Modern Pathology (2021) 34:895–903
https://doi.org/10.1038/s41379-020-00712-7

ARTICLE

Artificial neural networks and pathologists recognize basal cell carcinomas based on different histological patterns

Susanne Kimeswenger[1,2,3] · Philipp Tschandl[4] · Petar Noack[5] · Markus Hofmarcher[6] · Elisabeth Rumetshofer[6] · Harald Kindermann[7] · Rene Silye[5] · Sepp Hochreiter[6] · Martin Kaltenbrunner[2,3] · Emmanuella Guenova[8,9] · Guenter Klambauer[6] · Wolfram Hoetzenecker[1]

Received: 1 August 2020 / Revised: 14 October 2020 / Accepted: 15 October 2020 / Published online: 13 November 2020

Abstract

Recent advances in artificial intelligence, particularly in the field of deep learning, have enabled researchers to create compelling algorithms for medical image analysis. Histological slides of basal cell carcinomas (BCCs), the most frequent skin tumor, are accessed by pathologists on a daily basis and are therefore well suited for automated prescreening by neural networks for the identification of cancerous regions and swift tumor classification.

In this proof-of-concept study, we implemented an accurate and intuitively interpretable artificial neural network (ANN) for the detection of BCCs in histological whole-slide images (WSIs). Furthermore, we identified and compared differences in the diagnostic histological features and recognition patterns relevant for machine learning algorithms vs. expert pathologists.

An attention-ANN was trained with WSIs of BCCs to identify tumor regions ($n = 820$). The diagnosis-relevant regions used by the ANN were compared to regions of interest for pathologists, detected by eye-tracking techniques.

This ANN accurately identified BCC tumor regions on images of histologic slides (area under the ROC curve: 0.993, 95% CI: 0.990–0.995; sensitivity: 0.965, 95% CI: 0.951–0.979; specificity: 0.910, 95% CI: 0.859–0.960). The ANN implicitly calculated a weight matrix, indicating the regions of a histological image that are important for the prediction of the network. Interestingly, compared to pathologists' eye-tracking results, machine learning algorithms rely on significantly different recognition patterns for tumor identification ($p < 10^{-4}$).

To conclude, we found on the example of BCC WSIs, that histopathological images can be efficiently and interpretably analyzed by state-of-the-art machine learning techniques. Neural networks and machine learning algorithms can potentially enhance diagnostic precision in digital pathology and uncover hitherto unused classification patterns.

Introduction

Digital pathology, i.e., the management and clinical interpretation of information retrieved from digitalized histology slides, aims to improve the safety, quality, and accuracy of

These authors contributed equally: Emmanuella Guenova, Guenter Klambauer, Wolfram Hoetzenecker

Supplementary information The online version of this article (https://doi.org/10.1038/s41379-020-00712-7) contains supplementary material, which is available to authorized users.

✉ Wolfram Hoetzenecker
wolfram.hoetzenecker@kepleruniklinikum.at

[1] Johannes Kepler University Linz, Kepler University Hospital Linz, Department of Dermatology, Linz, Austria

[2] Johannes Kepler University Linz, Institute of Applied Physics, Department of Soft Matter Physics, Linz, Austria

[3] Johannes Kepler University Linz, Linz Institute of Technology, Soft Materials Lab, Linz, Austria

[4] Medical University of Vienna, Department of Dermatology, Vienna, Austria

[5] Kepler University Hospital Linz, Department of Pathology and Microbiology, Linz, Austria

[6] Johannes Kepler University Linz, Institute for Machine Learning, Linz, Austria

[7] University of Applied Sciences, Upper Austria, Marketing and Electronic Business, Steyr, Austria

[8] University of Lausanne, Faculty of Biology and Medicine, Department of Dermatology, Lausanne, Switzerland

[9] University Hospital of Zurich, Department of Dermatology, Zurich, Switzerland

pathological diagnoses [1]. Digital pathology in combination with machine learning can extend the scope of digital pathology far beyond the possibilities of today [2]. Although there have been great advances in the field of medical imaging using artificial intelligence (AI), considerable challenges in the field of histopathology remain.

First, methods for machine learning in histopathology traditionally use downscaling of whole-slide images (WSIs), online repository WSIs, handcrafted features or manually annotated regions of interest (ROI) [3]. In contrast, real-life pathological cases consist of many WSIs accompanied by the patient's (single) diagnosis and demographic metadata (weakly labeled data). Since WSIs, due to their size, cannot be processed through a neural network as a whole at full resolution, one approach is to split the information into several tiles. In the case of cancer detection, however, only some tiles contain cancerous tissue [4, 5]. Thus, parts relevant for the diagnosis might be missed by this approach as tiles for classification are commonly chosen randomly [6]. Skin neoplasm WSI classification using machine learning is an emerging field with several publications in the recent past. A common approach is multiple instance learning (MIL), a method that benefits from the property that "bags" of "instances" can be classified labeled only on the bag—but not on the instance level (=weakly labeled). In histopathology a "bag" is a single WSI that is weakly labeled, e.g., as "basal cell carcinoma" (BCC) or "non-BCC." The WSIs are divided into non-overlapping smaller images (tiles) that represent the "instances" of the "bag" (see also tiling in Supplementary Fig. 2) [5, 7, 8]. Campanella et al. [5] successfully used an MIL method for the classification of prostate carcinoma, BCC of the skin, and breast cancer metastases using a recurrent neural network as a classifier. In another study, deep learning outperformed 11 pathologists in the classification of histopathological melanoma images [9]. However, the second challenge, namely interpretability, remains. Interpretability and the process of learning and decision-making of AI in comparison to humans is a key question in modern health care. Interestingly, it has been shown that human and machine attention do not coincide in natural language processing [10, 11]. A recent study that compares human and artificial attention mechanisms in various applications demonstrates that in addition to differences, the closer the artificial attention is to human attention, the better the performance [11]. Such studies are important for making deep networks more transparent and explainable for higher-level computer vision tasks.

