Diagnostic Pathology

RESEARCH Open Access

Explainable convolutional neural networks for assessing head and neck cancer histopathology

Marion Dörrich[1], Markus Hecht[2,3], Rainer Fietkau[3,4], Arndt Hartmann[4,5,6], Heinrich Iro[7], Antoniu-Oreste Gostian[4,6,7], Markus Eckstein[4,5,6†] and Andreas M. Kist[1*†]

Abstract

Purpose Although neural networks have shown remarkable performance in medical image analysis, their translation into clinical practice remains difficult due to their lack of interpretability. An emerging field that addresses this problem is Explainable AI.

Methods Here, we aimed to investigate the ability of Convolutional Neural Networks (CNNs) to classify head and neck cancer histopathology. To this end, we manually annotated 101 histopathological slides of locally advanced head and neck squamous cell carcinoma. We trained a CNN to classify tumor and non-tumor tissue, and another CNN to semantically segment four classes - tumor, non-tumor, non-specified tissue, and background. We applied Explainable AI techniques, namely Grad-CAM and HR-CAM, to both networks and explored important features that contributed to their decisions.

Results The classification network achieved an accuracy of 89.9% on previously unseen data. Our segmentation network achieved a class-averaged Intersection over Union score of 0.690, and 0.782 for tumor tissue in particular. Explainable AI methods demonstrated that both networks rely on features agreeing with the pathologist's expert opinion.

Conclusion Our work suggests that CNNs can predict head and neck cancer with high accuracy. Especially if accompanied by visual explanations, CNNs seem promising for assisting pathologists in the assessment of cancer sections.

Keywords Explainable AI, Histopathology, Head and neck cancer, Semantic segmentation, Classification

[†] Markus Eckstein and Andreas M. Kist contributed equally as senior authors.

*Correspondence:
Andreas M. Kist
andreas.kist@fau.de
[1] Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen 91052, Germany
[2] Department of Radiotherapy and Radiation Oncology, Saarland University Medical Center, Homburg/Saar 66421, Germany
[3] Department of Radiation Oncology, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen 91054, Germany
[4] Comprehensive Cancer Center EMN, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen 91054, Germany
[5] Institute of Pathology, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen 91054, Germany
[6] Bavarian Cancer Research Center (BZKF), Bavaria, Germany
[7] Department of Otolaryngology - Head and Neck Surgery, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen 91054, Germany

Introduction

Head and Neck Squamous Cell Carcinoma (HNSCC) is a malignancy that can develop in several regions, such as the oral cavity, pharynx, or larynx [1]. Worldwide, HNSCC was the seventh most common cancer in 2020 [2]. HNSCC patients have a poor prognosis, and their disease is often diagnosed in an advanced stage [1]. Although the 5-year survival has improved over the last decades [1], it is still very low, ranging between 25% and 60% [3].

HNSCC is diagnosed by pathologists who assess tissue sections and provide important information for treatment choice and prognosis. Thin tissue slices are stained and evaluated using light microscopes, but recently can also be digitally analyzed as Whole Slide Images (WSIs). As WSIs are high-quality imaging data, many computer vision algorithms are being developed to reduce the workload of pathologists and improve the accuracy of the diagnosis.

Artificial Intelligence (AI) has seen a lot of attention recently, also in the medical field. Especially Convolutional Neural Networks (CNNs) have shown remarkable performance in the analysis of medical images, including WSIs [4]. However, translating AI systems into clinical practice remains difficult due to their black-box nature. Algorithms applied in the diagnosis of cancer need to be very reliable and trustworthy. Explainable AI methods can be used to tackle this problem and improve the transparency of neural networks, for example by offering visual explanations of predictions [5]. Explainable AI techniques help both developers and physicians to better understand AI algorithms, their abilities, and their limitations [6].

We aimed to investigate the ability of CNNs to classify and semantically segment head and neck cancer tissue. To this end, we manually annotated tissue in the WSIs in two distinct classes, namely tumor and non-tumor, and applied state-of-the-art CNNs. Additionally, we aimed to explore which features were responsible for both networks' decisions, using the two Explainable AI methods Grad-CAM [7] and HR-CAM [8].

Methods

Data source

The histopathological slices were collected in the context of the CheckRad-CD8 trial [9, 10]. In this trial, a cancer treatment that consists of induction therapy followed by radioimmunotherapy was developed [10]. Patients with locally advanced HNSCC of the oral cavity, oropharynx, hypopharynx, or larynx were selected in eight clinical centers in Germany [10]. Their diagnosis was confirmed by a biopsy of the primary tumor. The patients first received induction chemoimmunotherapy with double checkpoint blockade. Based on their response, determined by the increase of the intratumoral CD8+ cells, patients were selected for subsequent radioimmunotherapy [10]. The dataset used in this work consists of tissue sections resulting from the pre-therapeutic biopsies of 101 patients. One slide per patient was used. The slides were stained using hematoxylin and eosin (HE) and digitized as WSIs. Out of 57 patients with oropharyngeal cancer, 30 were HPV-related as determined by p16 expression. Patient characteristics are shown in Fig. 1.

