates whether tumor proliferative activity, conventionally assessed using Ki-67 immunohistochemical (IHC) staining, can be predicted directly from hematoxylin and eosin (H&E)-stained histopathological images. Instead of manually assigning proliferation labels to H&E image patches, the corresponding Ki-67-stained tissue is used to derive the supervisory information automatically.

An overview of the complete workflow is shown in Fig. 1. Paired H&E and Ki-67 images are first used to obtain proliferation information from the Ki-67 staining. The resulting proliferation index is converted into a three-class label and transferred to the spatially corresponding H&E patch. The deep learning models are subsequently trained and evaluated using only the H&E images as input.

A. Dataset and Label Generation

The experiments were conducted using histopathological images from the publicly available dataset of registered H&E and Ki-67 image pairs introduced in [9]. The dataset contains corresponding H&E- and Ki-67-stained tissue sections, enabling proliferation information obtained from Ki-67 staining to be spatially transferred to the associated H&E regions. For the present study, 66 registered H&E/Ki-67 WSI pairs were used, with approximately 2,000–3,000 corresponding image patches extracted from each pair.

Two independently generated variants of Ki-67-derived annotations were considered. The original annotations were

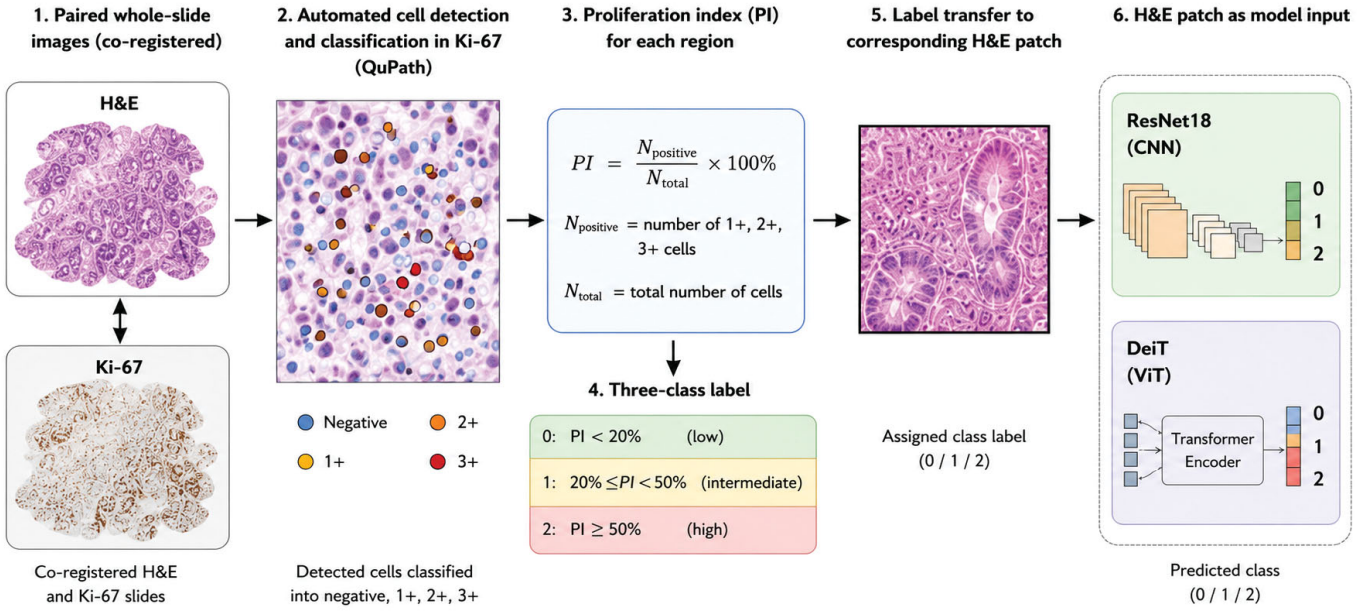


Fig. 1. Overview of the proposed workflow. Proliferation information is obtained from the Ki-67 image using automated cell detection and converted into three classes representing low, intermediate, and high proliferative activity. The resulting label is assigned to the spatially corresponding H&E patch, which is then used as the input to the ResNet18 and DeiT classification models.

obtained using the automated Ki-67 quantification pipeline presented in [10]. In this approach, foreground tissue was first separated from the background using fuzzy C-means clustering. K-means clustering was subsequently applied to distinguish DAB-stained Ki-67-positive cells from hematoxylin-stained negative cells. Morphological processing was used to refine the resulting masks, with the sequence and parameters of the morphological operations optimized using a genetic algorithm. Finally, a distance-transform-based procedure was applied to separate overlapping nuclei and estimate the numbers of positive and negative cells.

