Automated detection of the HER2 gene amplification status in Fluorescence *in situ* hybridization images for the diagnostics of cancer tissues

Falk Zakrzewski[1], Walter de Back[2], Martin Weigert[3,4], Torsten Wenke[5], Silke Zeugner[1], Robert Mantey[6], Katrin Friedrich[1], Ingo Roeder[2,6], Daniela E. Aust[1,6], Pia Hönscheid[1,6] & Gustavo Baretton[1,6]*

1 Institute of Pathology, University Hospital Carl Gustav Carus (UKD), TU Dresden, Dresden, Germany
2 Institute for Medical Informatics and Biometry (IMB), Carl Gustav Carus Faculty of Medicine, TU Dresden, Dresden, Germany
3 Max Planck Institute of Molecular Cell Biology and Genetics (MPI-CBG), Dresden, Germany
4 Center for Systems Biology Dresden (CSBD), Dresden, Germany
5 ASGEN GmbH & Co. KG, Dresden, Germany
6 National Center for Tumor Diseases (NCT), Partner Site Dresden, Germany
***** *Corresponding author*

Abstract

Background: The human epidermal growth factor receptor 2 (HER2) gene amplification status is crucial for developing a clinical strategy e.g. for evaluation of an anti-HER2-therapy in breast or stomach cancer. Therefore, the detection of HER2 gene amplification status is highly relevant in histopathological diagnostics. Recently, the application of convolutional neural networks (CNNs) has shown large progress in the automation of classification and object detection in medical image analysis.

Methods: Here, we apply deep learning-based pipeline for the detection, localization and classification of interphase nuclei depending on their HER2 gene amplification state in Fluorescence *in situ* hybridization (FISH) images. Our pipeline combines two CNN architectures named RetinaNet which are trained on (1) the detection and classification of interphase nuclei into normal, low-grade and high-grade and on (2) the detection and classification of FISH signals into HER2 and into the centromere of chromosome 17 (CEN17). In the first step (RetinaNet-1) nuclei are localized image-wide and a first classification is applied. The nuclei classification conducted via RetinaNet-1 is controlled and supplemented by HER2/CEN17 FISH signal ratios for the same nucleus by RetinaNet-2. Finally, an image-wide decision on the HER2 gene amplification stage is performed.

Results: We demonstrate that the accuracy of this deep learning-based pipeline is on par with that of a pathologist. The pipeline accurately classifies FISH images as demonstrated on set of 57 validation images containing hundreds of nuclei. Consequently, high quality FISH images can now be analyzed at once regarding their image-wide HER2 gene amplification status in our lab.

Conclusions: The automatic pipeline is a first step towards assisting the pathologist in evaluating the HER2 status of tumors using FISH images, for analyzing FISH images in retrospective and for optimizing the documentation of each tumor sample by automatically annotating and reporting of the HER2 gene amplification specificities.

Background

The human epidermal growth factor receptor 2 (HER2) gene, also designated ERBB2 gene for the v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, encodes a member of the epidermal growth factor receptor family of receptor tyrosine kinases. Amplification of the HER2 gene is the primary mechanism of HER2 overexpression in tumors[1]. HER2 amplification occurs before HER2 protein overexpression and consequently, monitoring of the tumor HER2 gene amplification status has become routine in breast cancer[2–4] surveillance. A positive HER2 status in around 25% of breast cancers is associated with poorer prognosis, more aggressive disease, and an increased risk of disease recurrence [2,5–7]. Application of HER2-directed therapies such as treatment with anti-HER2 antibodies, e.g. trastuzumab, depends on the detection of the HER2 gene amplification and increases overall survival of individuals suffering from HER2 positive breast cancer [2,6–10]. In addition to breast cancer, HER2 status testing is also applied in gastric cancers as trastuzumab is similarly effective in prolonging survival in HER2 positive carcinoma of the stomach and of the gastroesophageal junction[2,11].

HER2 testing is commonly carried out by immunohistochemistry (IHC), chromogenic *in situ* hybridization (CISH), silver-enhanced *in situ* hybridization (SISH) or Fluorescence *in situ* hybridization (FISH). In interphase nuclei of investigated tumor material, HER2 gene amplification testing is preferentially conducted via FISH[12]. In FISH analysis a HER2 positive state is defined when a HER2/CEN17 ratio of more than 2.2 is detectable, whereas CEN17 is a centromeric probe for the centromere of chromosome 17 on which the HER2 gene resides. Negative HER2 FISH amplification is defined as HER2/CEN17 ratio of less than 1.8[12]. Without an internal control probe such as CEN17, HER2 positive FISH is defined when above six HER2 genes are detectable per interphase nucleus while the equivocal range is defined with an average copy number of four to six HER2 genes per nucleus. Normal nuclei harbor two or fewer HER2 genes[13].