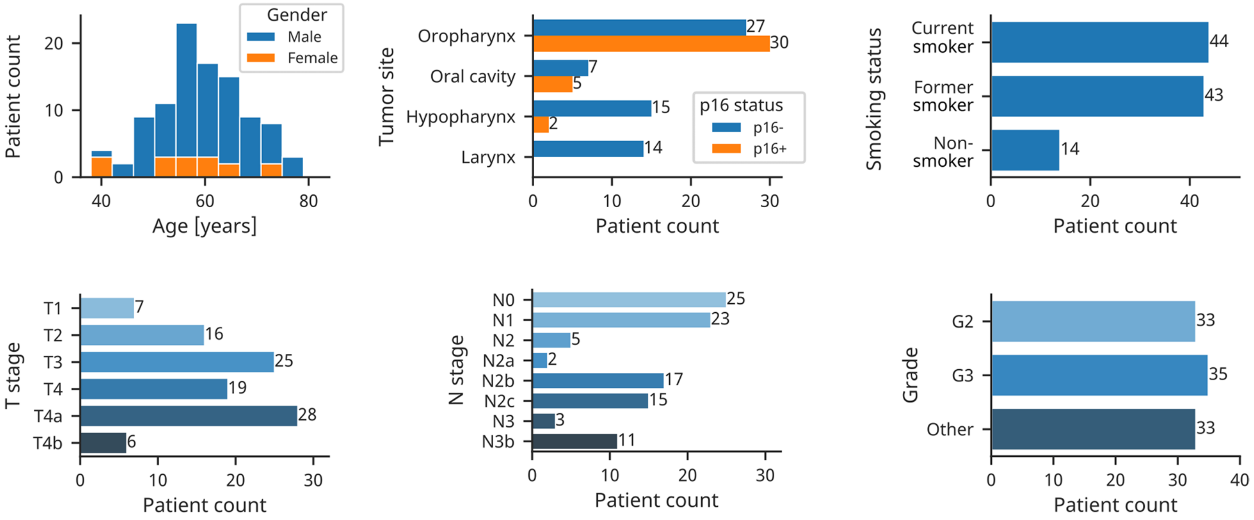

Fig. 1 Characteristics of 101 head and neck cancer patients from the CheckRad-CD8 trial. Tumor stages are given according to the UICC TNM eighth edition. Grade "Other" denotes patients with HPV-positive oropharyngeal tumor or missing tumor grade. TNM = tumor-node-metastasis

Data annotation

All 101 WSIs were manually annotated with supervision by an experienced pathologist. Using QuPath [11], an open-source software for whole-slide analysis, tissue regions were annotated in two distinct classes, namely tumor and non-tumor. The tumor class includes both tumor cells and surrounding tumor stroma. Tissues such as normal squamous epithelium, connective tissue, glands, muscle, and fat tissue were annotated as non-tumor. White background, damaged tissue, and large regions of blood or necrosis were not annotated. Furthermore, artifacts such as tissue folds were excluded from the annotation.

Data preprocessing

WSIs have very high resolutions and are commonly divided into small square tiles that can be fed to a CNN [12, 13]. The Python library PathML [14] was used to extract non-overlapping tiles with at least 30% annotated pixels. In most slides, more tumor than non-tumor tissue was present and some slides contained no non-tumor tissue. This resulted in a highly imbalanced class distribution. We decided to extract a maximum of 125 tumor and 500 non-tumor tiles from each slide, as depicted in Fig. 2. In this way, the majority class was undersampled and an overall balanced number of tiles per class was achieved [15, p. 221].

Along with each tile, the corresponding annotations were extracted as binary masks. These masks were further preprocessed for semantic segmentation, as shown in Fig. 2. Specifically, we added a background class for white background pixels and a fourth class for not annotated tissue, in the following referred to as class "other". The class "other" was added because some tiles contained tissue without any annotation, which should be separated from the background class. The resulting ground truth is a stack of four binary masks, where every pixel belongs to exactly one class. For the classification task, the class with the maximum pixel count was assigned to each tile.

Two important hyperparameters are the tile size, which is commonly between 10 and 250 μm in histopathology [14], and the input resolution, which affects the training speed and accuracy of CNNs. Tile size and resolution influence each other. For example, the impact of varying resolution on the accuracy is greater if images contain more complex information [16]. Thus, both were determined using a grid search with iterated 5-fold cross-validation [17, p. 136] and chosen based on the average validation accuracy of the classification network. In the grid search, the sizes 99.6, 149.4, and 199.1 μm and resolutions ranging from 64 to 512 pixels were considered.

Data augmentation has been shown to improve the generalization of CNNs trained on HE-stained histological images with stain variability [18]. Therefore, several transformations were randomly applied during training,

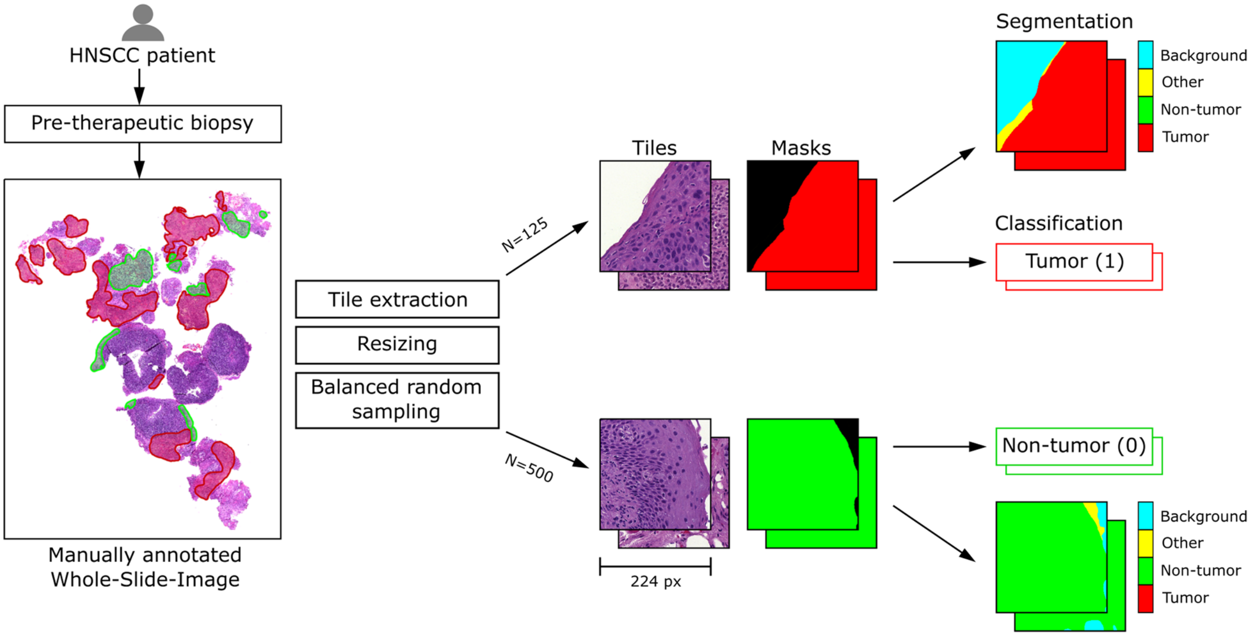

Fig. 2 Summary of our data pre-processing pipeline. Each pre-therapeutic WSI was manually annotated. Next, we extracted square tiles from the anntoated tumor (red) and non-tumor (green) regions. Based on the annotations, the ground truth masks for semantic segmentation and the class labels for classification were created

including rotation, mirroring, and variations in hue, saturation, brightness, and contrast. In the training of the classification network, blurring and additive gaussian noise were also applied. The Python library Albumentations [19] was used for transforming tiles and masks jointly. The pixel intensity was scaled to a range of 0 to 1 and then standardized by subtracting the mean value and dividing by the standard deviation of the training data [15, p. 126].