In the present study, we generated automated detection of BCCs, the most common skin tumor [12, 13], on WSIs via an artificial neural network (ANN) using MIL with an "attention" classifier that efficiently differentiates tumors and healthy skin on a slide (=bag) level. As there are no data on the differences in human and machine attention in dermatopathology, we closely studied the regions of the images that formed the basis for the predictions of the ANN and compared those with the diagnosis-relevant regions of pathologists using eye-tracking techniques.

Materials and methods

Data set

Sections of BCCs and normal skin were stained with hematoxylin and eosin (H&E, $n = 820$ slides) for routine diagnoses. H&E-stained images were scanned with Aperio scanners (Leica Biosystems Division of Leica Microsystems Inc., Buffalo Grove, USA) at maximum available resolution (2 pixels per micron). Images were retrospectively collected at the Kepler University Hospital and the Medical University of Vienna for analysis by machine learning methods, with consent by ethics votes number 1119/2018 (Ethics committee of the Federal State Upper Austria) and 2085/2018 (Ethics committee of the Medical University of Vienna), respectively. The images were collected, including metadata: diagnosis of the lesion, age of the patient, gender, and a pseudonymous patient identifier (to avoid lesions from the same patient ending up in both data sets (training or test set)). A total of 601 of the WSIs show BCCs, and 219 show only normal skin. The samples were categorized (BCC or non-BCC) independently by two board-certified pathologists. This set of 820 images was randomly split into 132 (16%) test images and 688 (84%) training images. Twenty percent of the training set was used for validation during hyperparameter tuning. The median size of the WSIs was 56,896 × 26,198 pixels, with heights ranging from 6884 to 47,939 and widths ranging from 7360 to 99,568 pixels.

Neural networks

We implemented neural network architectures based on two approaches [14]. The first approach represents a baseline architecture. It is a convolutional neural network (CNN) using downsized WSIs (1024 × 1024 pixels with white padding). The CNN consists of five blocks of convolution–convolution–maxpooling and utilizes scaled exponential linear unit (SELU) activation functions (Supplementary Fig. 1) [15]. The architecture and the hyperparameters of this CNN were optimized on a validation set using manual hyperparameter tuning. The network was trained with stochastic gradient descent (SGD).

The second ANN architecture is composed of two independent ANNs, one feature constructor CNN, and one classification ANN (Supplementary Fig. 2). WSIs were split