The calibrated annotations were generated using an alternative QuPath-based procedure. Cell detection and Ki-67 intensity classification were performed using the Positive Cell Detection algorithm in QuPath. Rather than relying on manually selected parameters, parameters controlling nucleus detection and the intensity thresholds separating negative, 1+, 2+, and 3+ cells were systematically calibrated using an expert-annotated reference subset. The resulting fixed parameter configuration was subsequently applied to the complete set of Ki-67 images without further manual adjustment. Cells classified as 1+, 2+, or 3+ were considered Ki-67 positive.

Despite the different annotation procedures, both variants ultimately provided the number of Ki-67-positive cells and the total number of detected cells for each image region. The proliferation index was therefore calculated consistently for both annotation variants as

$$PI = \frac{N_{\text{positive}}}{N_{\text{total}}} \times 100, \tag{1}$$

where N_{positive} denotes the number of Ki-67-positive cells and N_{total} denotes the total number of detected cells.

To establish discrete supervision targets for machine learning on H&E and Ki-67 stained slides, we categorized Ki-67 proliferation index values into three distinct classes: low ($< 20\%$), intermediate (20–50%), and high ($> 50\%$). Testicular seminoma is the most common testicular germ cell tumour and the most common malignant tumour among young men [11]. Prognosis depends on clinical stage at the time of diagnosis, tumour size, rete testis invasion, and vascular invasion. The proliferation index in seminoma tends to exceed 50% [12], but lower values below 20% are also observed. Although a high proliferation index in seminoma does not directly correlate with clinical stage or distant metastases [13], a significant inverse association with rete testis invasion has been observed when Ki-67 expression exceeds 50% of cells [14]. Based on these clinical features and observed expression ranges, the three thresholds ($< 20\%$, 20–50%, and $> 50\%$) were defined to capture distinct proliferative profiles within testicular seminoma.

The resulting proliferation index was converted into three target classes representing proliferative activity:

$$y = \begin{cases} 0, & PI < 20\%, \\ 1, & 20\% \le PI < 50\%, \\ 2, & PI \ge 50\%. \end{cases} \tag{2}$$

The derived class label was assigned to the spatially corresponding H&E image patch. Consequently, Ki-67 images were used exclusively for generating supervisory labels, whereas the

TABLE I. DISTRIBUTION OF PROLIFERATION CLASSES DERIVED USING THE ORIGINAL AND CALIBRATED ANNOTATIONS

Class	Original	Calibrated
Low ($< 20\%$)	13,956 (10.93%)	78,750 (61.66%)
Intermediate ($20\% \leq PI < 50\%$)	98,705 (77.28%)	40,596 (31.78%)
High ($\geq 50\%$)	15,070 (11.80%)	8,385 (6.56%)
Total	127,731	127,731

neural networks received only H&E images as input. The same proliferation-index thresholds were applied to both annotation variants, enabling the influence of the annotation-generation procedure on subsequent H&E classification to be evaluated independently of the class definition.

The two annotation procedures resulted in substantially different distributions of the derived proliferation classes, as summarized in Table I. The original annotations were strongly dominated by the intermediate-proliferation class, whereas calibration shifted the majority of patches toward the low-proliferation class. Both annotation variants remained class-imbalanced, motivating the use of weighted sampling during training and macro-averaged metrics for evaluation.

B. Classification Models and Training

Two deep learning architectures were evaluated for the three-class proliferation classification task. ResNet18 was used as a convolutional neural network baseline, while the Data-efficient Image Transformer (DeiT) represented the transformer-based approach. Both models were initialized using ImageNet-pretrained weights, and their classification heads were adapted to produce three output classes.

ResNet18 represents a convolutional architecture based on residual learning [15], whereas DeiT is a transformer-based architecture designed for data-efficient image classification [16]. Their comparison was used to investigate whether the different representation mechanisms influence the prediction of proliferation-related information from H&E morphology.

All H&E image patches were resized to 224×224 pixels and normalized using the ImageNet channel-wise mean and standard deviation. During training, random horizontal and vertical flipping was applied as data augmentation. Validation and test images were resized and normalized without augmentation.

Class imbalance in the training subsets was addressed using a weighted random sampler. The sampling weight of each patch was determined by the inverse frequency of its target class, and sampling with replacement was applied during training. This increased the probability of selecting samples from less represented classes without modifying the validation or test distributions.

Both networks were fine-tuned using cross-entropy loss for 100 epochs. ResNet18 was optimized using stochastic gradient descent (SGD) with a learning rate of 10^{-3} and momentum of 0.9, whereas DeiT was optimized using AdamW with a learning rate of 10^{-6}. Batch sizes of 64 and 32 were used for ResNet18 and DeiT, respectively. A dropout probability of 0.2 was applied in the modified classification heads. The model state achieving the highest macro-averaged validation F1-score during training was retained and subsequently evaluated on the corresponding test subset. The training configuration is summarized in Table II.

