Deep learning-based classification of early-stage mycosis fungoides and benign inflammatory dermatoses on H&E-stained whole-slide images: a retrospective, proof-of-concept study

Doeleman, T.; Brussee, S.; Hondelink, L.M.; Westerbeek, D.W.F.; Sequeira, A.M.; Valkema, P.A.; ... ; Schrader, A.M.R.

Deep Learning–Based Classification of Early-Stage Mycosis Fungoides and Benign Inflammatory Dermatoses on H&E-Stained Whole-Slide Images: A Retrospective, Proof-of-Concept Study

Thom Doeleman[1,2], Siemen Brussee[1], Liesbeth M. Hondelink[1], Daniëlle W.F. Westerbeek[1], Ana M. Sequeira[1], Pieter A. Valkema[1,3], Patty M. Jansen[1], Junling He[1], Maarten H. Vermeer[4], Koen D. Quint[4], Marijke R. van Dijk[2], Fons J. Verbeek[5], Jesper Kers[1,3] and Anne M.R. Schrader[1]

The diagnosis of early-stage mycosis fungoides (MF) is challenging owing to shared clinical and histopathological features with benign inflammatory dermatoses. Recent evidence has shown that deep learning (DL) can assist pathologists in cancer classification, but this field is largely unexplored for cutaneous lymphomas. This study evaluates DL in distinguishing early-stage MF from benign inflammatory dermatoses using a unique dataset of 924 H&E-stained whole-slide images from skin biopsies, including 233 patients with early-stage MF and 353 patients with benign inflammatory dermatoses. All patients with MF were diagnosed after clinicopathological correlation. The classification accuracy of weakly supervised DL models was benchmarked against 3 expert pathologists. The highest performance on a temporal test set was at ×200 magnification (0.50 μm per pixel resolution), with a mean area under the curve of 0.827 ± 0.044 and a mean balanced accuracy of 76.2 ± 3.9%. This nearly matched the 77.7% mean balanced accuracy of the 3 expert pathologists. Most (63.5%) attention heatmaps corresponded well with the pathologists' region of interest. Considering the difficulty of the MF versus benign inflammatory dermatoses classification task, the results of this study show promise for future applications of weakly supervised DL in diagnosing early-stage MF. Achieving clinical-grade performance will require larger multi-institutional datasets and improved methodologies, such as multimodal DL with incorporation of clinical data.

Keywords: Artificial intelligence, Cutaneous lymphoma, Digital pathology, Histology, Weakly supervised learning

Journal of Investigative Dermatology (2025) **145,** 1127–1134; doi:10.1016/j.jid.2024.07.036

INTRODUCTION

Mycosis fungoides (MF) is the most common primary cutaneous lymphoma (CL), accounting for almost 50% of all cases (Willemze et al, 2005). Clinically, patients typically present with multiple cutaneous erythematous patches and plaques in sun-protected areas. Histologically, small- to medium-sized, neoplastic T lymphocytes with cerebriform nuclei infiltrate the epidermis (epidermotropism) and papillary dermis. MF usually has an indolent protracted course, but progression to tumor stage and extracutaneous spread occur in 10–39% of patients within 10 years, and the 10-year disease-specific survival ranges from 83 to 97% (van Doorn et al, 2000). The diagnosis of early-stage MF (stages IA–IIA) can be very difficult owing to shared clinical and histopathological features with benign inflammatory dermatoses (BIDs) (ie, eczema, psoriasis, drug reactions), not uncommonly leading to misdiagnosis (Kelati et al, 2017). Differentiating early-stage MF from BIDs may even be impossible without clinicopathological correlation, reflected by a low interpathologist agreement rate (Guitart et al, 2001) and frequent diagnostic delay up to years between first symptoms and initial diagnosis (Cerroni, 2020). Therefore, easily accessible, adjunct diagnostic tools are needed for the early identification of patients with CL.

The transition into a digital workflow in pathology has offered a wide range of possibilities. This includes the use of deep learning (DL)–based tools to assist pathologists in cancer classification, outcome prediction, and grading tasks (Hondelink et al, 2022; Litjens et al., 2016; Mun et al, 2021).

[1]Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands; [2]Division of Laboratories, Pharmacy and Biomedical Genetics, Department of Pathology, University Medical Center Utrecht, Utrecht, The Netherlands; [3]Department of Pathology, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, The Netherlands; [4]Department of Dermatology, Leiden University Medical Center, Leiden, The Netherlands; and [5]Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden, The Netherlands

Correspondence: Thom Doeleman, Department of Pathology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA, Leiden 2300 RC, The Netherlands. E-mail: t.doeleman@lumc.nl

Abbreviations: AUC, area under the curve; BID, benign inflammatory dermatosis; CL, cutaneous lymphoma; CLAM, clustering-constrained attention multiple instance learning; DL, deep learning; MF, mycosis fungoides; ROI, region of interest; WSI, whole-slide image

Received 23 December 2022; revised 6 June 2024; accepted 15 July 2024; accepted manuscript published online 19 September 2024; corrected proof published online 29 October 2024

The use of DL techniques on whole-slide images (WSIs) of skin biopsies could support decision making for pathologists with expertise in other fields and allow for timely referral of patients with suspected CL to an expert center. However, to date, DL-based applications for histopathology of CL have hardly been explored, possibly explained by several specific challenges in this field: the difficulty of collecting a sufficient number of cases of these rare and heterogeneous diseases, overlapping histopathological features between CL and BIDs, and a dispersed presence of diagnostic features preventing easy labelling of regions of interest (ROIs) in a conventional supervised learning setting.

This paper presents a proof-of-concept study. We trained weakly supervised DL models using the clustering-constrained attention multiple instance learning (CLAM) method (Lu et al, 2021) to distinguish between early-stage MF and BIDs in skin biopsies using only WSIs from H&E-stained slides. The primary objective was to achieve a diagnostic performance in the binary classification task similar to expert opinion, measured by balanced accuracy and calculation of the area under the curve (AUC) by plotting the sensitivity against the specificity.

RESULTS

Patient characteristics of development dataset

The development dataset consisted of 724 WSIs, including 362 WSIs from 176 patients with MF, 99 WSIs from 65 patients with BIDs with clinical/histological suspicion of MF but without MF diagnosis after clinicopathological correlation, and 263 WSIs from 218 patients with BIDs without clinical/histological suspicion of MF (Table 1). The median number of WSIs per patient was 2 (range = 1−8) for MF and 1 (range = 1−6) for BID cases. The median age at diagnosis (60 years) was equal between the groups, but the MF cohort comprised more males than the BID cohort (66.5 vs 51.9%, $P < .01$). Regarding the anatomical distribution, the skin biopsies more often originated from the head-and-neck region (9.4 vs 1.9%; $P < .01$) and upper extremity (25.4 vs 16.0%, $P < .01$) in BID cases, whereas the majority originated from the trunk in MF cases (49.7 vs 34.8%, $P < .01$).

Patient characteristics of temporal test dataset

The temporal test dataset consisted of 200 WSIs, including 103 WSIs of 57 patients with MF, 53 WSIs from 31 patients with BIDs with clinical/histological suspicion of MF but without MF diagnosis after clinicopathological correlation, and 44 WSIs from 42 patients with BIDs without clinical/histological suspicion of MF (Table 2). The median number of WSIs per patient was 2 (range = 1−5) for MF and 1 (range = 1−6) for BID cases. In contrast to the development dataset, the median age at diagnosis was higher in the BID cohort (68 vs 60 years, $P < .05$), whereas the proportion of males to females was similar between the cohorts. Regarding the anatomical distribution, the skin biopsies more often originated from the head-and-neck region (17.5 vs 8.2%, $P < .01$) and trunk (65.0 vs 35.1%, $P < .01$) in MF cases.

Classification accuracy of expert pathologists

Three pathologists with expertise in CL independently scored all WSIs of the temporal test set (n = 200 WSI) on the basis of H&E only, without any clinical or additional pathological data (Supplementary Figure S1). When splitting the scores in either (uncertain, favor) BID (0−50%) or (uncertain, favor) MF (51−100%), the pathologists achieved a balanced accuracy of 83.2, 80.4, and 69.5% (mean = 77.7%). There was moderate interpathologist agreement (2-rating categories, 3 raters) with a multirater kappa value of 0.473.

Table 1. Patient and Biopsy Characteristics for WSI Development Dataset (n = 724 WSIs)

Characteristic	MF	BIDs	*P*-Value
Number of patients	176	283	N.A.
Number of WSIs	362	362	N.A.
Median number of WSIs per patient (range)	2 (1−8)	1 (1−6)	<.0001[1]
Sex, male, n (%)	117 (66.5)	147 (51.9)	.0022[1]
Age at time of biopsy, y, median ± SD	60 ± 16	60 ± 18	.754
Biopsy location, n (%)			
Head and neck	7 (1.9)	34 (9.4)	<.0001[1]
Trunk	180 (49.7)	126 (34.8)	<.0001[1]
Upper extremity	58 (16.0)	92 (25.4)	.0018[1]
Lower extremity	100 (27.6)	82 (22.7)	.1236
Unknown	17 (4.7)	28 (7.7)	.0910
Age of H&E slide, y, median (range)	5 (1−20)	4 (2−9)	.0210[1]
MF stage at time of biopsy, n (%)			
IA	182 (50.3)	N.A.	N.A.
IA/B	10 (2.8)	N.A.	N.A.
IB	141 (39.0)	N.A.	N.A.
IIA	0 (0)	N.A.	N.A.
IIB[2]	22 (6.1)	N.A.	N.A.
III[2]	1 (0.3)	N.A.	N.A.
Unknown	6 (1.7)	N.A.	N.A.