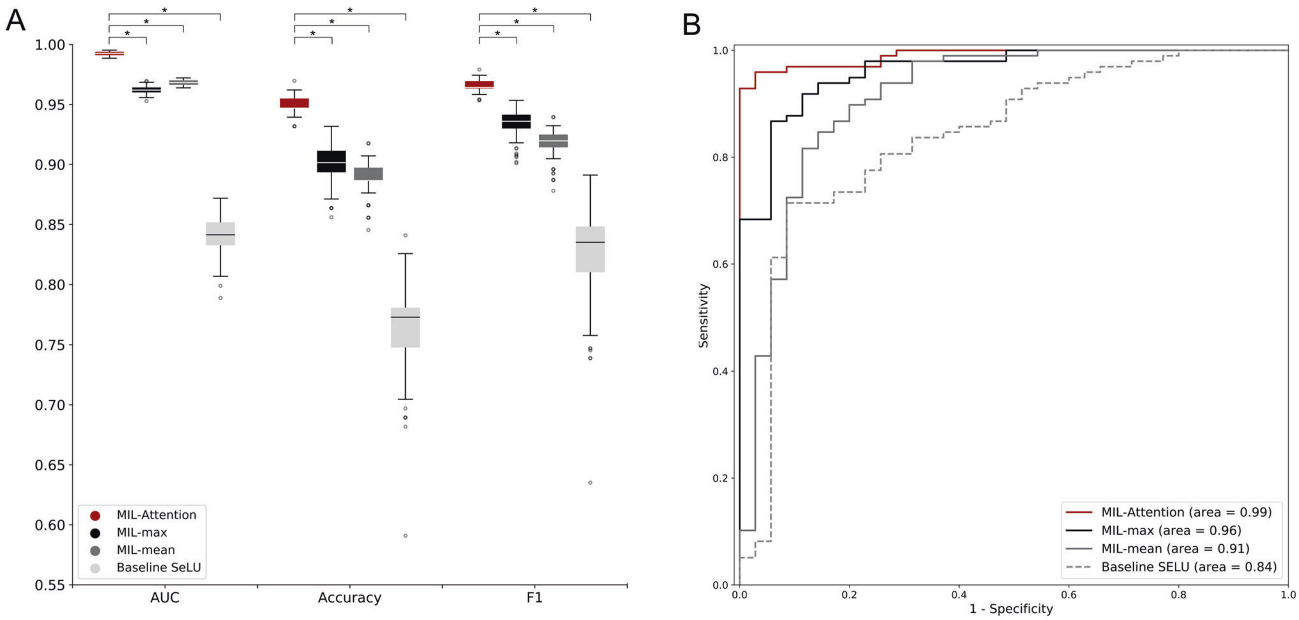

Fig. 1 Comparison of the metrics of four different MIL-based and baseline ANNs (MIL with attention (MIL-attention), MIL with maxpooling (MIL-max), MIL with mean pooling (MIL-mean), and the baseline SELU CNN (baseline SELU)). A Four different ANNs were tested on a test set of histologic section of basal cell carcinomas (BCC, $n = 97$) and normal skin (non-BCC, $n = 35$) to identify tumorous lesions. Subsequently, ANNs were compared with regard to area under curve (AUC), accuracy, and F1-score (measure of a test's accuracy that is not sensitive to imbalanced data sets) of 100 retrained ANNs. Indicated we see median (lines), interquartile range (bars), most extreme, non-outlier data points (whiskers), outliers (points). **B** Receiver operating characteristics (ROC) curves of (median performing out of 100 times retrained) MIL-based and baseline methods were calculated based on the test set of histologic section of basal cell carcinomas (BCC, $n = 97$) and normal skin (non-BCC, $n = 35$). $*p < 0.05$; MIL multiple instance learning, ROC receiver operating characteristic, SELU self-normalizing linear unit.

into tiles of size 224 × 224 pixels, as this is the image input size for the VGG11 NN architecture. Empty tiles were removed via pixel statistics (the average color intensity c_p of all pixels for each tile was calculated, the maximum $c_{\max}$ of each WSI was calculated, and all patches with c_p higher than $0.95 \times c_{\max}$ were removed and considered empty). Nonempty tiles in mini-batches of 32 were used as input for the feature constructor CNN. Each tile was normalized to 0 mean and unit variance. We used a VGG11 network pretrained on ImageNet [16] as the feature constructor CNN. The softmax function was removed, and the 1000-dimensional output of each tile was saved as "representation." The representations of all tiles from a WSI were used as mini-batch input for the classification ANN, which is based on MIL. The classification ANN was either (a) a mean of the resulting network predictions, (b) a maximum of the resulting network predictions, or (c) an attention classifier according to Ilse et al. [17]. The classification ANNs were trained with SGD. Hyperparameters were adapted using a manual hyperparameter search.

Eye tracking

Five "BCC" and four "non-BCC" cases were randomly selected from the test set for the eye-tracking study. WSIs and two magnifications were shown to four board-certified general pathologists. Eye traces were recorded using an iView X™ RED Laptop System (60/120 Hz) (SensoMotoric Instruments (SMI) GmbH, Germany) and analyzed using Experiment Suite 360° Professional (SMI GmbH, Germany) and Python 3.4.

Analysis of results

The results were analyzed using Python 3.4. The Jaccard similarity score was calculated using the package sklearn (version 0.21.2). Dice distances were calculated using the SciPy (version 1.2.1) package. For the Jaccard and Dice indices, discrete (0/1) values were used, i.e., a pixel was set to 1 if the pathologist looked at it for at least 7 ms; otherwise, it was set to 0. For computer attention, a pixel was set to 1 if it was higher than the mean value of all nonempty (preselected) tiles; otherwise, it was set to 0.

Statistics

We assessed the statistical significance of our results using hypothesis testing, with retraining the networks 100 times. Means and standard deviations of accuracy, F1-score (nonsensitive to unbalanced data sets) and AUC (area

Table 1 Metrics of different ANN methods (MIL with attention (MIL-attention), MIL with maxpooling (MIL-max), MIL with mean pooling (MIL-mean), and the baseline SELU CNN (baseline SELU)).

Data type	Method	Accuracy	F1-score	AUC	Sensitivity	Specificity
tiles	**MIL-attention**	**0.95** (0.94–0.96)	**0.97** (0.96–0.97)	**0.99** (0.99–0.99)	**0.96** (0.95–0.97)	**0.93** (0.90–0.95)
tiles	MIL-max	0.90 (0.89–0.92)	0.93 (0.92–0.94)	0.96 (0.96–0.96)	0.94 (0.91–0.98)	0.78 (0.69–0.87)
tiles	MIL-mean	0.88 (0.87–0.89)	0.92 (0.91–0.92)	0.91 (0.91–0.92)	0.93 (0.92–0.95)	0.72 (0.69–0.75)
downscaled WSIs	baseline SELU	0.76 (0.72–0.80)	0.83 (0.78–0.87)	0.84 (0.82–0.86)	0.78 (0.68–0.87)	0.73 (0.57–0.89)

AUC area under the ROC curve, *CNN* convolutional neural network, *MIL* multiple instance learning, *WSI* whole-slide image.

Bold values indicate name and metrics of the proposed attention-ANN method.

under the curve) of the ROC (receiver operating characteristic) curve were calculated using the results of these 100 retrained networks. Metrics were calculated as follows (TN = true negative, TP = true positive, FN = false negative, FP = false positive):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{all samples}},$$

$$\text{F1 score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}.$$

The significance of the metrics was calculated using a two-tailed Wilcoxon signed-rank test. The significance of the similarity metrics (Jaccard and Dice scores) was calculated using a two-tailed independent sample t-test. The results were considered statistically significant at p values < 0.05. Correlations between two variables were calculated using linear regression.