In clinical practice, the analysis is determined by the pathologist by observation of the FISH slide using the fluorescence microscope. The decision making relies on the individual expert knowledge of the pathologist and is dependent on standardization of the methodology, lab-dependent routines and finally on the quality of the FISH images (background signals, artifacts, tissue quality, and fluorescence microscope-dependent parameters). Pathologists analyze the HER2 gene amplification status of a tumor sample via evaluation in comparison to control samples. Testing criteria define HER2 positive status when (on observing within an area of tumor that amounts to > 10% of contiguous and homogeneous tumor nuclei) there is evidence of HER2 gene amplification based on counting at least 20 nuclei within this area[14]. By counting and classification of at least 20 interphase nuclei from different areas of the FISH slides a diagnostic decision is possible regarding a positive or negative state of HER2 gene amplification and its HER2 grade (low or high). The diagnostic relies on ratios of HER2 to CEN17 signals per nuclei on which the subsequent classification of the corresponding tumor sample is conducted.

While there are already automation methods for extracting features from microscopic images such as spot detection and counting[15,16], during the last years a notable increase in deep learning applications for classification tasks of pathological microscopic images were developed and successfully conducted on a wide field of applications[17]. Image classification tasks are commonly applied via Convolutional Neural Networks (CNNs) which rely on convolutional and non-linear transformations of the input data for a high-level abstraction classification[18]. Deep learning approaches such as CNNs have been already performed on pathology image classification, tumor classification, on imaging mass spectrometry data[16], in the identification of

metastatic cancer areas[17] or annotation of pathological images[19]. Recently, CNNs were applied for signal detection and counting in nuclei from FISH images (SpotLearn)[20] and segmentation of chromosomes in multicolor FISH images[21]. SpotLearn includes two supervised machine learning-based analysis workflows for the high-throughput segmentation and classification of large and diverse sets of FISH signals. FISH signals are detected with high accuracy in three separate fluorescence microscopy channels[20]. We aimed to develop a pipeline based on CNNs that works in one only channel because in our certified routine diagnostic workflow FISH signals are captured using a graded filter recording the different HER2 gene and CEN17 signals in one step which cannot be differentiated by SpotLearn. Furthermore, we targeted the localization of nuclei and FISH signals using fast one-stage-detectors rather than applying a segmentation algorithm workflow as applied in SpotLearn.

Therefore, we generated a lab-specific pipeline for automatic classification of FISH images comprising of many interphase nuclei into normal, low and high-grade on the basis of CNNs to be specifically used at our institute. The pipeline consists of two trained RetinaNet[22] steps for an image-wide classification of the HER2 gene amplification status. While the first RetinaNet (RetinaNet-1) detects and pre-classifies the nuclei, in the second RetinaNet step (RetinaNet-2) the HER2 and CEN17 signals are counted for each nucleus providing detailed information on each pre-classified nucleus. RetinaNet is a state-of-the-art, real time object detection and classification network with the aim of fast, accurate recognition of a wide variety of objects[22]. It relies on a Feature Pyramid Network[14] backbone on top of a feed-forward ResNet[23] architecture. To this backbone, RetinaNet attaches two subnetworks: one for classifying anchor boxes and one for regressing from anchor boxes to ground-truth object boxes[22]. Together with this architecture the new loss function, focal loss, is used that acts as a more effective alternative to previous approaches for dealing with class imbalance. RetinaNet is potentially more efficient than other state-of-the-art one-stage detectors because of the focal loss of RetinaNet, which applies a modulating term to the cross entropy loss in order to focus learning on hard negative examples, achieving state-of-the-art accuracy and speed of two-stage detectors[24]. To use these advantages for the detection on pathological samples, we applied RetinaNet in our deep learning-based system targeting the automation of FISH image evaluation regarding the HER2 grade detection at high accuracy and compared the performance of our system with the pathological assessment.

Material and Methods

Preparation of slides, Fluorescence *in situ* hybridization (FISH) and image capturing

Formalin-fixed Paraffin-embedded (FFPE) cancer tissue is delivered from clinical institutions from all over Germany (up to 20 patients per week). FFPE tissue is cut into small pieces (2µm) on a slide and dehydrated using first a xylene washing step subsequent flowed by a series of ethanol steps (100%, 96%, 70%). After drying the slide at room temperature slides are incubated with sodium thiocyanate followed by a wash step using distilled water. Subsequently, slides are incubated with pepsin and hydrochloric acid, washed using distilled water and dried at room temperature. Probes (PathVysion HER-2 DNA Probe Kit II, Abbott Inc.) are hybridized at 37°C in a wet chamber overnight. Washing of slides occurs in 2x saline-sodium citrate (SSC) buffer and DAPI counterstaining is conducted. Images are taken using fluorescence microscope (Axioskop 2, Zeiss Inc.) using a graded filter (Filter Set 23 (488023-0000-000), emission: 515-530 nm + 580-630 nm, Zeiss Inc.), recording HER2 gene

signals, CEN17 signals and a small subset of DAPI signals at once.