Abbreviations: BID, benign inflammatory dermatosis; MF, mycosis fungoides; N.A., not applicable; WSI, whole-slide image.
[1]Statistically significant *P*-values.
[2]Only biopsies included from patch or plaque lesion in patients with advanced-stage MF (stage IIB and III).

Classification accuracy of DL-based models

Of 3 evaluated feature extractors, the CLAM-based method combined with the UNI foundation model showed highest overall performance, using ×200 magnification with 256 × 256-pixel patches (16,384 μm^2) (Supplementary Figure S2). On the internal test set, a mean AUC of 0.955 ± 0.014 and a mean balanced accuracy of 90.6 ± 1.8% were achieved. On the temporal test set, a mean AUC of 0.827 ± 0.044 and a mean balanced accuracy of 76.2 ± 3.9% were achieved (Figure 1a and b).

Agreement between pathologists and DL-based models

Comparison of the pathologists' level of suspicion for MF with the prediction scores of the best performing model (discretized into bins 0−25, 25−50, 50−75, and 75−100%) revealed fair agreement, measured by linear-weighted kappa, between the best fold model and the 3 individual pathologists. The kappa values were k = 0.394, k = 0.264, and k = 0.215, respectively, for the scoring of the temporal test set. The confusion matrices are shown in Figure 1c. The performance of the DL-based model was significantly better on the

Table 2. Patient and Biopsy Characteristics for WSI Temporal Test Dataset (n = 200 WSIs)

Characteristic	MF	BIDs	*P*-Value
Number of patients	57	70	N.A.
Number of WSIs	103	97	N.A.
Median number of WSIs per patient (range)	2 (1–5)	1 (1–6)	.0143[1]
Sex, male, n (%)	34 (59.6)	46 (65.7)	.4814
Age at time of biopsy, y, median ± SD	60 ± 18	68 ± 19	.0159[1]
Biopsy location, n (%)			
Head and neck	18 (17.5)	8 (8.2)	.0030[1]
Trunk	67 (65.0)	34 (35.1)	<.0001[1]
Upper extremity	12 (11.7)	21 (21.6)	.0574
Lower extremity	18 (17.5)	26 (26.8)	.1118
Unknown	6 (5.8)	8 (8.2)	.5029
Age of H&E slide, y, median (range)	1 (0–3)	2 (1–3)	<.001[1]
MF stage at time of biopsy, n (%)			
IA	66 (64.1)	N.A.	N.A.
IA/B	0 (0)	N.A.	N.A.
IB	36 (35.0)	N.A.	N.A.
IIA	1 (1.0)	N.A.	N.A.
IIB[2]	0 (0)	N.A.	N.A.
III[2]	0 (0)	N.A.	N.A.
Unknown	0 (0)	N.A.	N.A.

Abbreviations: BID, benign inflammatory dermatosis; MF, mycosis fungoides; WSI, whole-slide image.
[1]Statistically significant *P*-values.
[2]Only biopsies included from patch or plaque lesion in patients with advanced-stage MF (stage IIB and III).

easy cohort of the temporal test set, consisting of 79 WSIs, than on the difficult cohort, consisting of 107 WSIs, with accuracy rates of 86.1 and 70.2%, respectively ($P = .01$).

Review of attention heatmaps

To evaluate the regions with the highest diagnostic value, that is, the regions that mostly contributed to the model's predictions, attention heatmaps were generated for all cases in the temporal test set (n = 200). For the best performing fold at ×200 magnification, there was good correspondence with the pathologists' ROIs in 127 of 200 (63.5%) WSI heatmaps (Figure 2), moderate correspondence with the pathologist's ROIs in 62 of 200 (31.0%) WSI heatmaps, and poor correspondence with the pathologists' ROIs in 11 of 200 (5.5%) WSI heatmaps. For unknown reasons, in cases with moderate and poor correspondence, high attention was (partly) focused on dermal collagen, the pilosebaceous unit, or other areas without obvious diagnostic relevance to the pathologist's eye. In subgroup sensitivity analyses on the temporal test set, no significant differences in performance were observed between biopsy locations or between sexes (Supplementary Figure S3).

DISCUSSION

In this single-center, proof-of-concept study, we trained DL-based models using CLAM to distinguish early-stage MF from BIDs using H&E-stained slides of skin biopsies. With 10-fold cross-validation at ×200 magnification and a 256 × 256-pixel patch size using the UNI feature extractor, our models achieved mean WSI-level AUCs of 0.955 and 0.827 and mean WSI-level balanced accuracies of 90.6 and 76.2% on the internal test set (n = 71 WSI) and temporal test set (n = 200 WSI), respectively. The final balanced accuracy of 76.2% was similar to those of 3 CL expert pathologists, who achieved a mean balanced accuracy of 77.7% on the temporal test set (n = 200).

This study demonstrates the presence of a trainable, discriminative signal based on DL for the binary MF versus BID classification task with only H&E-stained WSIs as input. Previously, researchers from Switzerland (Scheurer et al, 2020) published a DL-aided diagnostic tool for predicting MF or eczema using WSIs with corresponding segmentation maps of the epidermis and spongiosis as input. However, on a dataset of 307 WSIs from 93 patients, the classification accuracy showed an almost negligible difference compared with a dummy classifier. In addition, there was a separate study by Turkish colleagues (Karabulut et al, 2023) employing a DL model to distinguish between MF and non-MF lymphocytes, albeit without performing slide-level classification.

In this study, we managed to establish a relatively large dataset for its type, including 465 WSIs of 233 patients with early-stage MF, and demonstrated that slide-level distinction from BIDs with overlapping histopathological features was feasible without the use of manual annotations. We believe that a strength of our study is the inclusion of not only textbook examples of MF and BIDs but also ambiguous cases that were sent in for consultation but did not receive an MF diagnosis after clinicopathological correlation, increasing the resemblance of the dataset to a real-world pathology practice.

Considering the difficulty of the MF versus BID classification task, as reflected by the relatively poor accuracy and high uncertainty of the expert pathologists, the results of this study show promise for future applications of weakly supervised DL to aid in diagnosis of patients with early-stage MF. To ensure easy applicability in clinical practice, we opted to train and validate our models using only H&E-stained slides. We envision an H&E-based DL model as a screening tool for pathologists without CL expertise, to optimize the referral process of suspected patients with MF to a center of expertise, ultimately diminishing the diagnostic delay. In addition, such a model could reduce the need for (excessive) immunohistochemistry panels on skin biopsies, saving time, resources, and tissue.

In the field of digital pathology, it is crucial to focus on the explainability of artificial intelligence–based models (Border and Sarder, 2021), which is facilitated by the generation of attention heatmaps to visualize the relative contribution of tissue regions to the model's predictions. In our study, the majority (63.5 vs 31.0%) of the attention heatmaps of the best fold test set corresponded well to moderately with the pathologist's ROIs, with high attention localized at the dermal–epidermal junction and superficial dermis containing the lymphoid infiltrates. With these regions being the ROIs in MF and MF-mimicking BIDs, this suggests that the model's predictions were based on relevant tissue regions. In addition, we demonstrated that the performance of the DL-