Deep neural networks

The classification network is an EfficientNet-B0 [20] pre-trained on ImageNet [21]. We modified this architecture by adding a global average pooling layer, a dense layer with 1024 neurons, Dropout [22], and another dense layer on top. The final dense layer contains a single neuron using a sigmoid activation function. The segmentation network is based on a U-Net architecture [23]. We modified the architecture by replacing the default encoder with EfficientNet-B0 pre-trained on ImageNet [21]. Specifically, the encoder is composed of an input layer and seven blocks of EfficientNet-B0. The decoder consists of five decoder blocks and the output layer, a 1x1 convolutional layer using a softmax activation. A decoder block applies upsampling, followed by two convolutions. Each decoder block is connected to an encoder block by a skip connection. The segmentation network is based on the U-Net architecture with EfficientNet-B0 backbone of the Segmentation Models library [24]. A similar architecture called Eff-Unet with EfficientNet-B7 encoder has been shown to outperform similar approaches [25].

Both networks were implemented and trained in Python using TensorFlow (version 2.8 with Keras API) [26]. The classification network was trained using Adam optimizer [27] to minimize the binary cross-entropy loss. The learning rate was set to 10^{-5} and the batch size to 128. The segmentation network was trained using Adam optimizer [27] with a learning rate of 10^{-4}. We chose to minimize the Jaccard loss function and set the batch size to 64. Additionally, we created two ensemble models. To this end, we converted the predictions of the segmentation network to tile-level predictions. This was implemented by treating the fraction of predicted tumor pixels as tumor probability. The first ensemble model simply averaged the predictions of both networks, which is also called voting. The second ensemble model was a logistic regression model, which was trained on the predictions for the test data using iterated 2-fold cross-validation.

Explainable AI

For establishing visual interpretability, we relied on two Explainable AI methods that are based on Class Activation Maps (CAMs) [5]. Both methods produce heatmaps, where patterns contributing most to a prediction are highlighted.

Gradient-weighted Class Activation Mapping (Grad-CAM) involves computing the gradient of the class score with respect to feature maps of the final convolutional layer [7]. These feature maps are weighted according to their importance for the predicted class score to produce a coarse localization map [7]. Grad-CAMs can also be created for segmentation networks by replacing the class score by a set of pixels in the output [28]. As recommended by Vinogradova et al., we obtained feature maps from the bottleneck layer [28].

The second method is High-Resolution Class Activation Mapping (HR-CAM) which aggregates feature maps from multiple layers to create a high-resolution localization map [8]. To compute HR-CAMs, the classifier of a frozen CNN is removed, and feature maps are obtained from several convolutional layers. These feature maps are fed to global average pooling and a top dense layer, which is trained to minimize a cross-entropy loss [8]. The heatmap is a weighted sum of the feature maps and the weights of the final dense layer. We obtained feature maps from several layers, as summarized in Table 1. The HR-CAMs for both the classification and segmentation network were created in the

Table 1 Names and output sizes of layers, from which feature maps were obtained. HR-CAM uses feature maps from several layers, whereas Grad-CAM only requires the last convolutional layer's output. The layer names correspond to the original layer names of EfficientNet-B0 [20]

Method	Classification CNN layers	Segmentation CNN layers	Output size [px]
HR-CAM	block_3a_expand_activation	block_3a_expand_activation	56×56
	block_4a_expand_activation	block_4a_expand_activation	28×28
	block_6a_expand_activation	block_6a_expand_activation	14×14
	block_7a_expand_activation	block_6d_expand_activation	7×7
Grad-CAM	top_activation	block_7a_expand_activation	7×7

same way. We re-trained both models to classify tiles containing at least 80% tumor or 80% non-tumor for 50 epochs.

Results

Dataset compilation

First, we determined the ideal settings for training CNNs. To this extent, square tiles were extracted from the WSIs at 51× magnification with a pixel size of 194 nm. Using our grid search approach (see Methods), the highest score was reached using a tile size of 199.1 µm (corresponding to 1024 original pixels) and resampled resolutions of 224 to 512 pixels. Therefore, we extracted tiles of size 199.1 µm and resized them to 224 pixels. The scores for different combinations of sizes and resolutions are shown in Additional file 1, Fig. S1. The 101 slides were split into three subsets. 70 slides were used for training, 10 for validation, and 21 for testing. In total, the dataset contained 20,195 tiles.

CNNs are suited to classify and segment head and neck cancer tissue

The networks were evaluated using a previously unseen test dataset of 21 patients. Figure 3 shows their performance in terms of receiver operating characteristic (ROC) curves and confusion matrices. For a better comparison of the two networks, the predictions of the segmentation network were converted to tile-level predictions (see Methods). The resulting ROC curves are shown in Fig. 3a, and the corresponding confusion matrices in Fig. 3d and e. We found that the segmentation network had higher sensitivity but was outperformed by the classification network regarding accuracy and area under the curve (AUC), as summarized in Table 2. The classification network achieved 89.9% and the segmentation network 85.9% accuracy on the test data.

The confusion matrix in Fig. 3f contains the original pixel-wise predictions of the segmentation network. It achieved a class-averaged Jaccard coefficient of 0.690, and 0.782 for the tumor class in particular. Figure 3f shows that only 33% pixels of class "other" were correctly classified.