Results

Artificial neural networks accurately differentiate basal cell carcinomas from normal skin in histological sections

Swift automated analysis for digital pathology is a challenge because it requires the processing of large data sets. To reliably and quickly process and classify downscaled WSIs (1024 × 1024 pixels) of BCCs ($n = 601$) or normal skin (non-BCC, $n = 219$), we established a CNN with SELU activation functions (Supplementary Fig. 1). This baseline method offered the advantage of using only ~0.1% of all available information in terms of evaluated pixels but resulted in a mean accuracy of 0.753 ± 0.053 (SD) for the classification between BCC vs. non-BCC (Fig. 1A and Table 1). To increase accuracy, we tiled WSIs into squares of small resolution, which were then processed by ANN methods. After comparison of four different ANN methods, we proved that MIL with attention-based pooling significantly outperforms MIL with maxpooling, MIL with mean pooling, and the baseline CNN with respect to the area under the ROC curve (AUC), F1-score, and accuracy (Fig. 1 and Table 1; architecture: Supplementary Fig. 2) [14, 17]. In the same WSI collection, MIL with attention-based pooling identified BCC regions with a much higher accuracy of 0.950 ± 0.008 (SD; Fig. 1A). The robustness of our ANN method was tested by 100 times repeated retraining of each method, resulting in small ranges of metrics, e.g., range of AUC: 0.8% (detailed description in the Supplementary Results, Fig. 1A, B, and Supplementary Fig. 3).

Table 2 represents a summary of WSIs that were misclassified by at least 1 of 100 retrained attention-ANNs. BCCs were mainly misclassified due to small parts of BCC specimen on the image. All the misclassified non-BCC images showed at least one of the following characteristics: (1) solar elastosis, (2) inflammation, (3) scar, (4) fibrosis, (5) high vascularization. These features might serve as indicators for nearby neoplasms (e.g., the probability of nonmelanoma skin cancers rises in the proximity of solar elastosis; inflammation is commonly close to (particularly ulcerated) BCCs; scars can imitate the sclerosing tissue around infiltrative growing tumors). On the other hand, two board-certified pathologists analyzed the dermal structures independently and were not able to find any direct indicators for malignancy in these WSIs.

In addition to BCC tumor samples, the collection of non-BCC samples (healthy skin) used in this study consisted of uninvolved skin from surgical excisions in proximity to various skin neoplasms (e.g., dog ears and tumor-free resection edges) and scars (from re-excision surgeries) of BCC, squamous cell carcinomas (SCC) and melanoma. To check whether spatial proximity to any of those skin tumors accounted for the classification bias of the non-BCC samples, we related the number of non-BCC samples neighboring BCC, SCC, or melanoma to the number of false and correctly classified samples. Although not statistically significant, we observed the trend that tumor-free non-BCC samples obtained from the skin in proximity to BCCs were more often classified as BCCs than any other group

Table 2 Details of misclassified images by the attention-ANN.

	# of networks misclassifying the image	True diagnosis	Mean error[a]	Detailed findings/diagnosis upon reevaluation[b]	Non-BCC was excised close to	Histologic description of non-BCC dermis
Case01	1	Non-BCC	0.109	Solar lentigo with many cells and vessels	BCC	Solar elastosis, rich in cells, highly vascularized
Case02	7	Non-BCC	0.344	Hair follicles, solar lentigo	BCC	Solar elastosis
Case03	9	Non-BCC	0.304	No BCC	mel	Inflammation, fibrosis
Case04	20	BCC	0.311	BCC buds		
Case05	24	BCC	0.407	Nodular BCC		
Case06	48	Non-BCC	0.492	Solar lentigo/seborrheic keratosis	BCC	Low-grade inflammation
Case07	67	Non-BCC	0.520	Inflamed hair follicle	BCC	Solar elastosis, inflammation, fibrosis
Case08	86	Non-BCC	0.681	Scar	mel	Scar
Case09	97	BCC	0.889	Very small nodular BCC, probably a floater		
Case10	99	BCC	0.753	Small multifocal-superficial BCC, pale staining		
Case11	99	Non-BCC	0.721	Scar, inflamed hair shafts	BCC	Scar, inflammation, solar elastosis
Case12	100	BCC	0.961	Multifocal-superficial BCC		

The table shows the true diagnosis of the histological section of misclassified images and the mean error[a]. Furthermore, it displays a detailed description of the findings/diagnosis upon reevaluation[b] of the histological sections, indicates the tumor nearby to non-BCC sections, and shows the histologic description of the dermis of false classified non-BCC WSIs.

BCC basal cell carcinoma, *mel* melanoma.

[a] Mean error represents the mean prediction error of the 100 times retrained ANN method.

[b] Reevaluation was independently performed by a board-certified pathologist and a board-certified dermatopathologist.