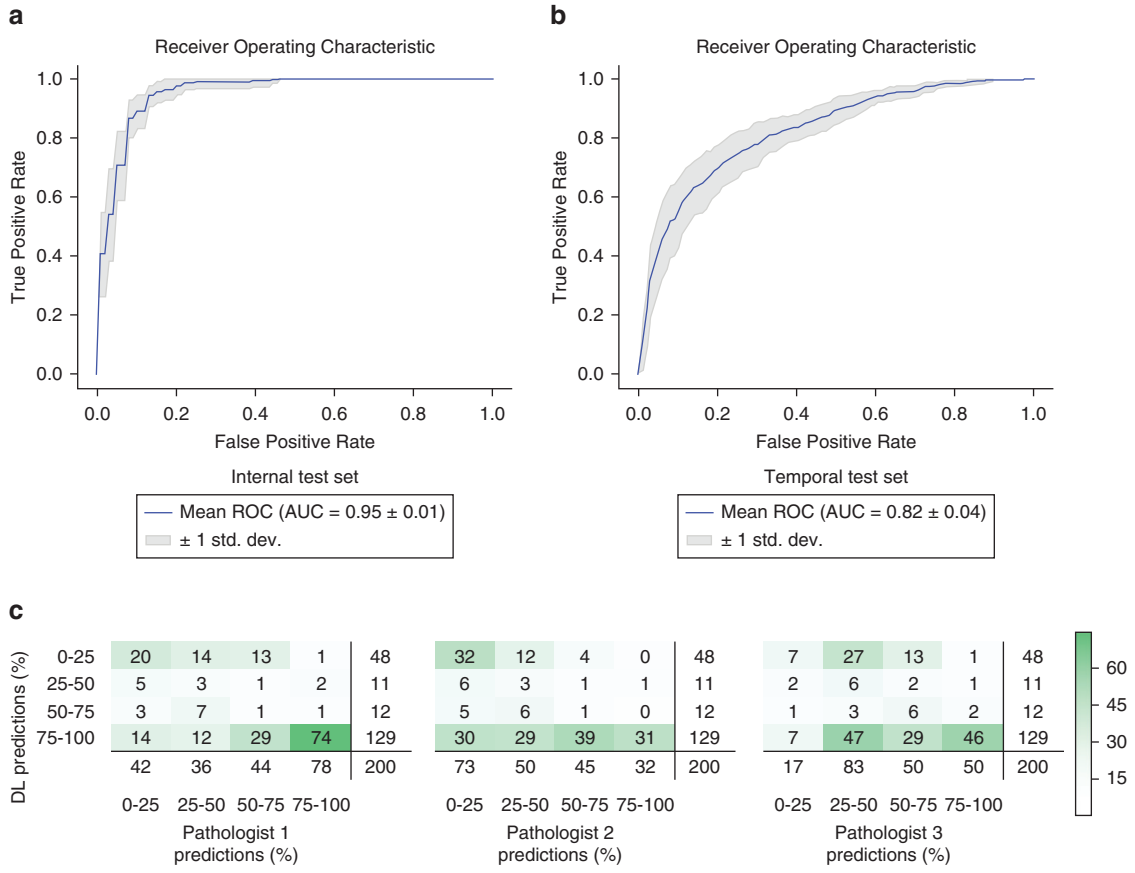


Figure 1. Performance evaluation of models for distinguishing MF from BIDs. (**a, b**) Unweighted macroaverage WSI-level ROC curves from 10-fold cross-validation for distinguishing MF from BIDs on (**a**) the internal test set (n = 71 WSI) and (**b**) the temporal test set (n = 200 WSI) using models trained at ×200 magnification with a patch size of 256 × 256 pixels. (**c**) Confusion matrices comparing the pathologists' level of suspicion for MF with the prediction scores of the best performing model (discretized into bins 0–25, 25–50, 50–75, and 75–100%) on the temporal test set (AUC = 0.908), which includes 200 WSIs comprising 103 MF and 97 BID cases. AUC, area under the curve; BID, benign inflammatory dermatosis; DL, deep learning; MF, mycosis fungoides; ROC, receiver operating characteristic; WSI, whole-slide image.

based model was significantly better on the easy cases than on the difficult cases of the temporal test set, as defined by the pathologists' scoring criteria.

The artificial comparison of the model with the performance of the 3 expert pathologists can be considered as a limitation of the study. To fairly assess the performance of the models, we impaired the diagnostic capabilities of the pathologists by forcing judgments to be made on 1 representative WSI, without any clinical information. In real-life practice, a pathologist would have access to (basic) clinical information or even photographs of the clinical presentation of the patient as well as immunohistochemical stainings and, in some instances, molecular testing. For future directions, it would be interesting to evaluate whether incorporation of such information could further improve the performance of the model through multimodal DL. In addition, in clinical practice, skin biopsies are evaluated at both low-power and high-power magnification (Subtil, 2019), whereas the DL models in this study were trained on single magnification levels. An increase in performance might also be seen when using multiresolution models, as has been shown with promising results for endometrial carcinoma subtype classification (Hong et al, 2021) and tissue segmentation in breast cancer (Van Rijthoven et al, 2021). However, most importantly, the performance of DL-based models significantly improves with increasing numbers of patients included in the training dataset. Large-scale experiments have shown that, although dependent on the tissue type and specific task, at least 10,000 WSIs are necessary for clinical-grade classification performance (Campanella et al, 2019), which is very difficult to achieve for rare diseases, such as MF. By pooling patient material from multiple institutions, considerably larger datasets can be generated, and this also enables the generation of an external cohort for model validation. DL algorithms trained on WSIs from a single institution, as in our study, often experience further performance deterioration when applied to WSIs from other institutions (Van der Laak et al, 2021). Even though we adopted the Macenko stain normalization method that can help to minimize uninformative color variations in WSIs from different institutions, performance deterioration is also to be expected from our model.

Possible confounding factors in our study include the higher proportion of male patients in the MF cohort, with possible sex-related differences in skin histology (Sandby-Møller et al, 2003). In addition, topographical differences in skin histology (Kakasheva-Mazhenkovska et al, 2011) may be relevant because the BID cohort contained more cases

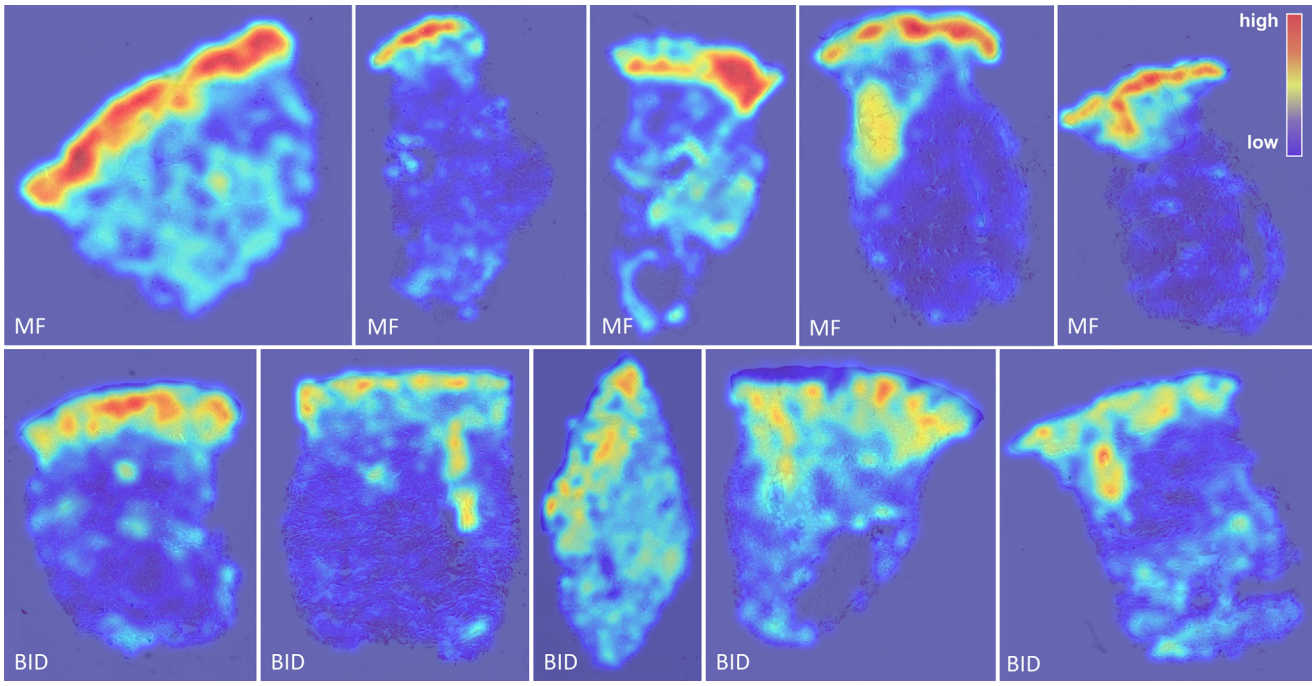


Figure 2. Attention heatmaps for the MF versus BIDs classification task. Representative attention heatmaps of 5 MF and 5 BID skin biopsies as generated by the best fold model trained at ×200 magnification. High attention is mostly clustered at the dermal–epidermal junction containing lymphocytic infiltrate, with good correspondence to the pathologist's ROIs. Red area = highest attention, blue area = lowest attention. BID, benign inflammatory dermatosis; MF, mycosis fungoides; ROI, region of interest.

from the head-and-neck region and upper extremity and fewer from the trunk than the MF cohort. However, in subgroup sensitivity analyses on the temporal test set, no significant differences in performance were observed between biopsy locations or between sexes.

In conclusion, although promising, our proof-of-concept study underscores the difficulty in distinguishing between early-stage MF and BIDs for both artificial intelligence models and human observers relying solely on WSIs from H&E slides. Nevertheless, our results support weakly supervised DL as a promising technique for the development of a screening tool to assist pathologists without CL expertise for timely referral of patients with suspected MF to a specialized CL center. It is expected that clinical-grade performance can be attained with the generation of larger and more diverse, multi-institutional datasets using a multimodal approach with incorporation of clinical information.

MATERIALS AND METHODS

WSI dataset construction

Representative skin biopsies from all patients with early stages of classic MF (stages IA–IIA) and only patch- or plaque-stage disease from patients with advanced stages of classic MF (stages IIB–III) were retrospectively included. Patients were identified from the registry of the Dutch Cutaneous Lymphoma Group, diagnosed after clinicopathological correlation between January 1, 2015 and December 7, 2023.