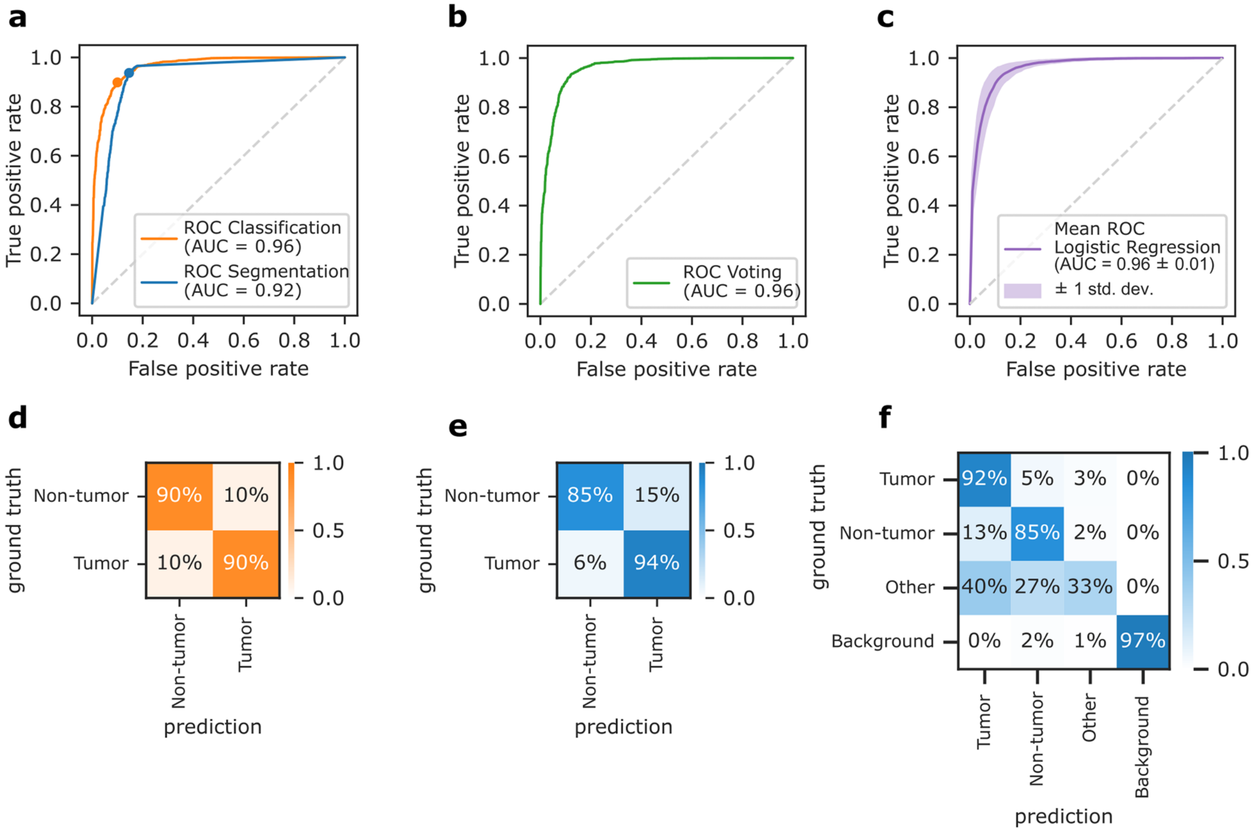

Fig. 3 Performance of the classification and segmentation network on test data. **a**-**c** Receiver operating characteristic (ROC) curves and area under the curve (AUC). **a** Comparison of both networks, where the segmentation network's pixel-level predictions were converted to tile-level predictions. **b** Model ensemble using voting. **c** Model ensemble using logistic regression. **d**-**f** Row-normalized confusion matrices. **d** Classification network. **e** Segmentation network with tile-level predictions. **f** Segmentation network with pixel-level predictions

Table 2 Quantitative performance evaluation of both networks and two model ensembles. The segmentation networks' predictions were converted to tile-level predictions in advance for better comparison. All performance metrics were computed for the full test dataset, except for ensemble logistic regression, where the mean values were obtained using iterated 2-fold cross-validation

	Accuracy	AUC	Sensitivity	Specificity
Classification network	89.9%	0.963	89.8%	90.0%
Segmentation network	85.9%	0.921	93.6%	85.4%
Ensemble averaging	87.1%	0.959	86.4%	87.8%
Ensemble logistic regression	89.7%	0.960	91.1%	90.0%

This class contains pixels that were not manually annotated. For example, it includes tissue edges due to imprecise annotations, or artifacts. However, it may also include some tumor or non-tumor tissue, which led to low scores. We next asked if combining the results of the classification and the semantic segmentation CNNs improves the prediction accuracy. We found that neither of the two tested ensemble models, i.e. averaging the predictions or fitting a logistic regression decision function, outperformed the pure classification network, as shown in Fig. 3b-c.

Both networks yielded predictions for individual tiles. These need to be merged to visualize them for full WSIs. Our workflow for inference and visualization is illustrated in Additional file 1, Fig. S2. To reduce inference time, foreground detection was applied prior to tile extraction. Additionally, we imported predictions into QuPath to enable viewing the tissues and predictions jointly. Figure 4a shows three WSIs of the test dataset, followed by the manual annotation in Fig. 4b and corresponding predictions. To produce the colormaps in Fig. 4c, foreground detection was performed, and the resulting tiles were fed to the classification network. Additionally, the tumor probabilities were converted to binary class labels. The tumor probability map for all tiles, including background tiles, can be found in Additional file 1, Fig. S3. Figure 4d depicts the predictions of the segmentation network. In the segmentation maps, the class with the maximum probability was assigned to each pixel.

Class Activation Maps highlight pathological patterns

Grad-CAMs and HR-CAMs were computed for all tiles of the test dataset. We viewed samples of correct predictions to explore important patterns. First, we examined tiles that were correctly predicted as tumor by the classification network. The Grad-CAMs confirmed that the classifier focused strongly on present tumor cells instead of other surrounding tissue, as shown in Fig. 5a. HR-CAMs of tumor predictions led to the assumption that the presence of atypical cells was an important feature for the classification network. In some samples, a strong focus was on cells characterized by hyperchromatic nuclei that appear dark in the image. Other relevant features might be the abnormal size or irregular shape of the nucleus. These characteristics are also the key features in the decision-making process for pathologists. Examples, where such atypical cells are highlighted, are shown in Fig. 5b. We found that mitotic figures did not seem to be a relevant feature for tumor prediction, although it is a characteristic that pathologists often consider.

Additionally, we looked at the correct predictions of non-tumor samples. The Grad-CAMs appeared to be rather difficult to interpret because it was not clear why specific regions were highlighted. However, the HR-CAMs revealed that neighboring epithelial nuclei, often forming a structured pattern, are an important feature for the classification network. This is shown in Fig. 5c.