Image-wide nuclei detection and classification FISH images

Our pipeline consists of two RetinaNets[24] for detection and classification of nuclei occurring in a single FISH image. The pipeline was applied in Keras[28] with TensorFlow[29] backend in a Python environment. A Keras implementation of RetinaNet was used which is available on GitHub[25]. Initial labeling of training data was manually performed using labelIMG[30]. Training FISH images of high quality (minor background, minor number of artifacts, minor signal blurring, minor number of overlapping nuclei) were randomly chosen from documented high quality images of breast cancer FISH diagnostics on the HER2 gene amplification status from the years 2015-2018 harbored at the Institute of Pathology at the clinical campus of Carl Gustav Carus Hospital of TU Dresden. Training and validation (randomly chosen 10% of all images) was performed on each RetinaNet step, respectively. An overview about the training data set is given in **Table 1** for the RetinaNet-1 and in **Table 2** for the RetinaNet-2. The nuclei images used as training data for the RetinaNet-2 was randomly chosen from the FISH images used for training the RetinaNet-1. The RetinaNet-1 detects and classifies nuclei in a FISH image. A potential nuclei (or artifact) is marked via a bounding box and is additionally extracted and stored as an individual image file. A report text file containing the number of detected nuclei and their classification as well as the number of uncertain cases and artifacts is generated. The number of normal, low-grade and high-grade nuclei per FISH image is used for calculation of two ratios (ratio-1 and ratio-2): ratio-1 is low-grade nuclei/number of all detected nuclei and ratio-2 is high-grade nuclei/number of all detected nuclei. A FISH image is defined to be low-grade when ratio-1 is at least 0.2 while a FISH image is classified to be high-grade when ratio-2 is at least 0.4. These thresholds can be modified by the pathologist according to individual specificities and criteria. On top of the RetinaNet-1, the RetinaNet-2 detects and classifies the FISH signals in a single nucleus. Detected FISH signals were classified into HER2 signal, HER2 cluster (representing many not differentiable single HER2 signals) and CEN17 signals. All signals were counted respectively and for each nucleus the HER/CEN17 ratio is calculated. As soon as a HER cluster is detected the HER2/CEN17 ratio is automatically set to 10. If no HER2 signals are detected, the HER2/CEN17 ratio is automatically set to 1. In case no CEN17 signal is detected the nucleus is classified as artifact. For each FISH image the average HER2/CEN17 ratio is calculated on the basis of all HER2/CEN17 ratios from all detected nuclei from this FISH image. A HER2/CEN17 ratio greater than 1.5 and lower than 6.0 indicates a low-grade status of the FISH image. A value greater than 6.0 indicates a high-grade status of the FISH image. The RetinaNet-2 works automatically on top of the RetinaNet-1 and reports its detections in a second report text file. Each annotated nucleus is automatically stored as image file. Details on the training process for both RetinaNets, respectively, were as follows. We used a focal loss function22 for classification, and a smooth L1 loss function for bounding box regression together with the Adam optimizer25 with a fixed learning rate of 10-4. A batch size of 1 was used due to GPU memory limitations. The network was trained for 50 epochs on a single NVIDIA GPU (GeForce GTX 1080Ti) and took approximately 48 hours.

Table 1. Details on training FISH images

	# nuclei				
# images	**normal**	**low-grade**	**high-grade**	**# uncertain**	**# artifact**
299	626	782	1,760	2,037	1,050

Table 2. Details on training interphase nuclei

	# FISH signals		
# images	**CEN17**	**HER2**	**HER2 cluster**
301	512	1,552	441

Results

The certified FISH protocol is used routinely on the daily diagnostics for breast and stomach cancer patients and implements a standard procedure, which has been in use since 16 years. To enable an automated, *in-house* detection service we trained our pipeline on breast cancer FISH image samples originating from this routine diagnostics to enable applicability of the pipeline at our Institute. Image capture of FISH microscope images were taken using a graded filter recording HER2 and CEN17 signals in one step. The pipeline was trained for the detection of the HER2 gene amplification status into normal, low- and high-grade stage of routine FISH images from breast cancer samples. It relies on the implementation of two RetinaNets[22] trained on individual tasks respectively: The RetinaNet-1 with Resnet-50[23] backend was trained on up to 300 FISH images containing thousands of nuclei for nucleus detection and classification into high-grade, low-grade and normal nuclei (as well as artifacts and uncertain cases). The RetinaNet-2 with Resnet-50 backend was trained on up to 300 single nuclei images containing thousands of FISH signals for detection, classification and counting of FISH signals in each nucleus (HER2 single signals, HER2 cluster and CEN17 single signals). On the basis of the predictions of the RetinaNet-1 and the RetinaNet-2 on all nuclei of a FISH image a final decision is possible on the image-wide HER2 gene amplification status of the FISH image. This decision-making process is comparable to pathological assessment as in a first step nuclei are image-wide localized and classified and secondly a confirmation of the classification is applied on the basis of HER2/CEN17 ratios for each nucleus. The two major steps of our pipeline are explained in detail in the following sections. The training and prediction is illustrated in **Figure 1**.