The Dutch Cutaneous Lymphoma Group is a collaboration of dermatologists and pathologists from all Dutch university medical centers, meeting 4–5 times each year to discuss new patients with suspected skin lymphoma. A diagnosis is synthesized after review of clinical photographs, clinical data, and a representative skin biopsy. All biopsies with tumor-stage disease and transformed large-cell variants of MF were excluded. In addition, patients with variants of MF, that is, folliculotropic MF, pagetoid reticulosis, and granulomatous slack skin, were excluded because these variants have other clinical and histological mimickers compared with patch and plaque stages of classic MF. WSIs of included patients with MF were not selected when it was unclear whether the MF diagnosis was based on that particular slide and when pathologists reviewed this tissue slide as nondiagnostic for MF. In addition to MF, we included representative skin biopsies from all patients with BIDs with either clinical and/or histological suspicion of MF but without MF diagnosis after clinicopathological correlation, sent in as consultation to the Leiden University Medical Center between January 1, 2015 and December 7, 2023. In addition, a diverse set of BIDs, diagnosed in the same study period, was selected (Supplementary Table S1). In these cases, there was no clinical or histopathological suspicion of MF, and immunohistochemistry typing of the lymphoid infiltrates and clinicopathological correlation was not performed in routine diagnostics.

From all included cases, H&E-stained slides from representative skin biopsies were collected from the archives of the Department of Pathology of the Leiden University Medical Center, including slides from in-house biopsies and locally stained slides retrieved from consultation biopsies. The H&E-stained slides were scanned using the Philips Ultra-Fast Scanner and the 3DHISTECH PANNORAMIC P480 scanner in a ratio of approximately 1:1 to create a WSI repository. The scanning resolution was 0.25 μm per pixel. Out-of-focus WSIs were rescanned. WSIs with troublesome fading of the H&E stain and WSIs without any tissue or missing the epidermis were excluded.

For WSI preprocessing, all images in the Philips iSyntax file format were converted to pyramidal TIFF using the Philips Pathology

Software Development Kit (available from https://www.usa.philips.com/healthcare/sites/pathology/about/sdk/). TD manually delineated the artifact-free and most representative tissue section on each slide to minimize redundant information in sequential sections. The ASAP 1.9 software (available from https://computationalpathologygroup.github.io/ASAP/) was used for this delineation, and the WSI was cropped to this section using a proprietary script provided by JK. Images in the MIRAX file format were cropped in QuPath 0.5.1 (available from https://github.com/qupath), exported as OME-TIFF images using Bio-Formats, and then converted to pyramidal TIFF using the libvips image processing library (available from https://github.com/libvips). For comparison, several models were trained using all available sections, that is, the entire tissue region on the slide. However, this approach did not significantly impact the results but did significantly increase the training duration of the models (data not shown).

In total, we established a 1-institution dataset of 924 H&E-stained WSIs from 585 patients, including 465 WSIs from 233 patients with early-stage MF, 152 WSIs from 96 patients with BIDs with clinical/histological suspicion of MF but without MF diagnosis after clinicopathological correlation, and 307 WSIs from 260 patients with BIDs without clinical/histological suspicion of MF (Figure 3). This dataset was divided into 2 cohorts: a development dataset encompassing the data used throughout the model development process, consisting of patients diagnosed from January 2015 to August 2021 (n = 724 WSI, 78.4% of total sample), and a temporal test set encompassing the data for further, independent validation, consisting of patients diagnosed from September 2021 to December 2023 (n = 200 WSI, 21.6% of total sample).

Expert opinion

To be able to compare the performance of our models with expert opinion, 3 pathologists with expertise in CL (PJ, AS, and MD) reviewed the same tissue sections as included in the temporal test set (n = 200 WSI), without clinical or additional histopathological information. All WSIs were scored in the following categories, dependent on certainty of the pathologist (between parentheses): BID (0–25%); uncertain, favor BID (26–50%); uncertain, favor MF (51–75%), and MF (76–100%). On the basis of the pathologists' scoring, the temporal test set was divided into 2 groups: (i) an easy cohort, containing cases where all 3 pathologists agreed on the diagnosis (binary decision) with at least 2 pathologists expressing certainty (either 0–25% for BIDs or 76–100% for MF), and (ii) a difficult cohort, containing the remainder (comprising cases where the pathologists either did not agree or there was substantial uncertainty).

DL model building and testing

To perform DL, we used CLAM (Lu et al, 2021) (code available at https://github.com/mahmoodlab/CLAM), a DL-based toolbox designed for weakly supervised, multiple-instance learning classification tasks in computational pathology. This was implemented using the Slideflow python package (version 2.3.0). WSIs, comprising numerous patches, were provided only with slide-level labels. These labels are utilized to train a classification model for

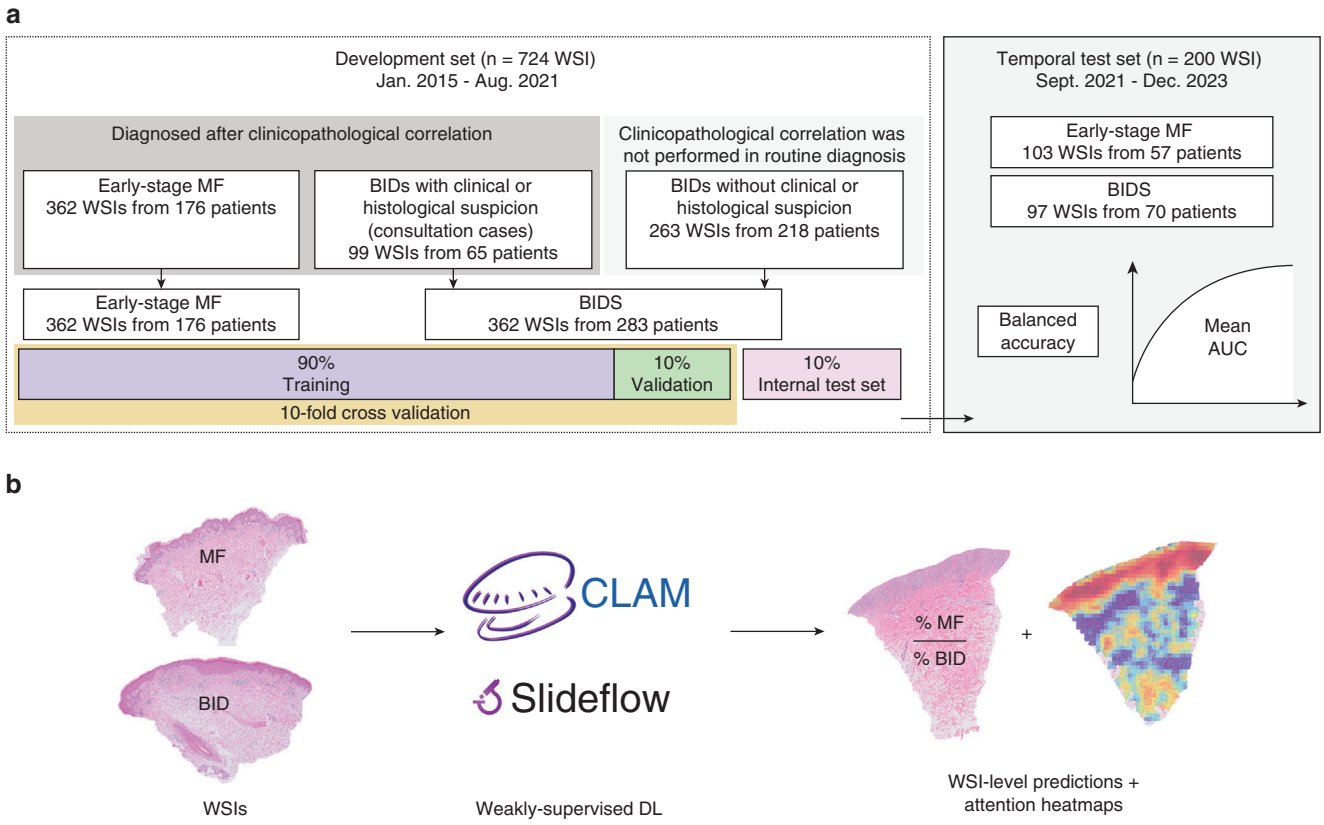


Figure 3. Flow diagram of study design. (**a**) Composition of the development and temporal test WSI datasets. (**b**) Overview of development of a weakly supervised DL model using (Lu et al, 2021) implemented in Slideflow (Dolezal et al, 2024). The CLAM and Slideflow logos are used with permission from their respective teams. AUC, area under the curve; Aug, August; BID, benign inflammatory dermatosis; CLAM, clustering-constrained attention multiple instance learning; Dec, December; DL, deep learning; Jan, January; MF, mycosis fungoides; Sept, September; WSI, whole-slide image.

predicting diagnostic labels on unseen WSIs. Among the 3 evaluated feature extractors, the UNI foundation model (code available at https://github.com/mahmoodlab/UNI) exhibited the highest performance and was consequently selected for reporting the results of this study (Supplementary Materials and Methods).