We found that Grad-CAMs and HR-CAMs highlighted similar patterns, with a mode correlation of 70.00% for the classification network. The mode correlation of 36.67% for the segmentation network was much lower. The corresponding distributions of the Pearson correlation coefficient can be found in Additional file 1, Fig. S4. To investigate whether decisions of the classification and the segmentation network were based on similar features, the correlation of their CAMs was also computed. We found that CAMs generated for the two distinct networks had a low correlation. In many cases, they focused on different image regions. The mode correlation of Grad-CAMs was 23.33%. HR-CAMs showed a higher mode correlation of 56.67%.

Figure 6 demonstrates the decision-making process as highlighted by Grad-CAM, and why we found different class predictions across the networks. For example, the tile in the first column shows carcinoma. The segmentation network predicted 65% tumor pixels based on present tumor cells. In contrast, the classification network predicted non-tumor because it focused on a region of cells that closely resemble healthy epithelial cells. The example in the second column in Fig. 6 was classified as non-tumor because of the presence of a blood vessel, but the segmentation network detected cancer cells in the tile.

Discussion

In this work, we created a manually annotated dataset from HE-stained slides of locally advanced HNSCC and trained CNNs for classification and semantic

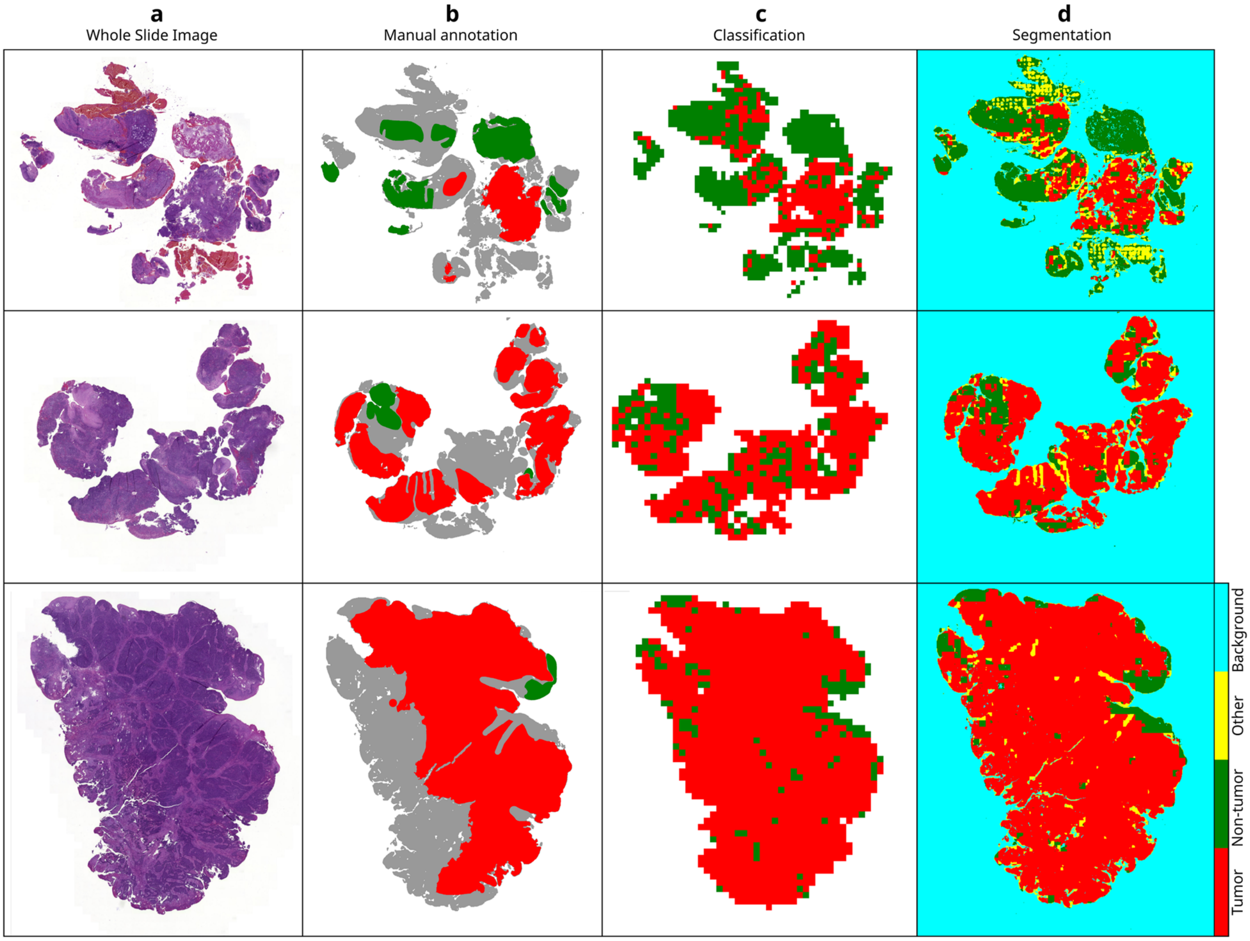

Fig. 4 Qualitative assessment of predictions for three WSIs of the test dataset. **a** WSIs. **b** Manual annotation. Not annotated tissue is shown in gray. **c** Predictions of the classification network, created by assigning class labels with a threshold of 0.5 and removing background tiles for better visibility. **d** Predictions of the segmentation network

segmentation. Both networks achieved a comparable performance as well as a high accuracy on unseen data, demonstrating their suitability for the detection of head and neck cancer.

The techniques Grad-CAM and HR-CAM were used to create visual explanations. The CAMs showed that both networks learned pathological patterns. For example, the presence of atypical cells with features such as irregular shapes or hyperchromatic, enlarged nuclei seemed to be important. Previously, Grad-CAM has shown that nuclear features contributed most to predictions of head and neck cancer [29] or to predictions of molecular subtypes of muscle-invasive bladder cancer [30]. These features agree with features used by expert pathologists, although they consider more characteristics such as the tissue structure or the number of mitoses, and take additional information into account, such as overall tumor morphology, growth patterns, and tumor architecture. Integrating the detection of such features for model training could be considered, for example, to enable cancer subtyping.