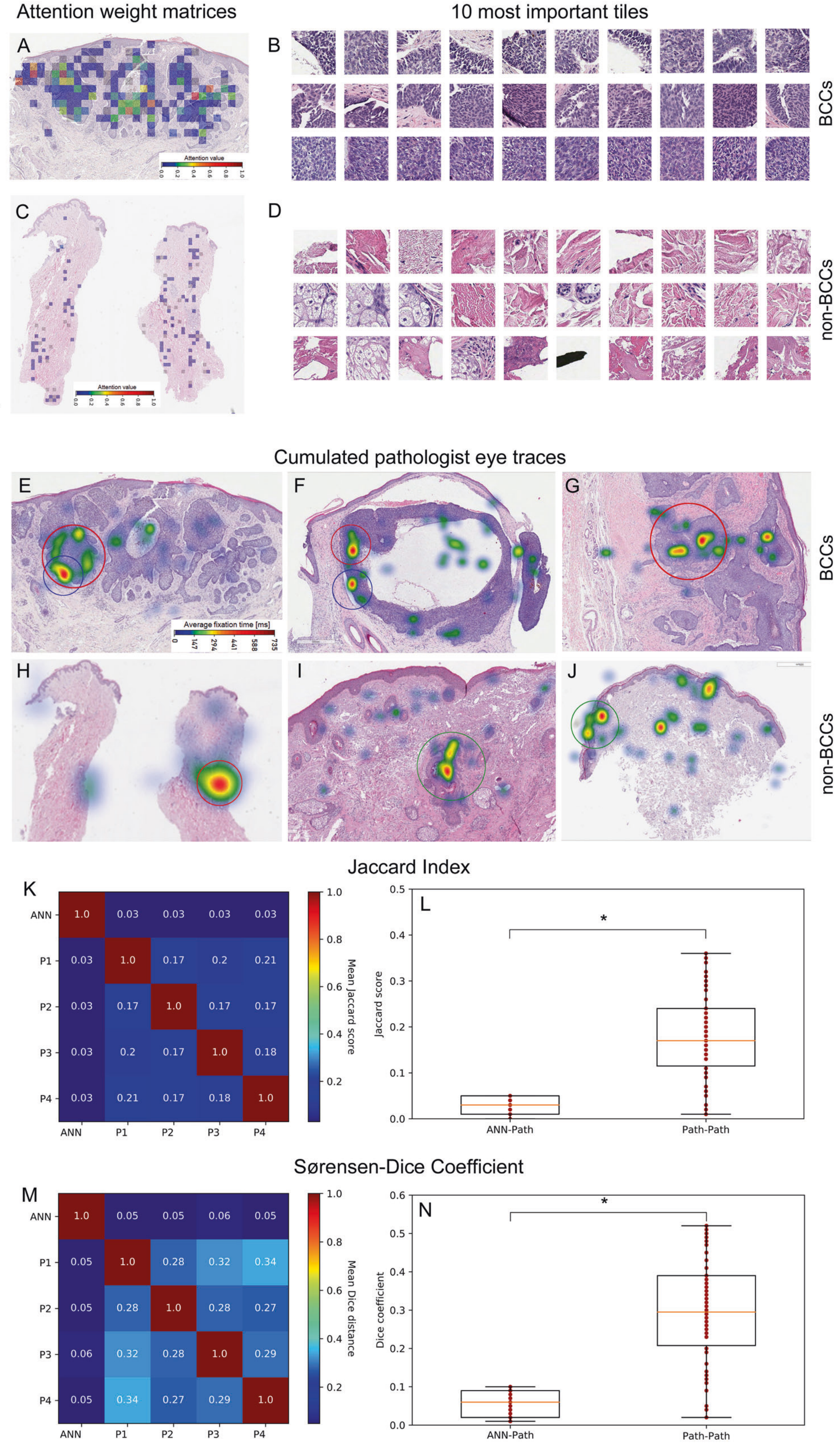
Attention weight matrices
10 most important tiles
A
Attention value
0.0
0.2
0.4
0.6
0.8
1.0
B
BCCs
C
Attention value
0.0
0.2
0.4
0.6
0.8
1.0
D
non-BCCs
Cumulated pathologist eye traces
E
Average fixation time [ms]
0
147
294
441
588
735
F
G
BCCs
H
I
J
non-BCCs
Jaccard Index
K
ANN 1.0 0.03 0.03 0.03 0.03
P1 0.03 1.0 0.17 0.2 0.21
P2 0.03 0.17 1.0 0.17 0.17
P3 0.03 0.2 0.17 1.0 0.18
P4 0.03 0.21 0.17 0.18 1.0
ANN P1 P2 P3 P4
Mean Jaccard score
1.0
0.8
0.6
0.4
0.2
L
*
Jaccard score
0.5
0.4
0.3
0.2
0.1
0.0
ANN-Path
Path-Path
Sørensen-Dice Coefficient
M
ANN 1.0 0.05 0.05 0.06 0.05
P1 0.05 1.0 0.28 0.32 0.34
P2 0.05 0.28 1.0 0.28 0.27
P3 0.06 0.32 0.28 1.0 0.29
P4 0.05 0.34 0.27 0.29 1.0
ANN P1 P2 P3 P4
Mean Dice distance
1.0
0.8
0.6
0.4
0.2
N
*
Dice coefficient
0.6
0.5
0.4
0.3
0.2
0.1
0.0
ANN-Path
Path-Path