TABLE II. TRAINING CONFIGURATION OF THE EVALUATED MODELS

Parameter	ResNet18	DeiT
Input size	224×224	224×224
Pretraining	ImageNet	ImageNet
Optimizer	SGD	AdamW
Learning rate	10^{-3}	10^{-6}
Batch size	64	32
Epochs	100	100
Momentum	0.9	–
Dropout	0.2	0.2
Loss function	Cross-entropy	Cross-entropy
Class balancing	Weighted sampling	Weighted sampling

The two architectures were trained independently using both the original and calibrated annotation variants under the data-partitioning protocols described in the following section.

C. Experimental Protocol and Evaluation

To evaluate the impact of the original and calibrated annotation procedures across different data distributions and examine how the separation of source slides influences measured classification performance, two evaluation protocols were implemented:

1) **Mixed Patch-Level Split (Optimistic Setting):** Image patches originating from all available WSI pairs were pooled and randomly shuffled prior to partitioning, assigning 81% of the patches to training, 9% to validation, and 10% to testing. This setting evaluates supervisory label performance under standard patch mixing, serving as a baseline for local patch-level discrimination. However, because partitioning was performed at the patch level, patches originating from the same source slide could appear across different subsets. This intra-slide data leakage creates an optimistic evaluation scenario. The exact same train, validation, and test patch assignments were used for both annotation variants and model architectures to ensure direct comparability.

2) **Per-Slide Split (Slide-Level Generalization):** To test whether calibrated annotations improve prediction on completely unseen clinical specimens, partitioning was enforced strictly at the source-slide level before individual image patches were loaded. Complete source slides were assigned exclusively to either the training ($\approx 80\%$), validation ($\approx 10\%$), or test ($\approx 10\%$) subset, preventing patches from the same slide from occurring in multiple subsets. To account for slide selection variability, this slide-level partitioning was evaluated across three independent random splits. Because the total patch count varied between individual source slides, the resulting proportions of image patches in the subsets varied slightly across slide partitions. Only source slides for which both original and calibrated annotations were available were included in this evaluation.

To account for the potential dependence of slide-level performance on the particular composition of the training, vali-

dation, and test subsets, the per-slide evaluation was repeated using three independently generated slide partitions. The same three partitions were used for ResNet18 and DeiT and for both annotation variants, ensuring that all model and annotation comparisons were performed on identical source-slide assignments. Only the random seed controlling the assignment of source slides was varied between the three partitions, while the training seed, model-specific hyperparameters, preprocessing, class-balancing procedure, and optimization settings were kept unchanged.

Performance on the test subsets was assessed using accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1-score. For the per-slide evaluation, the metrics were summarized across the three partitions using the mean and standard deviation. Macro averaging assigns equal importance to all three proliferation categories irrespective of their frequency and therefore complements overall accuracy in the presence of class imbalance. Confusion matrices were additionally analyzed to characterize class-specific prediction behavior.

IV. Experimental Results

The experiments compared ResNet18 and DeiT using the original and calibrated annotation variants under mixed patch-level and per-slide evaluation. The mixed evaluation measured classification performance when patches originating from the same source slide could occur in different subsets, whereas the per-slide evaluation assessed performance on previously unseen source slides. The latter experiment was repeated across three independent slide partitions to account for variability associated with the composition of the training, validation, and test subsets.

A. Mixed Patch-Level Classification

Table III summarizes the performance obtained using the mixed patch-level split. With the original annotations, ResNet18 achieved an accuracy of 73.20% and a macro F1-score of 51.18%, whereas DeiT reached 77.00% accuracy and a macro F1-score of 56.54%. DeiT therefore exceeded ResNet18 by 5.36 percentage points in macro F1-score.

The calibrated annotations considerably improved the macro-averaged metrics of both architectures while having almost no effect on overall accuracy. For ResNet18, accuracy changed only from 73.20% to 73.22%, whereas macro F1-score increased from 51.18% to 63.62%, representing an improvement of 12.44 percentage points. Similarly, DeiT accuracy changed from 77.00% to 76.87%, while macro F1-score increased from 56.54% to 67.36%, an improvement of 10.82 percentage points. Macro precision and recall followed the same pattern.

The highest balanced classification performance in the mixed evaluation was obtained by DeiT using the calibrated annotations, with a macro F1-score of 67.36%. More importantly, the substantial improvement in macro-averaged metrics obtained with the calibrated annotations was observed for both architectures. The discrepancy between the nearly unchanged accuracy and the considerably improved macro F1-score indicates that the effect of the annotation-generation procedure was primarily reflected in the balance of classification performance across the three proliferation categories rather than in the overall proportion of correctly classified patches.