The reliability and meaningfulness of explanations created for the classification network were highlighted by a high correlation between Grad-CAMs and HR-CAMs. We found that the classification network and the segmentation network learned partly overlapping, but yet distinct patterns with a low overall correlation. This suggests taking both architectures into account. We found that Explainable AI techniques were very useful for investigating wrong or differing predictions of the two networks. CAMs can not only help in model development but also assist pathologists in reviewing predictions, making it easier for them to understand the CNN's abilities and to detect errors [6]. Moreover, it has been shown that presenting Grad-CAMs as additional information along with WSIs can improve the classification accuracy of pathologists [30].

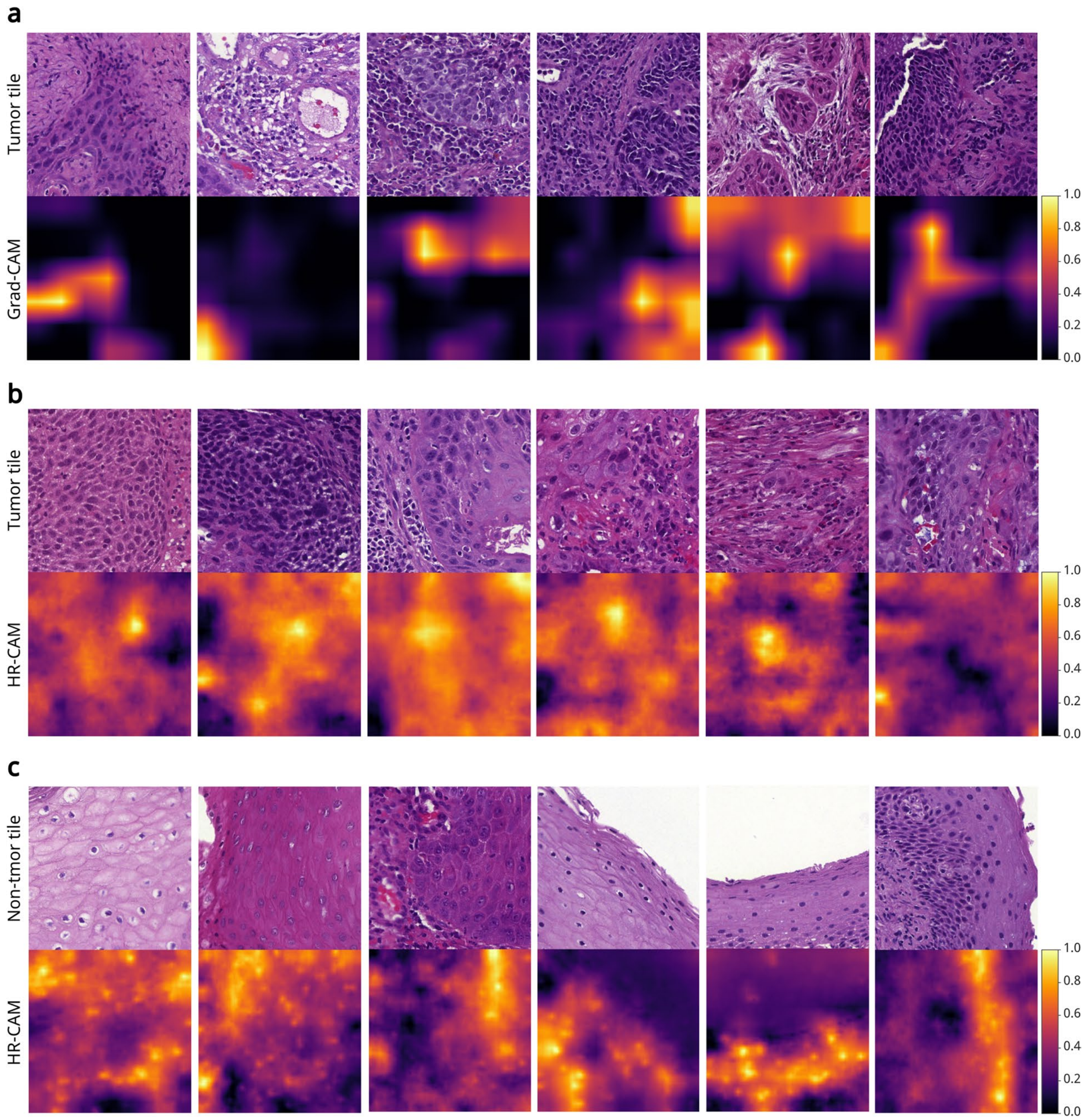

Fig. 5 Exemplary tiles of the test dataset with corresponding Grad-CAMs and HR-CAMs, generated for the classification network. The Grad-CAMs (7×7 pixels) were resized to 224×224 pixels. Values close to one indicate a high importance of the respective image region. **a** Tiles containing both tumor and surrounding tissue with Grad-CAMs. **b** Tiles containing atypical cells with HR-CAMs. **c** Tiles containing squamous epithelium with HR-CAMs

One limitation of our work results from coarse annotations. Most WSIs contained large, connected regions of either tumor or non-tumor tissue. Thus, it was very rare that one tile contained both classes. The segmentation network tended to classify most pixels in a tile either as tumor or non-tumor. This resulted in undesirable, tile-shaped class boundaries, as shown in Fig. 4d. Additionally, Halicek et al. recommend favoring a binary classification task in combination with such coarse annotations [29]. Therefore, the classification network is more suitable than the segmentation network to be trained on our dataset.