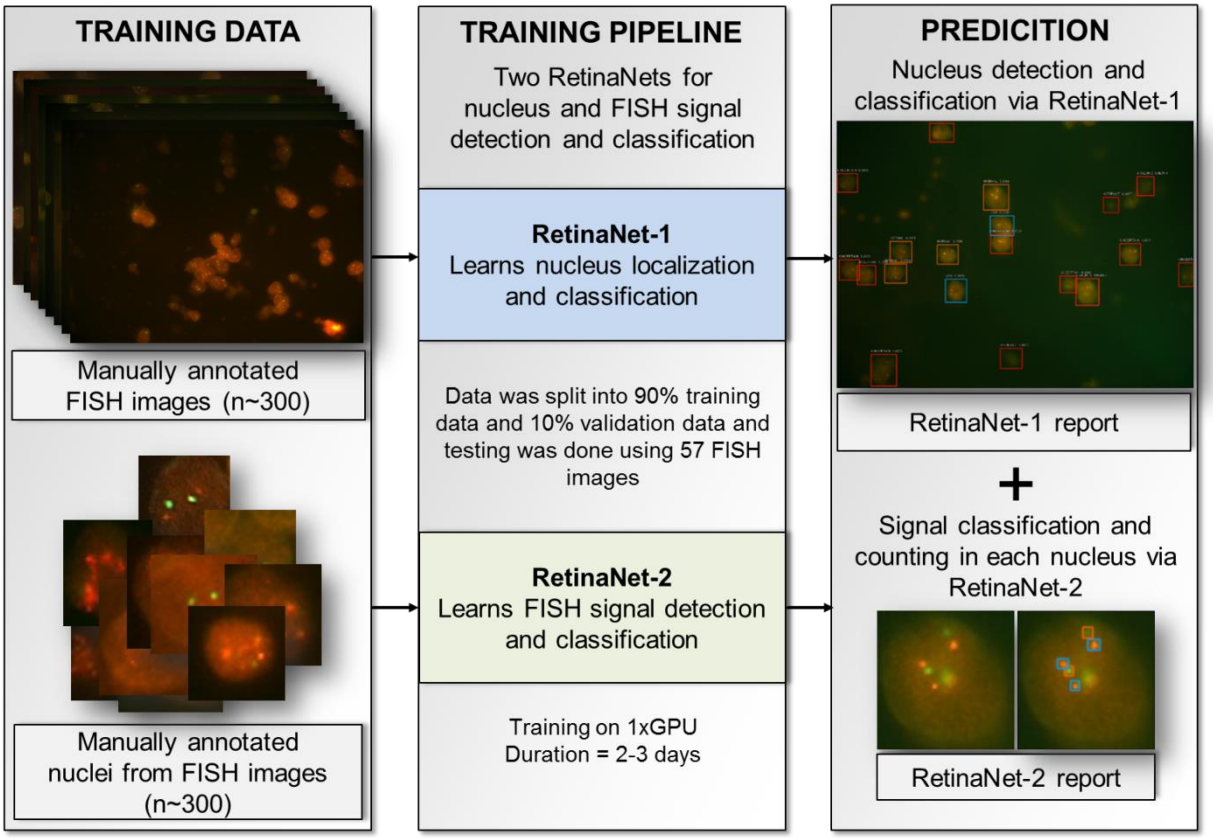


Figure 1. Overview of the detection pipeline of the HER2 gene amplification stage in FISH images from breast cancer samples.

Acquisition and manual labelling of the training data sets

FISH slides were prepared from FFPE tumor samples as described in the methods section. Probes against the HER2 gene and against the centromere of chromosome 17 were performed using the PathVysion HER-2 DNA Probe Kit II (Abbott Inc). Images were taken using the fluorescence microscope Axioskop 2 (Zeiss Inc.) using a graded filter (Filter Set 23 (488023-0000-000), emission: 515-530 nm + 580-630 nm, Zeiss Inc.), recording HER2 gene signals, CEN17 signals and a small subset of DAPI signals at once at a magnification of 100x. Images were captured using the Image –Pro MC 6.0 software and saved in JPEG file format with a size of 1200 x 1600 pixel. Our pipeline is optimized on these FISH images generated. We used up to 300 routine FISH images of high quality (minor or no background noise, minimal number of artifacts, no overlapping nuclei) from breast cancer samples (see detailed characterization of images in material and methods section). These FISH images represent a randomly selection of all FISH images of high quality routinely stored for training and documentation purposes from routine diagnostics of all analyzed breast cancer tumor samples which have been processed during the last three years at our institute. The manual annotation was conducted by a pathologist via labelling (bounding boxes) high-grade,

low-grade, normal nuclei as well as artifacts and uncertain nuclei for validation and test FISH images. HER2 and CEN17 FISH signals as well as cluster of HER2 signals were manually labelled (bounding boxes) by a pathologist in up

to 300 nuclei randomly chosen from all nuclei occurring within the ~300 previously mentioned FISH images.