The development dataset (cohort January 2015−August 2021) was subdivided into an internal test subset, a validation subset, and a training subset. We specified a sample of 10% as the internal test set, ensuring no patients were shared between the test and training subsets. Then, we applied 10-fold Monte Carlo cross-validation to the remaining 90% of the development dataset, where each fold used 10% of the data for validation. The splits were selected such that the relative proportion of labels was the same in the training split as in the validation split, and there was no leakage between patients. Multiple WSIs from the same patient were kept within the same partition. The temporal test set (cohort September 2021−December 2023) was used to evaluate the models' generalization on unseen data from a different time period. There was no external test set from a different institute. To minimize uninformative color variations in the data, we normalized the images using the Macenko stain normalization method. All training was conducted in a Linux-based cluster environment using Quadro RTX 6000 (4608/24GB) graphics processing units. Performance was quantified using balanced accuracy and the AUC of receiver operating characteristic on the internal and temporal test sets.

Attention heatmaps

CLAM is able to produce interpretable heatmaps that allow users to visualize, within each WSI, the relative contribution of each tissue area to the model's slide-level predictions. For each case in the temporal test set, an attention heatmap was generated and evaluated and scored by TD for correspondence to the pathologist's ROI, regarded as pathological changes, in a 3-tiered scale: good (high attention to pathological changes in the majority of patches), moderate (high attention to pathological changes in approximately half of the patches), or poor (high attention to pathological changes in the minority of patches). In addition, we extracted the top 5 patches with the highest predictive value for both MF and BID diagnoses (Supplementary Figures S4 and S5).

Statistics

The degree of interobserver agreement between pathologists was assessed with Fleiss' multirater kappa statistic. The degree of agreement between the DL model and the individual pathologists was assessed with Cohen's linear weighted kappa. For categorical variables (sex and biopsy location), the *P*-value was obtained from the Pearson chi-square test. For continuous variables with non-normal distribution (patient age, slide age, and number of WSIs), the *P*-value was obtained from the Mann−Whitney *U* test. All *P*-values below .05 were regarded statistically significant. Analyses were performed using IBM Statistical Package for the Social Sciences, version 28, and Microsoft Excel, version 2403.

ETHICS STATEMENT

The study was performed in accordance with the Declaration of Helsinki. Written informed consent was not obtained owing to the rarity of CLs and the long inclusion period of over 20 years, during which many patients were no longer under care or had passed away. Patients who objected to the use of their medical data or residual materials were excluded. Ethical approval was granted by the non-Medical Research Involving Human Subjects Act committee of division 4 of the Leiden University Medical Center (nWMO-D4-2022-019), and review by the Medical Ethics Committee was not required.

DATA AVAILABILITY STATEMENT

The whole-slide image dataset related to this article cannot be made publicly available owing to legal restraints but is made available to interested research partners upon reasonable request. Requests for sharing of data should be addressed to the corresponding author, TD, within 15 years of the date of publication of this article and should include a scientific proposal. The model code is available at https://github.com/Sbrussee/MF_BID_MIL.

ORCIDs

Thom Doeleman: http://orcid.org/0000-0003-4568-9510
Siemen Brussee: http://orcid.org/0009-0005-8858-1544
Liesbeth Hondelink: http://orcid.org/0000-0002-7173-7443
Daniëlle Westerbeek: http://orcid.org/0000-0001-6004-8653
Ana Marta Sequeira: http://orcid.org/0000-0001-9134-5255
Pieter Valkema: http://orcid.org/0000-0001-7189-4441
Patty Jansen: http://orcid.org/0000-0002-2361-9303
Junling He: http://orcid.org/0000-0003-0563-4822
Maarten Vermeer: http://orcid.org/0000-0002-5872-4613
Koen Quint: http://orcid.org/0000-0002-3267-5630
Marijke van Dijk: http://orcid.org/0009-0003-6344-5736
Fons Verbeek: http://orcid.org/0000-0003-2445-8158
Jesper Kers: http://orcid.org/0000-0002-2418-5279
Anne Schrader: http://orcid.org/0000-0003-3483-0558

CONFLICT OF INTEREST

The authors state no conflict of interest.

ACKNOWLEDGMENTS

This work was supported by an unrestricted Fellowship grant of Stichting Hanarth Fonds in The Netherlands. We would like to acknowledge Rein Willemze and all members of the Dutch Cutaneous Lymphoma Group for establishing the registry of the Dutch Cutaneous Lymphoma Group that served as the basis for the whole-slide image dataset developed for this study.

AUTHORS CONTRIBUTIONS

Conceptualization: TD, JK, AMS; Data Curation: TD, DW; Investigation: TD, DW, PJ, MvD, AMS; Methodology: SB, LH, AMS, TD; Software: SB, AMS; Supervision: AMS, MV; Writing - Original Draft Preparation: TD, LH, DW, AMS; Writing - Review and Editing: PV, PJ, JH, MV, KQ, MvD, FV, JK, LH

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at www.jidonline.org, and at https://doi.org/10.1016/j.jid.2024.07.036.

REFERENCES

Border SP, Sarder P. From what to why, the growing need for a focus shift toward explainability of AI in digital pathology. Front Physiol 2021;12:821217.

Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301−9.

Cerroni L. Skin lymphoma: the illustrated guide. Chichester: John Wiley & Sons; 2020.

Dolezal JM, Kochanny S, Dyer E, Ramesh S, Srisuwananukorn A, Sacco M, et al. Slideflow: deep learning for digital histopathology with real-time whole-slide visualization. BMC Bioinformatics 2024;25:134.

Guitart J, Kennedy J, Ronan S, Chmiel JS, Hsiegh YC, Variakojis D. Histologic criteria for the diagnosis of mycosis fungoides: proposal for a grading system to standardize pathology reporting. J Cutan Pathol 2001;28:174−83.

Hondelink LM, Hüyük M, Postmus PE, Smit VTHBM, Blom S, von der Thüsen JH, et al. Development and validation of a supervised deep learning algorithm for automated whole-slide programmed death-ligand 1 tumour proportion score assessment in non-small cell lung cancer. Histopathology 2022;80:635−47.

Hong R, Liu W, DeLair D, Razavian N, Fenyö D. Predicting endometrial cancer subtypes and molecular features from histopathology images using multi-resolution deep learning models. Cell Rep Med 2021;2:100400.

Kakasheva-Mazhenkovska L, Milenkova L, Gjokik G, Janevska V. Variations of the histomorphological characteristics of human skin of different body regions in subjects of different age. Prilozi 2011;32:119−28.

Karabulut YY, Dinç U, Köse EÇ, Türsen Ü. Deep learning as a new tool in the diagnosis of mycosis fungoides. Arch Dermatol Res 2023;315:1315–22.

Kelati A, Gallouj S, Tahiri L, Harmouche T, Mernissi FZ. Defining the mimics and clinico-histological diagnosis criteria for mycosis fungoides to minimize misdiagnosis. Int J Womens Dermatol 2017;3:100–6.

Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.

Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 2021;5:555–70.

Mun Y, Paik I, Shin SJ, Kwak TY, Chang H. Yet another automated Gleason grading system (YAAGGS) by weakly supervised deep learning. NPJ Digit Med 2021;4:99.

Sandby-Møller J, Poulsen T, Wulf HC. Epidermal thickness at different body sites: relationship to age, gender, pigmentation, blood content, skin type and smoking habits. Acta Derm Venereol 2003;83:410–3.

Scheurer J, Ferrari C, Berenguer Todo Bom L, Beer M, Kempf W, Haug L. Semantic segmentation of histopathological slides for the classification of cutaneous lymphoma and eczema. In: Papież B, Namburete A, Yaqub M, Noble J, editors. Annual Conference on Medical Image Understanding and Analysis. Cham: Springer; 2020. p. 26–42.

Subtil A. Diagnosis of cutaneous lymphoid infiltrates: a visual approach to differential diagnosis and knowledge gaps. Cham: Springer; 2019.

Van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nat Med 2021;27:775–84.

van Doorn R, Van Haselen CW, van Voorst Vader PC, Geerts ML, Heule F, de Rie M, et al. Mycosis fungoides: disease evolution and prognosis of 309 Dutch patients. Arch Dermatol 2000;136:504–10.

Van Rijthoven M, Balkenhol M, Siliņa K, Van Der Laak J, Ciompi F. HookNet: multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. Med Image Anal 2021;68:101890.

Willemze R, Jaffe ES, Burg G, Cerroni L, Berti E, Swerdlow SH, et al. WHO-EORTC classification for cutaneous lymphomas. Blood 2005;105:3768–85.

SUPPLEMENTARY MATERIALS AND METHODS

Materials

Preprocessing and training procedures were conducted using the Slideflow python package (version 2.3.0) (Dolezal et al, 2024). The code used is available on GitHub. Training was conducted in a Linux-based cluster environment using Quadro RTX 6000 (4608/24GB) graphics processing units.