TABLE III. Performance under the mixed patch-level evaluation. All values are reported in percent.

Annotations	Model	Acc.	Prec.	Rec.	F1
Original	ResNet18	73.20	51.79	50.63	51.18
Original	DeiT	**77.00**	58.06	55.54	56.54
Calibrated	ResNet18	73.22	64.62	63.13	63.62
Calibrated	DeiT	76.87	**68.67**	**66.70**	**67.36**

B. Generalization to Unseen Slides

A substantially different performance pattern was observed when complete source slides were separated between the training, validation, and test subsets. Table IV presents the mean and standard deviation across the three independent slide partitions.

For both architectures and both annotation variants, slide-level evaluation resulted in considerably lower macro F1-scores than mixed patch-level evaluation. With the original annotations, ResNet18 decreased from 51.18% to $34.17 \pm 5.07\%$, while DeiT decreased from 56.54% to $33.85 \pm 3.79\%$. With the calibrated annotations, the decrease was larger: ResNet18 dropped from 63.62% to $31.49 \pm 5.61\%$, and DeiT from 67.36% to $31.19 \pm 6.38\%$.

The per-slide results also showed considerable variability in overall accuracy between the three partitions. ResNet18 accuracy ranged from 53.56% to 75.74% with the original annotations and from 33.27% to 59.72% with the calibrated annotations. DeiT showed a similar pattern, ranging from 51.57% to 74.25% and from 34.16% to 57.74%, respectively. This variability indicates that the measured accuracy was strongly influenced by the composition of the held-out source slides.

Macro F1-score was less variable than accuracy but remained consistently below the corresponding mixed-split performance. For ResNet18, macro F1-score ranged from 28.69% to 38.71% with the original annotations and from 25.26% to 36.15% with the calibrated annotations. For DeiT, the corresponding ranges were 29.88–37.43% and 25.03–37.77%. Thus, despite differences among individual slide partitions, none of the evaluated per-slide experiments approached the macro F1-scores obtained using mixed patch-level evaluation.

In contrast to the mixed evaluation, the two architectures performed very similarly under the per-slide protocol. Their mean macro F1-scores differed by only 0.32 percentage points with the original annotations and by 0.30 percentage points with the calibrated annotations. The advantage of DeiT observed under mixed patch-level evaluation therefore did not persist when performance was evaluated across unseen source slides.

TABLE . PERFORMANCE UNDER THE PER-SLIDE EVALUATION, REPORTED AS MEAN ± STANDARD DEVIATION ACROSS THREE INDEPENDENT SLIDE PARTITIONS. ALL VALUES ARE REPORTED IN PERCENT.

Annotations	Model	Accuracy	Macro Precision	Macro Recall	Macro F1
Original	ResNet18	66.21 ± 11.41	39.08 ± 5.74	35.70 ± 2.56	34.17 ± 5.07
Original	DeiT	64.19 ± 11.55	36.54 ± 3.63	35.21 ± 1.85	33.85 ± 3.79
Calibrated	ResNet18	47.84 ± 13.43	35.62 ± 1.99	35.26 ± 2.61	31.49 ± 5.61
Calibrated	DeiT	49.58 ± 13.36	36.83 ± 3.17	34.92 ± 3.19	31.19 ± 6.38

C. Influence of Annotation Procedure and Evaluation Protocol

The interaction between annotation calibration and evaluation protocol represents a notable finding of the experiments. Under mixed patch-level evaluation, calibrated annotations improved macro F1-score by 12.44 percentage points for ResNet18 and 10.82 percentage points for DeiT. In contrast, the corresponding mean per-slide macro F1-scores were similar for the two annotation variants. ResNet18 achieved $34.17 \pm 5.07\%$ with the original annotations and $31.49 \pm 5.61\%$ with the calibrated annotations, whereas DeiT achieved $33.85 \pm 3.79\%$ and $31.19 \pm 6.38\%$, respectively. Consequently, the improvement associated with annotation calibration in the mixed setting did not translate into improved classification performance on unseen slides.

Figure 2 summarizes this difference by comparing the macro F1-scores obtained under the two evaluation protocols. The mixed results are represented by the performance of the corresponding single patch-level experiment, whereas the per-slide results are shown as the mean and standard deviation across the three independent slide partitions.

The confusion matrices further revealed substantial differences in class-specific prediction behavior between individual slide partitions. In several experiments, predictions were strongly concentrated in one of the proliferation categories. This effect was particularly evident for the high-proliferation class, which was poorly recognized in several per-slide experiments. At the same time, some partitions contained a strongly dominant class, allowing relatively high overall accuracy to coexist with low macro-averaged performance.

Figure 3 presents confusion matrices from one representative slide partition for both architectures and annotation variants. While the precise distribution of errors varied between experiments, the matrices illustrate the class-specific imbalance that is not apparent from accuracy alone.