RetinaNet-1: Detection and classification of interphase nuclei in FISH images

Training was performed on the manually labelled FISH images (n = 299) containing in total thousands of high-grade, low-grade and normal nuclei, as well as uncertain cases and artifacts (**Tab. 1**). The data was augmented using rotations, translations, shearing, scaling and horizontal and vertical flip. We used a focal loss function[22] for classification, and a smooth L1 loss function for bounding box regression, and the Adam optimizer[25] with a fixed learning rate of 10^{-4}. A batch size of 1 was used due to GPU memory limitations. The network was trained for 50 epochs on a single NVIDIA GPU (GeForce 1080Ti) and took approximately 48 hours.

The trained RetinaNet-1 automatically detects, localizes (via bounding boxes) and counts the number of normal, low-grade and high-grade nuclei as well as unidentifiable objects (uncertain cases and artifacts) image-wide. Each detected nuclei was stored as an individual image file using the detected bounding box and used in the RetinaNet-2 for FISH signals detection (see section below) as well as for potential manual re-evaluation and documentation purposes. Per FISH image, two ratios were calculated which allow conclusion about whether the image represents a positive (low or high grade) or normal state. The low-grade ratio (= ratio-1) indicates whether the FISH image is classified as HER2 low-grade and the high-grade ratio (= ratio-2) reports how likely it is that the FISH image is classified as high-grade. These ratios were calculated as follows: number of low-grade nuclei or high-grade nuclei divided by the sum of all detected and classified nuclei, respectively. As threshold we used values greater or equal to 0.2 for a nucleus to be a low-grade nucleus and 0.4 for a nucleus to be a high-grade nucleus. However, these thresholds are manually customizable according to the pathologist's definitions on specific ratios of the classified nuclei. Finally, the absolute occurrence of each class and the ratio-1 and ratio-2 were denoted in a report text file. Two exemplarily FISH images (low-grade and high-grade) complemented with the visualization of RetinaNet-1 object detection and classification results are shown in **Figure 2**.