Image handling

For each whole-slide image (WSI), a tissue mask was created by applying the following sequential steps on the thumbnail images: (i) Gaussian filtering to remove artifacts and out-of-focus areas, (ii) transformation from RGB (red, green, and blue) to HSV (hue, saturation, and value) color space, (iii) contrast-limited adaptive histogram equalization for contrast enhancement on the V channel, and (iv) Otsu thresholding (Otsu, 1979). Furthermore, areas unsuitable for patching were identified using blur detection (Wu et al, 2015).

The area within the tissue mask suitable for patching was downsampled into tiles according to specified tile size and magnification. To minimize uninformative color variations in the data, we normalized the images using the Macenko stain normalization method (Macenko et al, 2009). This feature extraction resulted in a feature bag for each tile of each WSI in our dataset.

Dataset split

Our dataset consisted of a development dataset (n = 724 WSIs) and an independent temporal test set (n = 200 WSIs). This temporal test set (cohort September 2021−December 2023) was used to evaluate the models' generalization on unseen data from a different time period. There was no external test set from a different institute. The development dataset (cohort January 2015−August 2021) was subdivided into an internal test subset, a validation subset, and a training subset. We specified a sample of 10% as our internal test set such that no patients were shared between the test and training sets. Then, we applied 10-fold cross-validation to the remaining 90% of our internal dataset, where, for each fold, 10% of the data was used for validation. The splits were selected such that the relative number of labels was the same in the training split as in the validation split, and there was no leakage between patients. If there were multiple WSIs from 1 patient, they were placed in the same subset.

Training procedure

During training, we used the Leslie Smith LR Range test (Smith and Topin, 2019) to find an optimal learning rate on the basis of the data. For scheduling the learning rate, we used the 1 cycle scheduler (Smith and Topin, 2019). We trained for 32 epochs and used a batch size of 64 and a bag size of 512. For clustering-constrained attention multiple instance learning (CLAM), we used the small model size (1024, 512, 256) and default single-attention-branch CLAM model. For CLAM's bag loss, the cross-entropy bag loss was used, with a bag weight of 0.7. Adam (Kingma and Ba, 2014[1]) was used as the optimization algorithm. For CLAM's instance loss, cross-entropy was also used. We sampled 8 positive/negative patches during instance-level training. In each validation split, we trained a CLAM model on the training set and measured the performance both on the internal test set and on our independent temporal test set. The results were averaged over the splits, and the SD was also calculated.

Feature extractors

During feature extraction, we generated bags of feature representations for each tile, in each WSI of our dataset. To extract features, we used the following models: (i) ImageNet-pretrained ResNet50, a ResNet50 model pretrained on the ImageNet dataset (Deng et al, 2009; He et al, 2016); (ii) CTransPath, a transformer-based model trained using contrastive self-supervised learning on 15 million H&E-stained patches from The Cancer Genome Atlas and Pathology AI Platform WSIs across cancer types (Wang et al, 2022); and (iii) pretrained UNI foundation model, a general-purpose self-supervised model for pathology, pretrained using >100 million images from over 100,000 diagnostic H&E-stained WSIs (>77 TB of data) across 20 major tissue types (Chen et al, 2024).

Evaluation metrics

For model evaluation, we used the balanced accuracy and area under the receiver operating characteristic curve score. Given predicted label $\hat{y}_i$ and true label i for any WSI i, let us define false positives (denoted as FPs), true negatives (denoted as TNs), true positives (denoted as TPs), and false negatives (denoted as FNs) as follows:

$$\mathrm{FP} = \sum_{i=1}^{N} 1\left(y_i = 0 \wedge \hat{y}_i = 1\right), \quad \mathrm{TN} = \sum_{i=1}^{N} 1\left(y_i = 0 \wedge \hat{y}_i = 0\right),$$
$$\mathrm{TP} = \sum_{i=1}^{N} 1\left(y_i = 1 \wedge \hat{y}_i = 1\right), \quad \mathrm{FN} = \sum_{i=1}^{N} 1\left(y_i = 1 \wedge \hat{y}_i = 0\right)$$

Using these definitions, we can define the false positive rate (denoted as FPR) and true positive rate (denoted as TPR) as follows:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{1}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

Now, we can define the balanced accuracy and area under the curve score as follows:

$$\text{Balanced Accuracy} = \frac{1}{2}\sum_{i=1}^{2} \frac{TP_i}{TP_i + FN_i} \tag{3}$$

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(fpr)\, d(fpr) \tag{4}$$

Where the area under the curve score integrates the decision threshold on the basis of the true positive and false positive rates. The roc_auc_score function from the scikit-

[1] Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv 2014.

learn machine learning library was used with default values to compute the area under the curve from prediction scores.

SUPPLEMENTARY REFERENCES

Chen RJ, Ding T, Lu MY, Williamson DF, Jaume G, Song AH, et al. Towards a general-purpose foundation model for computational pathology. Nat Med 2024;30:850–62.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Miami: IEEE Publications; 2009. p. 248–55.

Dolezal JM, Kochanny S, Dyer E, Ramesh S, Srisuwananukorn A, Sacco M, et al. Slideflow: deep learning for digital histopathology with real-time whole-slide visualization. BMC Bioinformatics 2024;25:134.

He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE Publications; 2016. p. 770–8.

Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Guan X, et al. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE international symposium on biomedical imaging: from nano to macro. Boston: IEEE Publications; 2009. p. 1107–10.

Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 1979;9:62–6.

Smith LN, Topin N. Super-convergence: very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications. SPIE; 2019. p. 11006.

Wang X, Yang S, Zhang J, Wang M, Zhang J, Yang W, et al. Transformer-based unsupervised contrastive learning for histopathological image classification. Med Image Anal 2022;81:102559.

Wu H, Phan JH, Bhatia AK, Cundiff CA, Shehata BM, Wang MD. Detection of blur artifacts in histopathological whole-slide images of endomyocardial biopsies. In: 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE Publications; 2015. p. 727–30.

a

No. of WSI		Ground-truth		
		BID	MF	Total
Path. 1	BID	**71**	7	78
	MF	26	**96**	122
	Total	97	103	**200**

b

No. of WSI		Ground-truth		
		BID	MF	Total
Path. 2	BID	**90**	33	123
	MF	7	**70**	77
	Total	97	103	**200**

c

No. of WSI		Ground-truth		
		BID	MF	Total
Path. 3	BID	**68**	32	100
	MF	29	**71**	100
	Total	97	103	**200**

Supplementary Figure S1. Pathologist's scoring of temporal test set (n = 200 WSIs). (**a**) Pathologist 1: balanced accuracy = 83.2%, sensitivity = 93.2%, and specificity = 73.2%. (**b**) Pathologist 2: balanced accuracy = 80.4%, sensitivity = 68.0%, and specificity = 92.8%. (**c**) Pathologist 3: balanced accuracy = 69.5%, sensitivity = 68.9%, and specificity = 70.1%. Interpathologist agreement for 2 rating categories: Fleiss' multirater kappa = 0.473.

a **CLAM-ResNet50**

	BA on temporal test set:			AUC on temporal test set:		
Tile size || Magnification	400x	200x	100x	400x	200x	100x
128x128	0,629 +- 0,014	0,626 +- 0,022	0,647 +- 0,028	0,644 +- 0,021	0,649 +- 0,028	0,677 +- 0,038
256x256	0,640 +- 0,017	0,655 +- 0,031	0,593 +- 0,051	0,674 +- 0,023	**0,694 +- 0,042**	0,596 +- 0,081
512x512	**0,657 +- 0,022**	0,584 +- 0,059	0,558 +- 0,038	0,688 +- 0,036	0,596 +- 0,067	0,526 +- 0,068
	BA on internal test set:			AUC on internal test set:		
Tile size || Magnification	400x	200x	100x	400x	200x	100x
128x128	0,845 +- 0,048	0,786 +- 0,036	0,822 +- 0,029	0,893 +- 0,048	0,841 +- 0,032	0,888 +- 0,022
256x256	**0,850 +- 0,042**	0,813 +- 0,025	0,714 +- 0,112	**0,918 +- 0,039**	0,872 +- 0,033	0,734 +- 0,173
512x512	0,772 +- 0,040	0,673 +- 0,129	0,605 +- 0,081	0,796 +- 0,051	0,710 +- 0,156	0,543 +- 0,137