Overall, the results demonstrate a pronounced difference between performance measured using randomly mixed image patches and performance obtained on patches originating from previously unseen source slides. Both annotation calibration and the DeiT architecture improved balanced classification performance in the mixed setting. However, neither advantage was retained under repeated slide-level evaluation, where both architectures achieved similar macro F1-scores and substantial variation was observed between individual source-slide partitions.

V. DISCUSSION

The experiments revealed a substantial difference between performance measured using randomly mixed image patches and performance obtained on previously unseen source slides. Under mixed patch-level evaluation, both architectures achieved considerably higher macro F1-scores, and the use of calibrated Ki-67 annotations provided a clear improvement over the original annotations. However, neither the advantage of annotation calibration nor the superior mixed-split performance of DeiT was retained under repeated slide-level evaluation. These findings highlight the importance of considering the source-slide structure of histopathological datasets when evaluating models intended to generalize beyond the images represented during training.

A. Effect of the Annotation-Generation Procedure

Under mixed patch-level evaluation, the calibrated annotations increased the macro F1-score from 51.18% to 63.62% for ResNet18 and from 56.54% to 67.36% for DeiT. At the same time, overall accuracy remained nearly unchanged. The calibration procedure substantially changed the distribution of the derived proliferation classes, shifting the dominant category from the intermediate-proliferation class in the original annotations to the low-proliferation class in the calibrated annotations.

Because macro F1 calculates an unweighted average across classes, this substantial shift toward a different class distribution plays a direct role in driving the higher macro F1-score, independently of overall accuracy. Therefore, the higher macro F1-score obtained under calibrated annotations should not automatically be interpreted as definitive evidence of superior annotation quality or improved feature representation. Rather, it reflects how performance metrics respond when the supervisory target distribution is altered.

The improvement under mixed evaluation is consistent with the role of the Ki-67 processing pipeline in the proposed approach. Although the networks are trained exclusively on H&E images, their target labels are derived from cell detections performed on the corresponding Ki-67 images. Inconsistencies in Ki-67 cell detection propagate directly into the supervisory signal used for H&E classification. The altered mixed-split performance indicates that the choice of Ki-67 annotation procedure substantially influences downstream patch-level label assignment and class balance.

Importantly, this patch-level effect does not translate into improved generalization to independent slides. Under per-slide evaluation, the mean macro F1-scores obtained with the

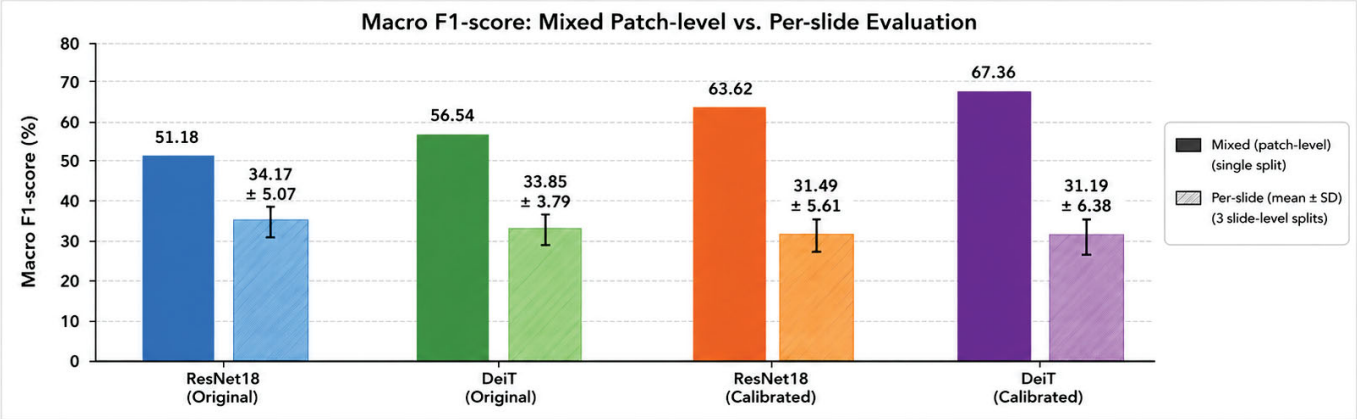


Fig. 2. Comparison of macro F1-score under mixed patch-level and per-slide evaluation for ResNet18 and DeiT using the original and calibrated annotations. Per-slide results are reported as mean ± standard deviation across three independent slide partitions.