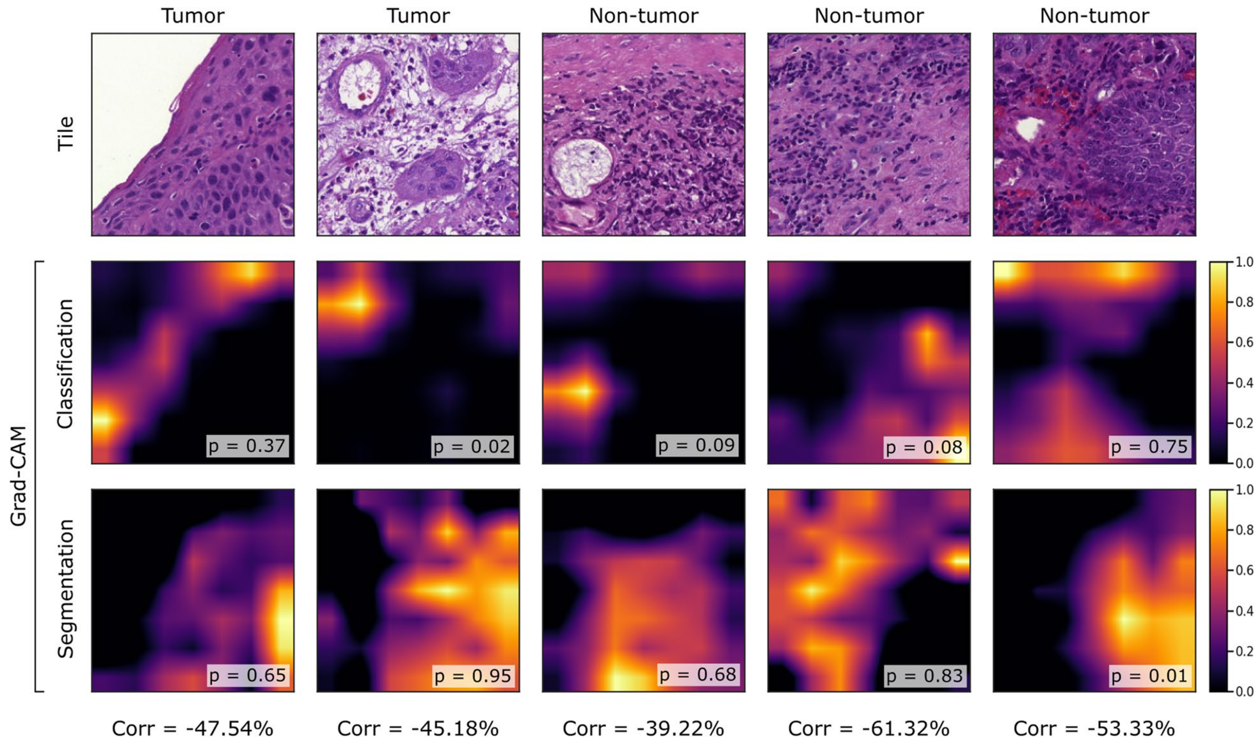

Fig. 6 Exemplary test tiles with a negative correlation between Grad-CAMs generated for the classification network and segmentation network. For each pair of Grad-CAMs, the Pearson correlation coefficient (Corr) was computed. For classification, the predicted tumor probability (p) and for segmentation, the fraction of predicted tumor pixels (p) is given. On the top, the binary ground truth labels, i.e. the predominant class in the tile based on the maximum pixel count, are shown

Tile-shaped class boundaries were also observed in multi-class breast cancer segmentation by Ho et al. who tackled this problem using multiple magnifications and precise annotations [31]. We argue that our segmentation network could improve using a similar strategy. Moreover, we found that binary class labels were ambiguous when both classes were present in a single tile. For example, the tile in the fifth column of Fig. 6 contained both cancer cells and squamous epithelium. This is a disadvantage of the binary classification approach. To avoid too coarse annotations and ambiguous labels, annotating individual cells should be considered, although this is more costly and time-consuming.

A second limitation is that both networks were not explicitly trained to distinguish tissues and artifacts. We observed that artifacts, such as written text or dust on the slide, were usually classified as non-tumor. However, some artifacts occurred as not annotated regions in the tiles, causing them to be labeled as class "other" by the segmentation network. This applied for tissue-fold artifacts and blood, for example. The left-most WSI in Fig. 4a contains blood, which is classified as "other" in Fig. 4d. An example of tissue-fold artifacts can be found in Additional file 1, Fig. S2. Still, we recommend applying a preceding artifact removal. Alternatively, artifacts could be annotated as an additional class for CNN training.

Another limitation relates to the use of HR-CAM for the segmentation network. HR-CAM relies on adding a global average pooling layer and a dense layer on top of the trained network [8]. The resulting model can only be trained for a classification task and not for semantic segmentation. Most likely, the resulting HR-CAMs highlight features important for the classification task and are therefore biased. This influenced the correlation between HR-CAMs and other CAMs, and led to a skewed distribution of correlation values, as shown in Additional file 1, Fig. S2c-d. For explaining the predictions of a segmentation network, Grad-CAM is more suitable than HR-CAM.

Conclusion

Our CNNs were able to detect head and neck cancer in unseen slides with high accuracy. This work contributed to the understanding of which features are learned from histological images, by comparing Grad-CAMs and HR-CAMs of the CNNs trained for different tasks. The Explainable AI techniques confirmed that the networks predicted the tumor class based on present pathological patterns, possibly focusing on nuclear features of atypical cells. This is consistent with how pathologists analyze tissue. Thus, CNNs seem promising in assisting pathologists in the assessment of cancer sections, especially in combination with visual explanations.

In the future, we will explore an alternative annotation strategy, namely annotating cell nuclei with a larger number of distinct classes. Future studies may also help to identify more nuclear features such as cell size, nucleoli, and cytoplasmic features in addition to the features we highlighted, to help identify cancer subtypes. Additional Explainable AI techniques such as occlusion experiments [32] or DeepLift [33] could be applied to further study class-discriminative features.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13000-023-01407-8.

Additional file 1: Figure S1. Mean validation accuracy for combinations of tile size and tile resolution, resulting from a grid search with iterated 5-fold cross-validation. Tiles were extracted at 512 pixels (99.6 µm), 768 pixels (149.4 µm), and 1024 pixels (199.1 µm), and then resized to different resolutions. **Figure S2.** Inference and visualization of predictions for an exemplary WSI in QuPath. (a) Classification map. (b) Segmentation map. (c), (d) and (e) Sections of both maps, superimposed on the image, in QuPath. Tumor is highlighted in red, non-tumor in green and "other" in yellow. Background tiles were omitted. **Figure S3.** Tumor probability map for three WSIs of the test dataset. (a) WSIs. (b) Manual annotation. (c) Tumor probability map, created from predictions of the classification network for all tiles, including background tiles. **Figure S4.** Correlation of CAMs of the two networks and using the two Explainable AI techniques. (a) Strong correlation between Grad-CAMs and HR-CAMs of the classification network. (b) Low correlation between Grad-CAMs of the classification and segmentation network. (c) Moderate correlation between Grad-CAMs and HR-CAMs of the segmentation network. (d) Bimodal correlation between HR-CAMs of the classification and segmentation network.