b **CLAM-CTransPath**

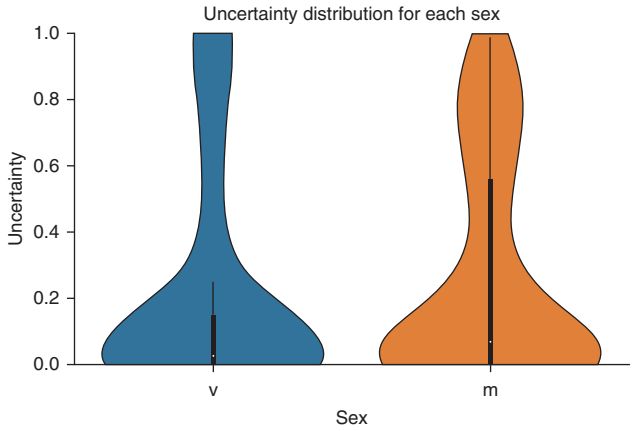
	BA on temporal test set:			AUC on temporal test set:		
Tile size || Magnification	400x	200x	100x	400x	200x	100x
128x128	0,580 +- 0,025	0,629 +- 0,053	0,679 +- 0,018	0,563 +- 0,028	0,629 +- 0,108	0,718 +- 0,018
256x256	0,656 +- 0,021	**0,685 +- 0,022**	0,684 +- 0,059	0,681 +- 0,034	**0,727 +- 0,028**	0,727 +- 0,081
512x512	0,626 +- 0,018	0,646 +- 0,105	0,593 +- 0,083	0,654 +- 0,019	0,672 +- 0,122	0,580 +- 0,118
	BA on internal test set:			AUC on internal test set:		
Tile size || Magnification	400x	200x	100x	400x	200x	100x
128x128	0,804 +- 0,050	0,754 +- 0,132	0,807 +- 0,032	0,854 +- 0,035	0,757 +- 0,137	0,892 +- 0,029
256x256	0,806 +- 0,026	0,849 +- 0,029	0,717 +- 0,092	0,867 +- 0,021	0,897 +- 0,029	0,756 +- 0,113
512x512	**0,882 +- 0,041**	0,732 +- 0,167	0,64 +- 0,136	**0,921 +- 0,037**	0,765 +- 0,187	0,652 +- 0,184

c **CLAM-UNI**

	BA on temporal test set:			AUC on temporal test set:		
Tile size || Magnification	400x	200x	100x	400x	200x	100x
128x128	0,628 +- 0,023	0,692 +- 0,013	0,743 +- 0,036	0,645 +- 0,025	0,741 +- 0,018	0,803 +- 0,030
256x256	0,690 +- 0,024	**0,762 +- 0,039**	0,574 +- 0,098	0,744 +- 0,021	**0,827 +- 0,044**	0,565 +- 0,136
512x512	0,733 +- 0,039	0,647 +- 0,087	0,605 +- 0,744	0,787 +- 0,041	0,663 +- 0,135	0,621 +- 0,109
	BA on internal test set:			AUC on internal test set:		
Tile size || Magnification	400x	200x	100x	400x	200x	100x
128x128	0,894 +- 0,020	**0,925 +- 0,032**	0,880 +- 0,028	0,951 +- 0,021	**0,975 +- 0,018**	0,940 +- 0,017
256x256	0,863 +- 0,018	0,906 +- 0,018	0,617 +- 0,137	0,924 +- 0,014	0,955 +- 0,014	0,608 +- 0,173
512x512	0,876 +- 0,027	0,680 +- 0,138	0,685 +- 0,123	0,941 +- 0,027	0,684 +- 0,207	0,696 +- 0,166

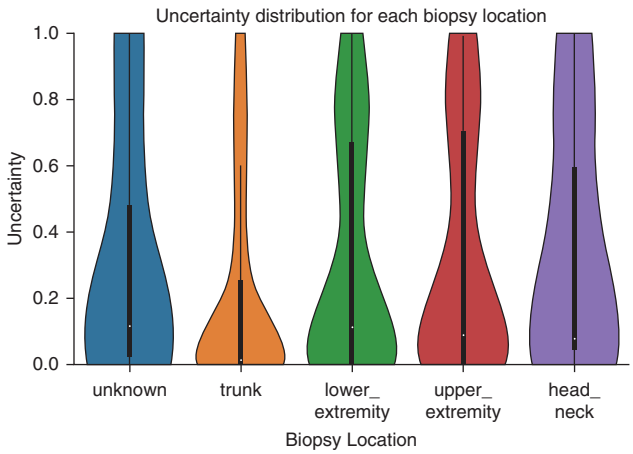
Supplementary Figure S2. Classification performance of 10-fold cross-validated models for distinguishing between MF and BIDs across various combinations of magnification level and patch size. This evaluation was conducted using the CLAM method with 3 different encoders: (**a**) the default ResNet50 encoder, (**b**) the CTransPath encoder, and (**c**) the UNI encoder. For the internal test set, each fold is presented with n = 71 WSIs, which consisted of a sample of 10% from the development WSI dataset (cohort January 2015−August 2021). For the temporal test set, each fold is presented with n = 200 WSIs (cohort September 2021−December 2023). Combination of magnification level and patch size with highest mean AUC and mean BA on the internal or temporal test set is provided in bold. AUC, area under the curve; BA, balanced accuracy; CLAM, clustering-constrained attention multiple instance learning; WSI, whole-slide image.

a



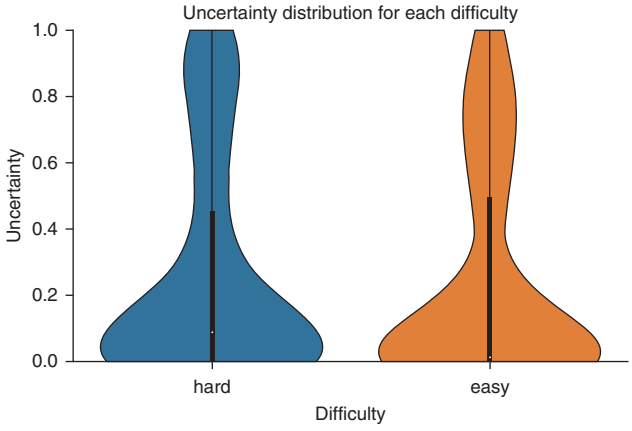
SEX	ACCURACY	UNCERTAINTY
MALE	0.736	0.263
FEMALE	0.813	0.206

b



BIOPSY_LOCATION	ACCURACY	UNCERTAINTY
HEAD_NECK	0.5	0.3240351824762653
LOWER_EXTREMITY	0.75	0.28744321300236336
TRUNK	0.801980198019802	0.18859459116116847
UNKNOWN	0.7142857142857143	0.31241516706715583
UPPER_EXTREMITY	0.7575757575757576	0.2948652549753461