Fig. 2 Regions of interest according to attention weight matrices (ANN) and eye tracking (pathologists). A Representative image of the attention weight matrix of a BCC section. **B** Representative images of the ten most important tiles for the MIL-attention method of three BCC WSIs. **C** Representative image of the attention weight matrix of a non-BCC sample. **D** Representative images of the ten most important tiles for the MIL-attention method of three non-BCC WSIs. **E–G** Representative images of the cumulated eye traces of four board-certified pathologists on three BCC samples. **H–J** Representative images of the cumulated eye traces of four board-certified pathologists on three non-BCC samples. **E–J** Blue circles represent the artefactual retraction gaps. Red circles highlight particular focus points of eye traces. Green circles highlight epidermis, glandular structures, and hair follicles. **K–N** Similarity measures between a single pathologist's eye trace and the attention weight matrix of a median performing ANN-attention model. **K** Heat map of Jaccard scores between pathologists and the ANN and pathologists to each other. **L** Scatter and bar chart of Jaccard scores between pathologists (Path-Path) and the ANN and pathologists to each other (ANN-Path; one scatter represents "path vs. path" in one image, $p = 5.81 \times 10^{-15}$). **M** Heat map of the Sørensen–Dice coefficient between pathologists and the ANN and pathologists to each other. **N** Scatter and bar chart of the Sørensen–Dice coefficient between pathologists (Path-Path) and the ANN and pathologists to each other (ANN-Path; one scatter represents "path vs. path" in one image, $p = 1.10 \times 10^{-16}$). P1–P4 pathologist 1 to pathologist 4, ANN artificial neural network, BCC basal cell carcinoma, WSI whole-slide image.

(Supplementary Fig. 4; not significant using binomial testing). This allowed us to hypothesize that ANNs consider stromal changes important during the recognition of BCCs in addition to direct tumor detection.

ANNs and pathologists identify basal cell carcinomas based on different recognition patterns

Interpretability and the process of decision-making of AI in comparison to humans is a key question in modern health care. In addition to classification prediction, the MIL-attention method outputs an attention weight matrix, which represents importance values for each tile. To address the issue of interpretability, we identified the local tiles (areas) of sections that are important for the classification of the network by analyzing its corresponding attention weights.

Analysis of the ROI according to attention weight matrices documented that ANN "scans" through the whole section for diagnosis (Fig. 2A, C). Unexpectedly, detailed close-ups of the most important tiles revealed a focus of the ANN not only on BCC tumor cells and tumor stroma (e.g., cytoplasm, nuclei, and basophilic staining) but also on adnexal (including sebaceous and vascular structures, and connective tissue in the areas of surrounding normal skin (Fig. 2B, D)).

To address the question of whether all areas relevant for BCC diagnosis by the ANN are also part of the BCC recognition pattern recognized by expert pathologists, we conducted an eye-tracking study with four board-certified pathologists who blindly diagnosed the same slides that were presented to the ANN. Cumulated eye tracing data of the four blinded pathologists unambiguously demonstrated that all four pathologists unconsciously focused on similar structures before making a diagnostic decision for BCCs (Fig. 2E–J and Supplementary Fig. 5).

Upon the qualitative data review, we identified three main differences between pathologists and neural networks. First, pathologists preferably focused on individual areas of the tumor (examples in Fig. 2E–J, red circles), while the ANN included the entire tumorous section equally in its decision (e.g., Fig. 2A, C). Second, the pathologists' attention concentrated on the artefactual retraction gap for diagnosis (e.g., Fig. 2E, F, blue circles), while the network does not attach as much importance to it (e.g., Fig. 2A, B). Third, in non-tumor sections, pathologists focus mainly on the epidermis, glands, and hair follicles (examples in Fig. 2H–J, green circles), while the ANN "scans" through the whole section and pays additional attention to connective tissue patterns (e.g., Fig. 2C, D). To quantify the difference in pattern recognition between the ANN method and pathologists, we applied the Jaccard index and the Sørensen–Dice coefficient, two commonly used statistics for the measurement of similarities between sample sets. Both metrics proved that the similarity of the interpersonal eye traces of pathologists is significantly higher than the similarity between the pathologists and the attention weight matrix of the ANN method (Fig. 2K–N, $p < 10^{-4}$). These results demonstrate that pathologists are trained to focus on specific structures with higher contrast and color intensity for diagnosing BCCs, while the ANN bases its decision on all types of regions.

Discussion

Due to technical progress, whole-slide imaging has become a standard method in (digital) pathology. It enables a geographically independent, collaborative diagnosis of difficult cases. Recently, it has been shown that the analysis of WSI is comparable to classical microscopy in terms of diagnostic accuracy [18, 19]. These developments have greatly advanced computer-aided diagnosis. As an example, computer-aided diagnosis is already used for the assessment of several receptors in breast cancer and Ki67 in carcinoid tumors [20, 21]. In the present study, we implemented an ANN that predicts whether histological WSIs contain BCCs or normal skin with high accuracy, sensitivity, and specificity. Compared to other methods applied in this field, we use "attention" as a classifier, which is an easy method that implicitly outputs priorities of different regions. For this method, no detailed tumor region annotation is required. We detected important tiles and structures relevant for the diagnosis and subsequently identified histologic structures that might be important for the diagnosis of BCCs. Eventually, we qualitatively compared the differences in diagnosis-relevant regions between the ANN and pathologists.