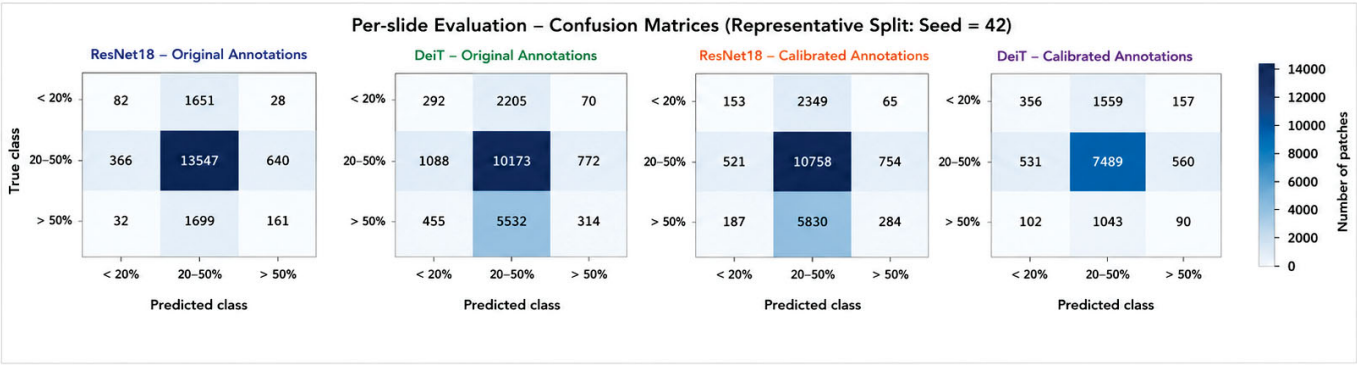


Fig. 3. Confusion matrices for a representative per-slide partition using ResNet18 and DeiT with the original and calibrated annotations. Rows represent reference proliferation classes and columns represent predicted classes.

calibrated annotations were $31.49 \pm 5.61\%$ for ResNet18 and $31.19 \pm 6.38\%$ for DeiT, compared with $34.17 \pm 5.07\%$ and $33.85 \pm 3.79\%$, respectively, using the original annotations. Considering the variability across the three partitions, these results provide no evidence of an advantage for either annotation variant under slide-level evaluation. This demonstrates that modifying the supervisory label distribution addresses only local patch-level assignments and does not by itself resolve the fundamental bottleneck of transferring H&E-derived proliferation patterns to previously unseen slides.

B. Patch-Level Performance and Slide-Level Generalization

The most pronounced finding of the study was the decrease in performance when the evaluation protocol was changed from randomly mixed patches to independent source slides. For the calibrated annotations, macro F1-score decreased from 63.62% to $31.49 \pm 5.61\%$ for ResNet18 and from 67.36% to $31.19 \pm 6.38\%$ for DeiT. A similar, although smaller, decrease was observed with the original annotations. Moreover, the reduction was reproduced across three independently generated slide partitions and with two substantially different model architectures.

A likely explanation is that patches originating from the same histological slide are not statistically independent. Neighboring or spatially related regions may share tissue morphology, staining characteristics, color distribution, acquisition properties, and other slide-specific features. When patches from the same source slide occur in both training and test subsets, a model can potentially benefit from these shared characteristics. Consequently, random patch-level splitting may provide an optimistic estimate of performance on truly independent histopathological material. Similar concerns regarding data leakage and the importance of patient- or slide-level separation have been raised in previous evaluations of deep-learning systems for whole-slide image analysis [7].

The repeated slide-level experiments further demonstrated that the measured performance depends on the particular slides included in the test subset. Accuracy showed especially large variation between partitions, whereas macro F1-score was more stable but consistently remained below the mixed-split values. The confusion matrices also showed that different partitions could result in substantially different prediction biases. These observations suggest considerable inter-slide heterogeneity in the dataset.

Several factors may contribute to such heterogeneity. Differences in tissue composition, tumor morphology, staining intensity, color characteristics, section preparation, and image acquisition may alter the appearance of H&E patches between slides. In addition, the labels used in this study are derived indirectly through registered Ki-67 sections rather than from the H&E images themselves. Imperfect local correspondence between serial tissue sections, differences in tissue structure between sections, and residual errors in cell detection may therefore introduce additional uncertainty into individual patch labels. The present experiments do not allow the contribution of these factors to be separated, and further analysis would be required to determine which sources of variation have the greatest effect on cross-slide generalization.

C. Comparison of CNN and Transformer Architectures

The comparison between ResNet18 and DeiT produced different conclusions depending on the evaluation protocol. Under mixed patch-level evaluation, DeiT consistently outperformed ResNet18. Using the calibrated annotations, DeiT achieved the highest macro F1-score of 67.36%, compared with 63.62% for ResNet18. A similar advantage was observed with the original annotations.

This result is consistent with previous work demonstrating the applicability of transformer-based architectures to histopathological image analysis [6]. The self-attention mechanism used by vision transformers provides a representation strategy different from the local convolutional operations of CNNs and may contribute to improved discrimination of proliferation-related morphological patterns when training and test data originate from overlapping source-slide distributions.