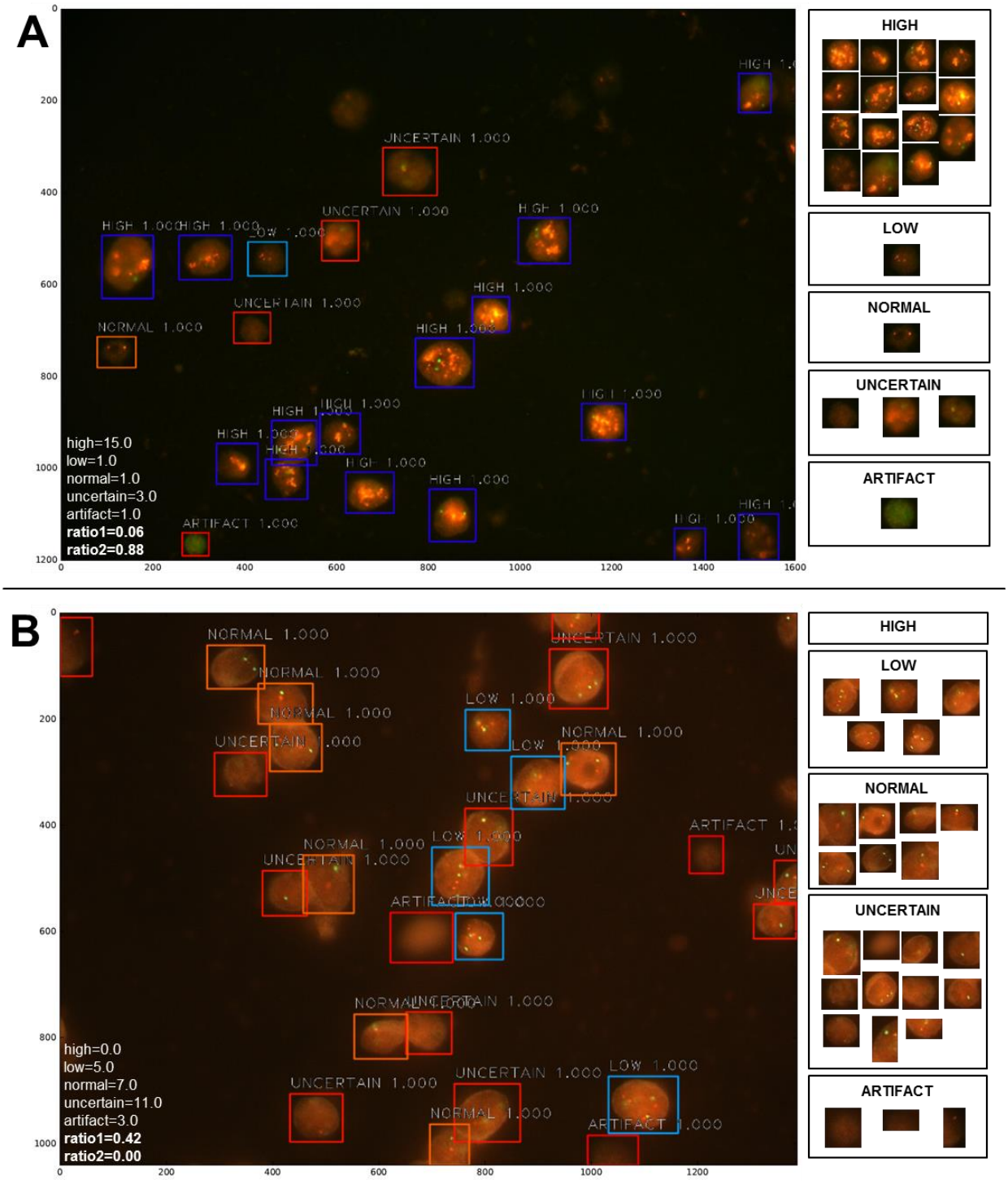


Figure 2. Application of RetinaNet pipeline on two Fluorescence *in situ* hybridization (FISH) images for interphase nuclei detection and classification.

(A) A high-grade stage was detected due to numerous high-grade nuclei. Only one nucleus was low-grade because it comprises four HER2 gene signals. One nucleus that was not detected is marked with a red arrow.

(B) A low-grade stage was detected due to five low-grade nuclei. Many nuclei were only classified as uncertain due to missing information on HER2 signals.

To validate the applicability and reliability of the first RetinaNet approach in routine diagnostics, 57 test high quality FISH images, containing 1,175 nuclei, were subject to image-wide nuclei detection and classification and compared to the annotation by a pathologist, considered as ground truth. The number of normal, high-grade, low-grade and unidentifiable nuclei (including artifacts and uncertain cases) were independently determined by the pathologist and by the RetinaNet-1 for each of the 57 FISH images. **Table 3** shows the results of the classification of the 1,175 nuclei as a confusion matrix. The classification performance is summarized using Cohen's kappa κ, a statistic measuring the degree of agreement between the predicted and the ground truth classification compared to a classification by chance. This results in κ=0.64, representing substantial agreement[26] over the whole validation set of nuclei (n=1,175). However, differentiation between normal from low-grade nuclei appears difficult, as shown by the prediction accuracy (acc) and reliability (rel) in **Table 3**, whereas high-grade nuclei were classified with high accuracy (acc=0.82) and reliability (rel=0.92). In addition, the accuracy of detection and classification of nuclei differed per image, ranging from poor accuracy (acc<0.5, 5 images) to near perfect classification (acc>0.85, 10 images) (**Suppl. Table 1**), with a mean accuracy of 0.73.

Nuclei in FISH images from our routine diagnostics might be of reduced quality compared *to up-to-date* fluorescence images as they have to be prepared under time limitation and a standardization procedure. High background noise, an increased number of artifacts and large differences in the number and shape of nuclei as well as overlapping nuclei all together influences the image quality of the captured nuclei. In addition, the quality depends on the input tumor material and available tissue type for analysis. To test the robustness of RetinaNet-1, we manually subdivided the nuclei from our investigated FISH images into the two groups "high quality" and "low quality" nuclei. Nuclei in the "high quality" group are characterized by clearly differentiable HER2 and CEN17 signals and by a uniform and regular nucleus shape without overlapping by further nuclei. In contrast, nuclei in the "low quality" group showed blurring of FISH signals, overlapping by further nuclei, very weak FISH signals or signal artifacts which made it difficult to adequately detect the signals. As shown in **Table 3**, Cohen kappa is reduced from substantial agreement (κ=0.64) for high quality nuclei to moderate agreement (κ=0.54) in the case of low quality nuclei. In particular, the accuracy of classification for high-grade nuclei is reduced between the high- and low quality nuclei (acc=0.93 vs. acc=0.55).

Table 3. Classification performance of RetinaNet-1 on validation images (n=57)

RetinaNet-1 on nuclei - all nuclei						
		Pathologist				
RetinaNet-1		normal	low	high	unid.	REL
	normal	**57**	40	1	43	*0,40*
	low	29	**85**	27	19	*0,53*
	high	0	5	**277**	18	*0,92*
	unid.	55	14	31	**474**	*0,83*
	ACC	*0,40*	*0,59*	*0,82*	*0,86*	***κ = 0.64***
RetinaNet-1 on nuclei - high quality nuclei						
		Pathologist				
RetinaNet-1		normal	low	high	unid.	REL
	normal	**41**	26	0	34	*0,41*
	low	24	**66**	11	13	*0,58*
	high	0	3	**224**	8	*0,95*
	unid.	29	6	5	**151**	*0,79*
	ACC	*0,44*	*0,65*	*0,93*	*0,73*	***κ = 0.65***
RetinaNet-1 on nuclei - low quality nuclei						
		Pathologist				
RetinaNet-1		normal	low	high	unid.	REL
	normal	**16**	14	1	9	*0,40*
	low	5	**19**	16	6	*0,41*
	high	0	2	**53**	10	*0,82*
	unid.	26	8	26	**323**	*0,84*
	ACC	*0,34*	*0,44*	*0,55*	*0,93*	***κ = 0.54***

RetinaNet-2: Counting the HER2 and CEN17 FISH signals per nucleus and detection of the image-wide HER2/CEN17 ratio

In order to increase the classification accuracy and to validate and control the nucleus-based classification performed by RetinaNet-1, we trained a second RetinaNet (RetinaNet-2) to localize and classify the individual HER2 and CEN17 signals in each nucleus detected by RetinaNet-1. It acts as a control mechanism (comparable to a second opinion) and provides additional detailed source of information on the number of HER2 and CEN17 signals per nucleus.