c

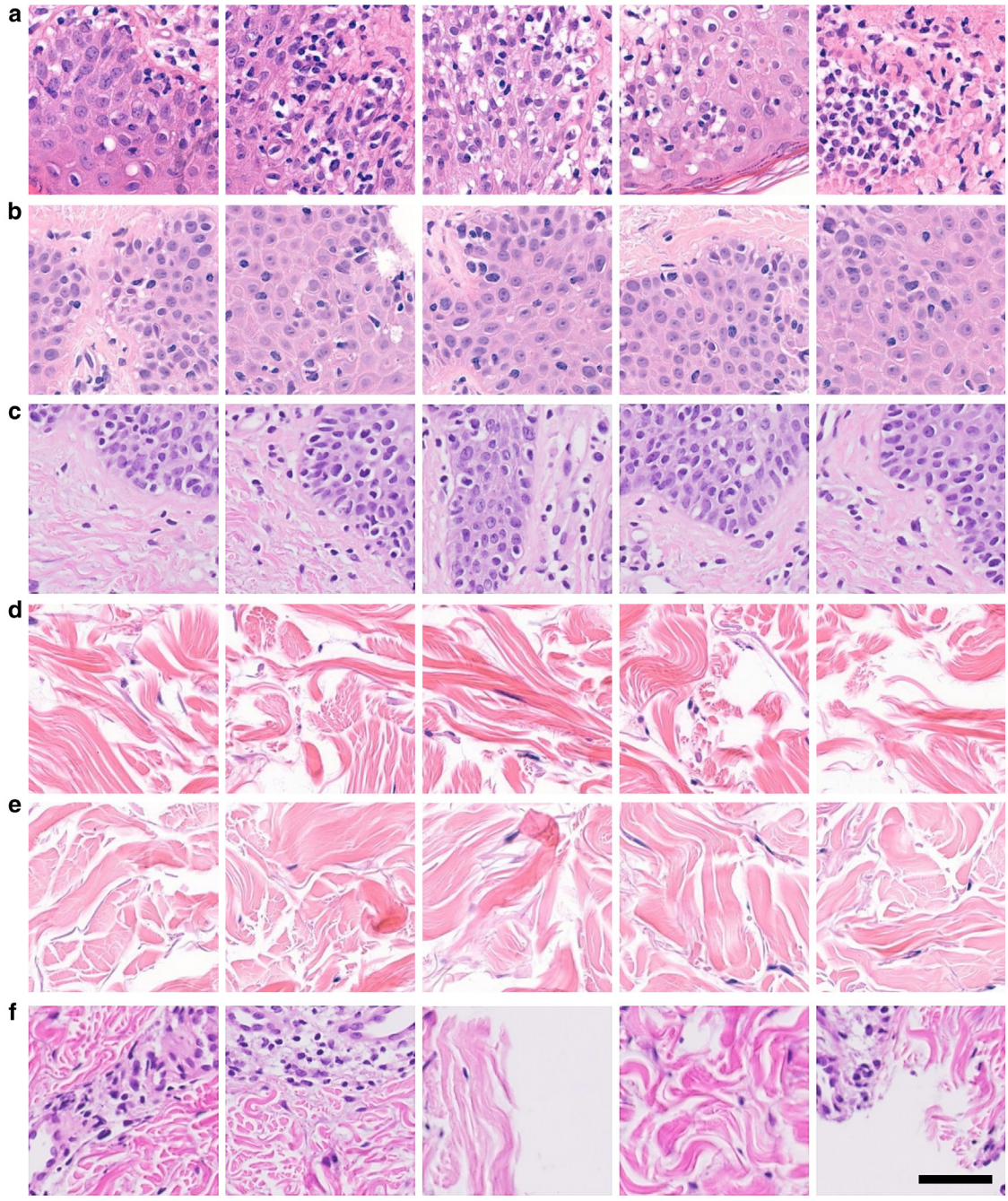


DIFFICULTY	ACCURACY	UNCERTAINTY
EASY	0.8607594936708861	0.22068918496630235
HARD	0.7024793388429752	0.25584922876017224

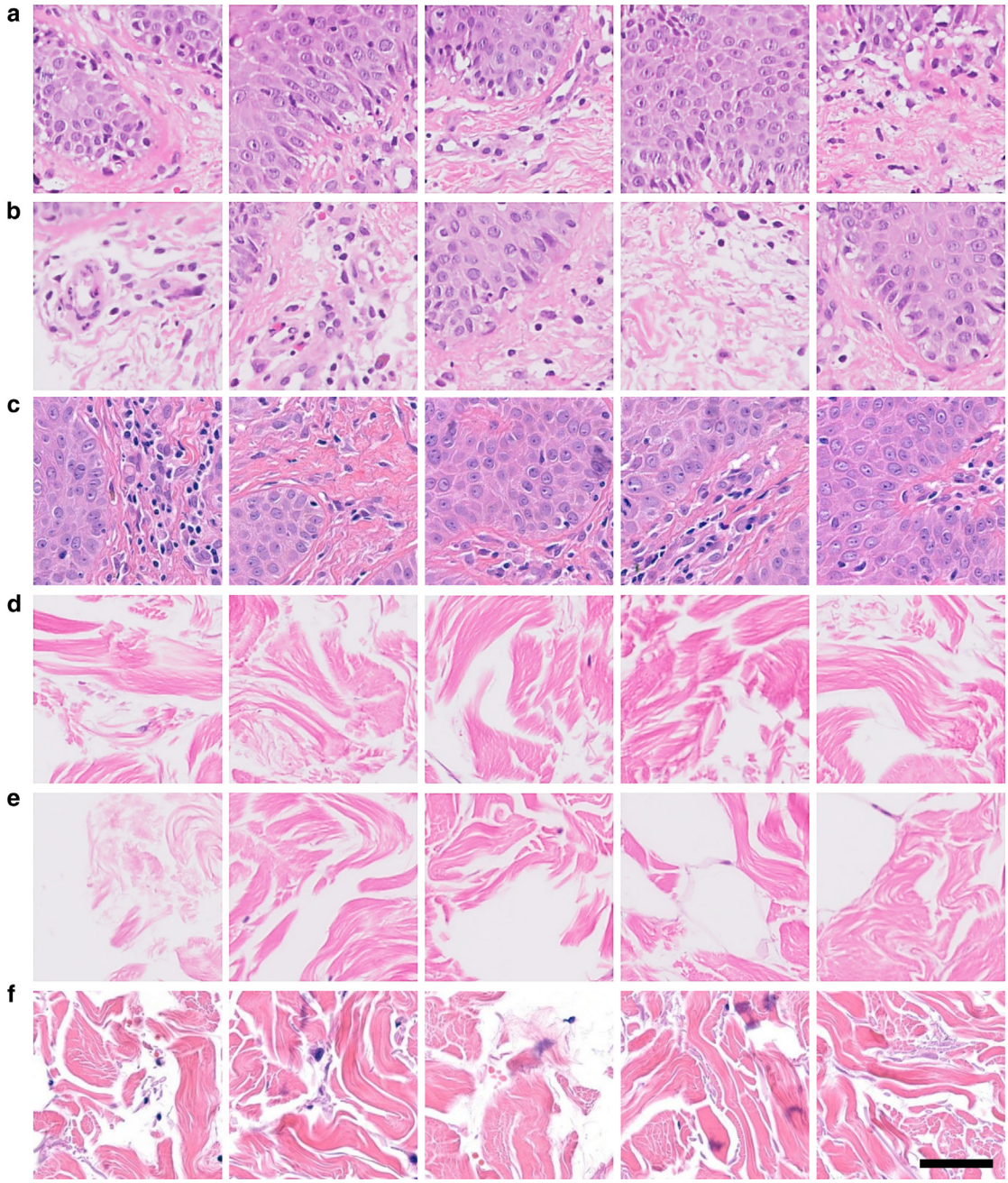
GROUP	VALUE	H-STATISTIC	P-VALUE
DIFFICULTY	uncertainty	5.132	0.023485177984857747
DIFFICULTY	accuracy	6.627	0.010043378670574536
BIOPSY_LOCATION	uncertainty	1.121	0.024261812509031016
BIOPSY_LOCATION	accuracy	4.138	0.3876516682533546
SEX	uncertainty	2.243	0.13420865750432068
SEX	accuracy	1.552	0.2129053168455239

$$\text{entropy} = -\left(p0 \cdot \log_2(P0) + P1 \cdot \log_2(P1)\right)$$

Supplementary Figure S3. Subgroup sensitivity analyses for distinguishing MF from BIDs on the temporal test set. Both accuracy and uncertainty were compared for the following subgroups: (**a**) sex (male vs female), (**b**) biopsy location (head and neck, upper extremity, lower extremity, trunk, and unknown), and (**c**) difficulty (easy vs difficult cohort). Entropy was used as a mathematical representation of uncertainty. This entropy value reflects the uncertainty in the model's predictions. Higher values would indicate more uncertainty (ie, the model is less confident about its predictions), whereas lower values would indicate less uncertainty (ie, the model is more confident about its predictions). The Kruskal–Wallis test was used to compare the samples. Statistically significant values are presented in bold. n = 200 WSIs. WSI, whole-slide image.



Supplementary Figure S4. Attention tiles for MF. (**a**–**c**) Top 5 most attended and (**d**–**f**) top 5 least attended tiles for 3 different representative MF WSIs from the temporal test set, plotted for the best fold model trained at ×200 magnification level and 256 × 256-pixel patch size. Bar = 50 μm. MF, mycosis fungoides; WSI, whole-slide image.



Supplementary Figure S5. Attention files for BIDs. (**a**–**c**) Top 5 most attended and (**d**–**f**) top 5 least attended tiles for 3 different representative BID WSIs from the temporal test set, plotted for the best fold model trained at ×200 magnification level and 256 × 256-pixel patch size. Bar = 50 μm. BID, benign inflammatory dermatosis; WSI, whole-slide image.

Supplementary Table S1. Composition of Training WSI Subset of BIDs without Clinical or Histological Suspicion of MF

Diagnosis or Differential Diagnosis	Number of Patients	Number of WSI
Atrophic lichen planus	3	3
(Chilblain) lupus erythematosus	2	2
Benign lichenoid keratosis	20	23
Cutaneous lupus erythematosus	21	28
Dermatomyositis	1	1
Erythema exsudativum multiforme	5	5
Exanthematous drug reaction, viral exanthema	1	1
Extragenital lichen sclerosus	8	10
Fixed drug eruption	1	1
Graft-versus-host disease	21	22
Graft-versus-host disease, drug reaction	2	2
Graft-versus-host disease, spongiotic dermatitis	2	2
Hypertrophic lichen planus	1	2
Large-plaque parapsoriasis	1	1
Lichen planus	15	18
Lichen planus pigmentosus (ashy dermatosis)	1	1
Lichen planus pigmentosus (ashy dermatosis), macular amyloidosis	1	2
Lichen planus, lichenoid drug reaction	2	2
Lichen planus, lichenoid drug reaction, lichenoid purpuric dermatosis	1	1
Lichen sclerosus	2	2
Lichenoid dermatitis	8	10
Lichenoid drug reaction	5	6
Lichenoid drug reaction, pityriasis lichenoides, lichen planus	1	1
Lupus erythematosus/lichen planus overlap syndrome	1	1
Perivascular dermatitis	1	1
Perivascular dermatitis, consistent with drug reaction	2	2
Pityriasis lichenoides chronica	5	7
Pityriasis lichenoides et varioliformis acuta	5	5
Psoriasiform dermatitis	1	1
Psoriasis	10	12
Spongiotic and perivascular dermatitis, consistent with drug reaction	1	2
Spongiotic dermatitis	52	59
Spongiotic dermatitis, lichen simplex chronicus	2	6
Spongiotic dermatitis, photoallergic reaction, drug reaction	1	1
Spongiotic dermatitis, spongiotic drug reaction	11	13
Spongiotic drug reaction	4	4
Toxic epidermal necrolysis-like cutaneous lupus erythematosus	1	1
Viral exanthema, graft-versus-host disease, drug reaction	2	2
Total	224[1]	263

Abbreviations: BID, benign inflammatory dermatosis; MF, mycosis fungoides; WSI, whole-slide image.

[1]Some patients provided biopsies with multiple (differential) diagnoses, so the total number exceeds the 218 patients included.