However, the architectural difference effectively disappeared under slide-level evaluation. With the original annotations, the mean macro F1-scores were $34.17 \pm 5.07\%$ for ResNet18 and $33.85 \pm 3.79\%$ for DeiT. With calibrated annotations, the corresponding values were $31.49 \pm 5.61\%$ and $31.19 \pm 6.38\%$. Thus, the stronger mixed-split performance of DeiT did not translate into better generalization to unseen slides.

This observation suggests that, for the present dataset, the main limitation is unlikely to be solved simply by replacing a convolutional architecture with a transformer. Both architectures appear to be affected by inter-slide variability when evaluated on previously unseen source slides. Consequently, improving cross-slide robustness may be more important than increasing model complexity or selecting a particular architecture.

D. Implications and Limitations

The results have implications for studies attempting to infer immunohistochemical biomarkers directly from routinely available H&E images. Previous studies have demonstrated the feasibility of predicting Ki-67-related information from H&E morphology using deep learning [4], [5], [8]. The present results indicate that proliferation-related patterns can be learned from H&E images under mixed patch-level evaluation. At the same time, they show that performance estimates can change markedly when source-slide independence is enforced.

Several limitations should be considered. First, the experiments were based on 66 registered H&E/Ki-67 WSI pairs. Although these provided a large number of image patches, the number of independent source slides is substantially smaller than the number of patch-level samples. This distinction is particularly important for slide-level evaluation and limits the diversity available for learning slide-invariant representations.

Second, the proliferation categories were derived from automatically detected Ki-67-positive cells in registered serial sections. Although the calibrated QuPath-based procedure was designed to improve the consistency of Ki-67 quantification, the resulting labels remain automatically generated rather than manually established for every image patch by expert pathologists. Registration inaccuracies and local differences between serial sections may also introduce label uncertainty. The calibrated annotation procedure also resulted in a substantially different distribution of the derived proliferation classes, with the dominant category shifting from the intermediate-proliferation class in the original annotations to the low-proliferation class in the calibrated annotations. This complicates direct attribution of the observed performance differences solely to annotation quality, since the two procedures change both individual patch labels and the overall class distribution.

Third, only ResNet18 and DeiT were evaluated. These architectures were selected to represent convolutional and transformer-based approaches rather than to provide an exhaustive comparison of available deep-learning models. Given that both architectures exhibited similar behavior under slide-level evaluation, evaluating additional architectures alone would be unlikely to address the generalization issue identified in this study.

Finally, the slide-level analysis was repeated across three partitions. This substantially reduces dependence on a single train–test division, but it does not replace validation on a completely independent external cohort. External validation across slides originating from different laboratories, scanners, or staining protocols would provide a stronger assessment of the clinical generalizability of the proposed approach.

Future work should therefore focus primarily on improving robustness across source slides. Potential directions include stain normalization or stain augmentation, domain-generalization approaches, histopathology-specific pretraining, and training on a larger and more heterogeneous collection of independent slides. The influence of slide-level class distributions and the correspondence between registered H&E and Ki-67 regions should also be investigated in greater detail. Ultimately, evaluation on an external dataset will be necessary to determine whether proliferation-related information learned from H&E morphology transfers across acquisition conditions and patient cohorts.

VI. Conclusion

This study presented a comparative evaluation of an established calibrated Ki-67 annotation procedure against an

original baseline to provide supervisory signals for three-class H&E proliferation classification. By evaluating both annotation strategies across ResNet18 and DeiT architectures, we demonstrated that calibrated annotations significantly boost model performance under mix patch-level evaluation (achieving a macro F1 of up to 67.36%).

Evaluating these models under a strict per-slide protocol further revealed that cross-slide generalization drops to 31–34% across all configurations, highlighting that while calibrated annotations refine local supervisory targets, overcoming inter-slide variability remains a key challenge for clinical deployment. Future work should build upon calibrated supervisory labels by incorporating domain generalization, stain normalization, and slide-level aggregation strategies.

These findings demonstrate that both annotation quality and evaluation methodology have a substantial influence on the apparent performance of H&E-based proliferation classification. The choice of Ki-67 annotation-generation procedure substantially affected the derived supervisory labels and the resulting balanced patch-level classification performance, but neither annotation variant ensured generalization to unseen histological slides. Moreover, performance obtained from randomly mixed patches may considerably overestimate the ability of a model to generalize to independent source material.

Future work should therefore focus primarily on cross-slide robustness rather than solely on increasing model complexity. Larger and more heterogeneous collections of independent slides, stain normalization and augmentation, domain-generalization strategies, histopathology-specific pretraining, and ultimately validation on external cohorts represent important directions toward reliable prediction of proliferation-related information directly from H&E images.