HER2 FISH signal detection was split into two classes: HER2 single signals and HER2 clusters. A HER2 cluster represents a region of the nucleus which is characterized by a high density of adjacent HER2 signals which often cannot be well distinguished into the underlying single signals and hence appeared as an accumulation. Training was performed on in total thousands of HER2 and CEN17 signals from the ~300 randomly selected nuclei from the ~300 training FISH images. Apart from the different input images, where we used images containing a single nucleus instead of a complete FISH image, the training process was identical to that described above for RetinaNet-1.

The nucleus-specific FISH signal detection and classification via the RetinaNet-2 automatically works on top of the initial nuclei detection and

classification performed via the RetinaNet-1. Each nucleus detected via the RetinaNet-1 is automatically fed into the RetinaNet-2 where FISH signals were classified and counted, documented in an image-wide report text file and visualized in an additional nucleus-specific image file. RetinaNet-2 predicts a bounding box and classifies each individual HER2 signal, HER2 cluster and CEN17 signal. Afterwards the boxes are counted and the ratio of HER2/CEN17 signals is calculated per nucleus. If no HER2 signals were detected the HER2/CEN17 ratio was automatically set to 1. In case no CEN17 signal was detected the nucleus was classified as uncertain. When a HER2 cluster was detected the HER2/CEN17 ratio was set to 10 as a HER2 clusters may contain a high but unknown number of HER2 signals. The average and image-wide HER2/CEN17 ratio was calculated on the basis of all detected nuclei harboring CEN17 and HER2 signals. This quantity was used to decide the image-wide HER2 gene amplification status of the corresponding FISH image.

To measure the performance of RetinaNet-2 for nucleus classification, 50 randomly selected nuclei were analyzed by the RetinaNet-2 and compared to the manual annotation by the pathologist. In six cases a different classification was revealed (**Tab. 4**). In three of the six cases, a normal nucleus was classified via the RetinaNet-2 while the pathologist detected a low-grade nucleus which was caused due to missed HER2 single signal detection via the RetinaNet-2. In two of the six cases the RetinaNet-2 detected a high-grade nucleus while the pathologist classified these nuclei as normal. The reason was that RetinaNet-2 detected a HER2 signal as HER2 cluster because of strong blurring of the single HER2 signal mimicking a HER2 cluster. In one of the six cases, a classification via the RetinaNet-2 was not possible although the same number of HER2 signals and CEN17 signals was found in comparison to the pathologist. However, because only one HER2 signal was identified, the RetinaNet-2 classified the nucleus as “uncertain”.

Table 4. Comparison of FISH signal detection of RetinaNet-2 with the pathologist on 50 validation interphase nuclei

Nucleus	CEN17 single signal		HER2 single signal		HER2 cluster		Nucleus classification	
	RetinaNet-2	Pathologist	RetinaNet-2	Pathologist[1]	RetinaNet-2	Pathologist[2]	RetinaNet-2	Pathologist
1	1	2	2	2	0	0	normal	normal
2	1	1	1	1	0	0	normal	normal
3	2	2	2	2	0	0	normal	normal
4	2	2	2	5	0	0	normal	low
5	2	2	3	3	0	0	low	low
6	1	1	2	3	0	0	normal	low
7	1	1	2	2	1	1	normal	normal
8	1	1	1	1	0	0	normal	normal
9	2	3	3	3	0	0	low	low
10	0	1	5	4	0	0	low	low
11	2	2	2	2	0	0	normal	normal
12	2	2	2	2	0	0	normal	normal
13	3	3	2	4	0	0	normal	low
14	2	2	4	4	0	0	low	low
15	1	1	1	2	0	0	normal	normal
16	2	2	2	2	0	0	normal	normal
17	2	2	3	3	0	0	low	low
18	2	2	2	2	0	0	normal	normal
19	2	3	3	3	0	0	low	low
20	4	4	5	5	0	0	low	low
21	2	2	1	1	0	0	uncertain	normal
22	1	1	1	1	0	0	normal	normal
23	2	3	3	3	0	0	low	low
24	2	2	2	2	0	0	normal	normal
25	2	4	4	4	0	0	low	low
26	1	2	4	3	0	0	low	low
27	3	3	3	4	0	0	low	low