Machine learning is an emerging field in medicine, e.g., for the diagnosis of dermatoscopic photographs, radiology images, skin lesion photographs, and unprocessed clinical photographs [22–25]. In addition, the number of promising methods for computer-aided diagnosis in digital pathology has increased. Recent studies have shown classification accuracies higher than 90% for the detection of different classes of skin lesions [26] and tumor/metastasis predictions [5] on WSIs using ANNs. Campanella et al. recently demonstrated that ANN architectures are capable of clinical-grade prediction of WSIs including various different diagnoses. The authors analyzed BCCs, among others, resulting in 100% diagnostic sensitivity with an acceptable false positive rate. Based on their data, they propose to remove 75% of the slides from the workload of a pathologist without loss of sensitivity [5]. While these results are intriguing, it should be noted that the attribution of medico-legal responsibility for errors occurring in AI-assisted workflows is not clearly regulated up to date. In our study we mainly focused on the interpretability of computer-aided diagnosis.

Our study addresses the challenge of evaluating real-life gigapixel data via machine learning and provides interpretable predictions. In this context, our project differs from others in this field, as it does not utilize publicly available or downsized data. Instead, it employs real-life data, retrospectively, collected from patients of two study centers. Our approach allows using weakly labeled input data and reduces the need for handcrafted annotations, such as segmenting the tumor area, to a minimum. Through this approach, we also bypass subjective local annotation features that may contain mistakes and are time-intensive for collecting from physicians. The methods of the current study represent a proof-of-concept that ANNs can deal with this kind of data efficiently.

There are multiple histomorphologic variants of BCCs that share similar histopathologic features with almost all variants. BCCs typically comprise islands or nests of basaloid cells surrounded by loose fibromucinous stroma, with a characteristic peripheral palisading of cells and a haphazard arrangement of cells in the center [27, 28]. Artefactual retraction gaps between the tumor and stroma, apoptotic cells, amyloid deposits in the stroma and a variable inflammatory infiltrate are often associated with BCCs [27, 28]. One key machine learning question in health care is its interpretability and the process of decision-making in comparison to humans, where the benefit of human-computer collaboration differs significantly between used methodologies [29]. Analyzing the different ROIs between pathologists and our ANN method, we identified several differences in attention patterns. We observed that neural networks distribute their attention over larger tissue areas, whereas pathologists focus on specific structures (e.g., peripheral palisading of tumor cells and retraction gap). In addition, the ANN integrates the connective tissue in its decision-making, which is different from the recorded eye traces of pathologists. In this context, the ANN method predicted normal skin that was close to BCCs more likely as “BCC” than skin close to melanomas or SCCs (Supplementary Fig. 4). Moreover, the tissue of the misclassified non-BCC images was interspersed with features that are also commonly seen in proximity to BCC (e.g., inflammation and solar elastosis; Table 2). Our results indicate that the tumor microenvironment of BCCs is also important for BCC diagnosis, in line with previous histopathological findings [30]. Consequently, we tested if the network is able to classify stroma of BCC WSIs as “BCC.” The ANN predicted the images to be “BCC” in 0–33% of cases (data not shown). Further studies with higher training sample numbers will be needed to address this issue in more detail.

Distinguishing BCCs with superficial growth pattern (superficial) to those with growth to the reticular dermis and deeper within the test set, we found that 23.3% (21/90) of test cases were superficial, whereas they accounted for 80% (4/5) of misclassified BCCs (by at least one of 100 retrained ANNs). Our findings demonstrate that performance metrics may differ significantly for tumor subtypes and should be reflected in reporting of future studies.

The attention patterns of pathologists are based upon learned behaviors for distinguishing a great number of different tumors, including various cutaneous cancer entities. In contrast, our ANN was only trained to separate BCCs from non-tumor skin. This difference may explain the distinct attention patterns of pathologists (e.g., use of higher magnification and focus on retraction artifact), which are different compared to the ANN. Future studies with ANNs that have to learn to distinguish multiple tumor entities are needed to better understand the different interpretation of the attention-based data between pathologists and ANNs.

Microscopically controlled surgery is considered the gold standard for the treatment of certain skin cancers [31]. In this context, skin sections of microscopically controlled surgery represent a significant daily workload of pathologists. Consequently, automated systems that prescreen WSIs for cancerous tissue might be time saving in daily practice. In this study, we developed an ANN that can detect BCCs in skin sections with high accuracy. However, there are still several limitations to this technique before it can be safely applied in daily clinical routines (e.g., pace of imaging procession and sample size, legal aspects).

Our results demonstrate that ANNs diagnose BCCs in partially different ways compared to human professionals, although the outcome—the correct histologic diagnosis—is comparable. As the interpretability of ANNs is the key for future applications, our data are a significant contribution to this rapidly emerging field.