ACKNOWLEDGMENT

This publication was supported by the European Union (EU) Research Program DIGITAL EUROPE through the TEF HEALTH-Testing and Experimentation Facility for Health, Artificial Intelligence (AI) and Robotics under Contract 101100700

REFERENCES

[1] P. Bankhead, M. B. Loughrey, J. A. Fernández *et al.*, "Qupath: Open source software for digital pathology image analysis," *Scientific Reports*, vol. 7, p. 16878, 2017.

[2] B. Acs, V. Pelekanou, Y. Bai, S. Martinez-Morilla, M. Toki, S. C. Y. Leung, T. O. Nielsen, and D. L. Rimm, "Ki67 reproducibility using digital image analysis: An inter-platform and inter-operator study," *Laboratory Investigation*, vol. 99, no. 1, pp. 107–117, 2019.

[3] C. Boyaci, W. Sun, S. Robertson, B. Acs, and J. Hartman, "Independent clinical validation of the automated ki67 scoring guideline from the international ki67 in breast cancer working group," *Biomolecules*, vol. 11, no. 11, p. 1612, 2021.

[4] Y. Liu, X. Li, A. Zheng, X. Zhu, S. Liu, M. Hu, Q. Luo, H. Liao, M. Liu, Y. He, and S. Chen, "Predict ki-67 positive cells in h&e-stained images using deep learning independently from ihc-stained images," *Frontiers in Molecular Biosciences*, vol. 7, p. 183, 2020.

[5] A. Akbarnejad, N. Ray, P. J. Barnes, and G. Bigras, "Toward accurate deep learning-based prediction of ki67, er, pr, and her2 status from h&e-stained breast cancer images," *Applied Immunohistochemistry & Molecular Morphology*, vol. 33, no. 3, pp. 131–141, 2025.

[6] G. L. Baroni, L. Rasotto, K. Roitero, A. Tulisso, C. Di Loreto, and V. Della Mea, "Optimizing vision transformers for histopathology: Pretraining and normalization in breast cancer classification," *Journal of Imaging*, vol. 10, no. 5, p. 108, 2024.

[7] C. Fell, M. Mohammadi, D. Morrison, O. Arandjelovic, P. Caie, and D. Harris-Birtill, "Reproducibility of deep learning in digital pathology whole slide image analysis," *PLOS Digital Health*, vol. 1, no. 12, p. e0000145, 2022.

[8] A. K. Chaurasia, P. W. Toohey, M. T. Bennett, H. C. Harris, and A. W. Hewitt, "Automated quantification of ki-67 expression in breast cancer from h&e-stained slides using a transformer-based regression model," *Breast Cancer Research*, vol. 27, no. 1, p. 206, 2025.

[9] D. Petríková, I. Cimrák, K. Tobiášová, and L. Plank, "Dataset of registered hematoxylin–eosin and ki67 histopathological image pairs complemented by a registration algorithm," *Data (Basel)*, vol. 9, no. 8, p. 100, Aug. 2024.

[10] M. J. Kulich, D. Petrikova, and I. Cimrak, "Automated ki-67 proliferation index estimation for deep learning applications in histopathology," in *Proceedings of the IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, 2026, in press.

[11] G. Krag Jacobsen, H. Barlebo, J. Olsen, H. P. Schultz, H. Starklint, H. Søgaard, and M. Vaeth, "Testicular germ cell tumours in denmark 1976-1980. pathology of 1058 consecutive cases," *Acta Radiol. Oncol.*, vol. 23, no. 4, pp. 239–247, 1984.

[12] H. M. Rabes, N. Schmeller, A. Hartmann, U. Rattenhuber, P. Carl, and G. Staehler, "Analysis of proliferative compartments in human tumors. II. seminoma," *Cancer*, vol. 55, no. 8, pp. 1758–1769, Apr. 1985.

[13] I. Gallegos, J. P. Valdevenito, R. Miranda, and C. Fernandez, "Immunohistochemistry expression of p53, ki67, CD30, and CD117 and presence of clinical metastasis at diagnosis of testicular seminoma," *Appl. Immunohistochem. Mol. Morphol.*, vol. 19, no. 2, pp. 147–152, Mar. 2011.

[14] B. C. Lourenço, C. Guimarães-Teixeira, B. C. T. Flores, V. Miranda-Gonçalves, R. Guimarães, M. Cantante, P. Lopes, I. Braga, J. Maurício, C. Jerónimo, R. Henrique, and J. Lobo, "Ki67 and LSD1 expression in testicular germ cell tumors is not associated with patient outcome: Investigation using a digital pathology algorithm," *Life (Basel)*, vol. 12, no. 2, Feb. 2022.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[16] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 10 347–10 357.