Acknowledgements

This work was funded in part by the Federal Ministry of Education and Research (BMBF) to AOG and ME (01KD2211B) and to AMK (01KD2211A). The clinical trial (CheckRad-CD8) was funded by AstraZeneca (ESR-16-12356). We acknowledge financial support by Deutsche Forschungsgemeinschaft and Friedrich-Alexander-Universität Erlangen-Nürnberg within the funding programme "Open Access Publication Funding".

Authors' contributions

A.M.K. and M.E. conceived the concept of this study and supervised the study. M.H., R.F., A.H., H.I., A.O.G., and M.E. provided the patient data and histological slides. M.D. annotated the data, analyzed the data, trained and evaluated neural networks, and applied Explainable AI techniques. M.D. and A.M.K. wrote the manuscript. M.H., A.O.G., and M.E. provided critical revision of the manuscript for important intellectual content.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request due to restricted access to the patient data of the CheckRad-CD8 trial.

Declarations

Competing interests

M.H. conflict of interest with Merck Serono (advisory role,speakers' bureau, honoraria, travel expenses, research funding); MSD (advisory role, speakers' bureau, honoraria, travel expenses, research funding); AstraZeneca (research funding, advisory role, speakers' bureau, honoraria, travel expenses); Novartis (research funding); BMS (advisory role, honoraria, speakers' bureau); Teva (travel expenses).

M.E. declares: Personal fees, travel costs and speaker's honoraria from MSD, AstraZeneca, Janssen-Cilag, Cepheid, Roche, Astellas, Diaceutics; research funding from AstraZeneca, Janssen-Cilag, STRATIFYER, Cepheid, Roche, Gilead; advisory roles for Diaceutics, MSD, AstraZeneca, Janssen-Cilag, GenomicHealth.

The other authors (M.D., A.O.G., R.F., H.I., A.H., and A.M.K.) do not declare any conflicts of interest with respect to this study.

Received: 19 July 2023 Accepted: 24 October 2023
Published online: 03 November 2023

References

1. Johnson DE, Burtness B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. Nat Rev Dis Prim. 2020;6(1):92.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;71(3):209–49.
3. Budach V, Tinhofer I. Novel prognostic clinical factors and biomarkers for outcome prediction in head and neck cancer: a systematic review. Lancet Oncol. 2019;20(6):e313–26.
4. Van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nat Med. 2021;27(5):775–84.
5. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. New York: IEEE; 2016. p. 2921–29.
6. Pocevičiūtė M, Eilertsen G, Lundström C. Survey of XAI in digital pathology. Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges. 2020;12090:56–88.
7. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. New York: IEEE; 2017. p. 618–26.
8. Shinde S, Chougule T, Saini J, Ingalhalikar M. HR-CAM: Precise localization of pathology using multi-level learning in CNNS. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22. Springer; 2019. p. 298–306.
9. Hecht M, Gostian AO, Eckstein M, Rutzner S, von der Grün J, Illmer T, et al. Safety and efficacy of single cycle induction treatment with cisplatin/docetaxel/durvalumab/tremelimumab in locally advanced HNSCC: first results of CheckRad-CD8. J Immunother Cancer. 2020;8(2):e001378.
10. Hecht M, Eckstein M, Rutzner S, von der Grün J, Illmer T, Klautke G, et al. Induction chemoimmunotherapy followed by CD8+ immune cell-based patient selection for chemotherapy-free radioimmunotherapy in locally advanced head and neck cancer. J Immunother Cancer. 2022;10(1). e003747.
11. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: Open source software for digital pathology image analysis. Sci Rep. 2017;7(1):1–7.
12. Tang Z, Chuang KV, DeCarli C, Jin LW, Beckett L, Keiser MJ, et al. Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. Nat Commun. 2019;10(1):2173.
13. Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K, Tsuneki M. Deep learning models for histopathological classification of gastric and colonic epithelial tumours. Sci Rep. 2020;10(1):1504.

14. Berman AG, Orchard WR, Gehrung M, Markowetz F. PathML: a unified framework for whole-slide image analysis with deep learning. medRxiv. 2021:2021–07. https://doi.org/10.1101/2021.07.07.21260138.
15. Raschka S, Mirjalili V. Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing Ltd; 2019.
16. Kannojia SP, Jaiswal G. Effects of varying resolution on performance of CNN based image classification: An experimental study. Int J Comput Sci Eng. 2018;6(9):451–6.
17. Chollet F. Deep learning with Python. Simon and Schuster; 2021.
18. Tellez D, Litjens G, Bándi P, Bulten W, Bokhorst JM, Ciompi F, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med Image Anal. 2019;58:101544.
19. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: fast and flexible image augmentations. Information. 2020;11(2):125.
20. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR; 2019. p. 6105–6114.
21. 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. IEEE; 2009. p. 248–55.
22. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–58.
23. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer; 2015. p. 234–241.
24. Iakubovskii P. Segmentation Models. GitHub; 2019. https://github.com/qubvel/segmentation_models. Accessed 30 May 2022.
25. Baheti B, Innani S, Gajre S, Talbar S. Eff-unet: a novel architecture for semantic segmentation in unstructured environment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New York: IEEE; 2020. p. 358–59.
26. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467. 2016.
27. Kingma DP, Ba J. Adam: A method for stochastic optimization. In: The International Conference on Learning Representations (ICLR). ICRL; 2015.
28. Vinogradova K, Dibrov A, Myers G. Towards interpretable semantic segmentation via gradient-weighted class activation mapping. In: Proceedings of the AAAI conference on artificial intelligence, vol. 34. Washington DC: AAAI; 2020. p. 13943–944.
29. Halicek M, Shahedi M, Little JV, Chen AY, Myers LL, Sumer BD, et al. Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. Sci Rep. 2019;9(1):14043.
30. Woerl AC, Eckstein M, Geiger J, Wagner DC, Daher T, Stenzel P, et al. Deep learning predicts molecular subtype of muscle-invasive bladder cancer from conventional histopathological slides. Eur Urol. 2020;78(2):256–64.
31. Ho DJ, Yarlagadda DV, D'Alfonso TM, Hanna MG, Grabenstetter A, Ntiamoah P, et al. Deep multi-magnification networks for multi-class breast cancer image segmentation. Comput Med Imaging Graph. 2021;88:101866.
32. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. Springer; 2014. p. 818–833.
33. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: International conference on machine learning. PMLR; 2017. p. 3145–3153.

Publisher's Note

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