Acknowledgements We want to thank Harald Kittler for input at the commencement of the study and Giuliana Petronio for invaluable help in managing the dermatologic image database at the Medical University of Vienna. Furthermore, we want to thank Gudrun Lang for excellent technical assistance. This research was supported by ERC grant number 75931 (GEL-SYS), REA grant number 828984 (LIONHEARTED), the Promedica Stiftung (1406/M and 1412/M), the Swiss Cancer Research Foundation (KFS-4243-08-2017), the Clinical Research Priority Program (CRPP) of the University of Zurich, the Swiss National Science Foundation (PMPDP3_151326), and the European Academy of Dermatology and Venereology (PPRC-2019-20).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Griffin J, Treanor D. Digital pathology in clinical use: where are we now and what is holding us back? Histopathology. 2017;70:134–45.
2. Parwani AV. Next generation diagnostic pathology: use of digital pathology and artificial intelligence tools to augment a pathological diagnosis. Diagn Pathol. 2019;14:19–21.
3. Arevalo J, Cruz-roa A, Arias V, Romero E, González FA. An unsupervised feature learning framework for basal cell carcinoma image analysis. Artif Intell Med. 2015;64:131–45.
4. Cruz-Roa A, Gilmore H, Basavanhally A, Feldman M, Ganesan S, Shih NNC, et al. Accurate and reproducible invasive breast cancer detection in whole-slide images: a deep learning approach for quantifying tumor extent. Sci Rep. 2017;7:1–14.
5. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck V, Silva K, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25:1301–9.
6. Dou Q, Chen H, Qin J, Heng P-A. Automatic lesion detection with three-dimensional convolutional neural networks. In: Feng DD, editor. BioMedical information technology. 2nd ed. Oxford: Elsevier Inc.; 2020. p. 265–93.
7. Sudharshan PJ, Petitjean C, Spanhol F, Oliveira LE, Heutte L, Honeine P. Multiple instance learning for histopathological breast cancer image classification. Expert Syst Appl. 2019;117:103–11.
8. Mercan C, Mercan E, Aksoy S, Shapiro LG, Weaver DL, Elmore JG. Multi-instance multi-label learning for whole slide breast histopathology. In: SPIE proceedings (9791) of medical imaging 2016: Digital Pathology. International Society for Optics and Photonics. San Diego, CA, USA; 2016. p. 979108.
9. Hekler A, Utikal JS, Enk AH, Solass W, Schmitt M, Klode J, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur J Cancer. 2019;118:91–6.
10. Das A, Agrawal H, Lawrence Zitnick C, Parikh D, Batra D. Human attention in visual question answering: do humans and deep networks look at the same regions? In: Conference on empirical methods in natural language processing. Association for Computer Linguistics. Austin, TX, USA; 2016. p. 932–7.
11. Lai Q, Wang W, Khan S, Shen J, Sun H, Shao L. Human vs machine attention in neural networks: a comparative study. 2019. http://arxiv.org/abs/1906.08764.
12. Chu DH. Development and structure of skin. In: Freedberg IM, Eison AZ, Wolff K, et al., editors. Fitzpatrick's dermatology in general medicine. 8th ed. New York: McGraw-Hill Professional; 2012.
13. Quevedo WCJ, Holstein TJ. General biology of mammalian pigmantation. In: Nordlund JJ, Boissy RE, Hearing VJ, et al., editors. The pigmentary system—physiology and pathophysiology. Carlton: Blackwell Publishing Ltd; 2006. p. 63–90.
14. Kimeswenger S, Rumetshofer E, Hofmarcher M, Tschandl P, Kittler H, Hochreiter S, et al. Detecting cutaneous basal cell carcinomas in ultra-high resolution and weakly labelled histopathological images. In: Machine learning for health workshop. NeurIPS; 2019. p. 1–6.
15. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-normalizing neural networks. In: Guyon I, Luxburg UV, Bengio S, et al., editors. Advances in neural information processing systems 30. NIPS'17; 2017. p. 972–81.
16. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations. ICLR. San Diego, CA, USA; 2015. p. 1–14.
17. Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. In: Proceedings of the 35th international conference on machine learning. ICML. Stockholm; 2018.
18. Mukhopadhyay S, Feldman MD, Abels E, Ashfaq R, Beltaifa S, Cacciabeve NG, et al. Whole slide imaging versus microscopy for primary diagnosis in surgical pathology—a multicenter blinded randomized noninferiority study of 1992 cases (Pivotal Study). Am J Surg Pathol. 2018;42:39–52.
19. Hanna MG, Reuter VE, Hameed MR, Tan LK, Chiang S, Sigel C, et al. Whole slide imaging equivalency and efficiency study: experience at a large academic center. Mod Pathol. 2019;32:916–28.
20. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol. 2019;20:e253–61.
21. Dimitriou N, Arandjelović O, Caie PD. Deep learning for whole slide image analysis: an overview. Front Med. 2019;6:1–7.
22. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
23. Han SS, Moon IJ, Lim W, Suh IS, Lee SY, Na JI, et al. Keratinocytic skin cancer detection on the face using region-based convolutional neural network. JAMA Dermatol. 2020;156:29–37.
24. Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol. 2019;2045:1–10.
25. Montagnon E, Cerny M, Cadrin-Chênevert A, Hamilton V, Derennes T, Ilinca A, et al. Deep learning workflow in radiology: a primer. Insights Imaging. 2020;11:1–15.
26. Ianni JD, Soans RE, Sankarapandian S, Chamarthi RV, Ayyagari D, Olsen TG, et al. Tailored for real-world: a whole slide image classification system validated on uncurated multi-site data emulating the prospective pathology workload. Sci Rep. 2020;10:1–12.
27. Weedon D, Marks R, Kao GF, Hartwood CA. World Health Organization classification of tumours. Pathology and genetics skin tumours. Lyon: IARC Press; 2006.
28. Saldanha P, Shanthala P, Upadhaya K. Cutaneous basal cell carcinoma: a morphological spectrum. Arch Med Heal Sci. 2015;3:24–8.
29. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020;26:1229–34.
30. Bertheim U, Hofer PÅ, Engström-Laurent A, Hellström S. The stromal reaction in basal cell carcinomas. A prerequisite for tumour progression and treatment strategy. Br J Plast Surg. 2004;57:429–39.
31. Mohs FE. Chemosurgery: a microscopically controlled methold of cancer excision. Arch Surg. 1941;42:279–95.