28	2	2	2	2	0	0	normal	normal
29	2	2	2	2	0	0	high	normal
30	1	2	3	n.d.	2	1	high	high
31	0	0	2	n.d.	2	1	high	high
32	2	2	4	4	0	0	low	low
33	3	3	4	n.d.	4	1	high	high
34	1	2	2	2	0	0	normal	normal
35	0	0	4	n.d.	1	1	high	high
36	1	1	2	n.d.	1	1	high	high
37	0	0	1	n.d.	1	1	high	high
38	2	1	2	n.d.	2	1	high	high
39	0	1	0	n.d.	3	1	high	high
40	1	1	2	n.d.	1	1	high	high
41	0	0	3	n.d.	2	1	high	high
42	0	0	0	n.d.	4	1	high	high
43	2	2	0	2	1	0	high	normal
44	0	0	1	n.d.	2	1	high	high
45	0	0	5	n.d.	2	1	high	high
46	1	2	6	n.d.	1	1	high	high
47	0	0	0	n.d.	5	1	high	high
48	1	1	4	n.d.	1	1	high	high
49	1	2	4	4	0	0	low	low
50	0	0	1	n.d.	6	1	high	high

n.d. HER2 signals were not quantified in the Ground Truth in case at least one HER2 cluster was identified. HER2 cluster were not counted in the Ground Truth because of the strong subjective aspect of this procedure. The value was set to 1 when at least one HER2 cluster occurred.

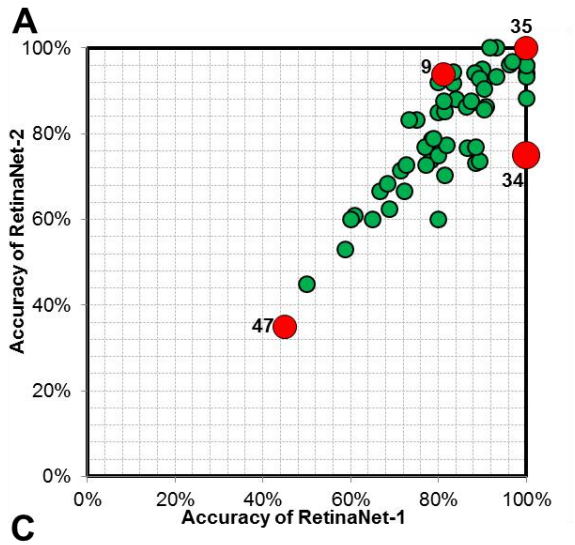
To validate the applicability and reliability of the RetinaNet-2 approach in the image-wide classification, 57 test FISH images (same images which were used for validation of the RetinaNet-1) were subject to their image-wide nuclei detection and classification compared to the ground truth annotated by the pathologist. The comparison was also conducted to “high quality” and “low quality” nuclei as previously done for the RetinaNet-1 to test the robustness on nuclei images of lower quality (**Tab. 5**). Again, we find a substantial agreement between our deep learning system and the human pathologist, but at a higher level of agreement, κ=0.76. In particular, the classification accuracy of normal nuclei has increased as compared to RetinaNet-1 (acc=0.40 and acc=0.72, respectively, **Tab. 3**). Whereas for low quality nuclei, we find only a moderate agreement (κ=0.55), similar to the performance of RetinaNet-1, for nuclei recorded at high quality, we find the classification performance of RetinaNet-2 κ=0.85, representing an almost perfect agreement with the pathologist. Nevertheless, a minor number of HER2 double or triple signals in very close vicinity were annotated as HER2 cluster leading to the wrong overall assumption that a high-grade nucleus occurred.

The accuracy of both RetinaNets was compared regarding the image-wide detection and classification of nuclei in the 57 test FISH images (**Suppl. Tab. 1**). Similar to RetinaNet-1, we found for RetinaNet-2 that the detection and classification performance differs between images. Interestingly, however, several images where RetinaNet-1 performed poorly were well-classified by RetinaNet-2 and vice versa, indicating these two approaches are complementary and are best used in combination (**Suppl. Tab. 1).** However, for most of the images the accuracy equals (**Fig. 3A**) but a few images show larger differences in their accuracies. Four example images were depicted where (1) the accuracy was 100% for both RetinaNets (**Fig. 3, image 35**), (2) the accuracy was low for RetinaNet-2 but high in RetinaNet-1 (**Fig. 3, image 16**), (3) the accuracy was lower for RetinaNet-1 compared to RetinaNet-2 (**Fig. 3, image 9**) and (4) the accuracy was similar low for both RetinaNets (**Fig. 3, image 47**). Nuclei in the FISH images are marked with a red arrow where a different classification was obtained by

Table 5. Classification performance of RetinaNet-2 on validation images (n=57).

RetinaNet-2 on nuclei - all nuclei						
		Pathologist				
		normal	low	high	unid.	REL
RetinaNet-2	normal	**102**	21	4	16	*0,71*
	low	1	**80**	24	2	*0,75*
	high	3	14	**287**	18	*0,89*
	unid.	35	29	21	**519**	*0,86*
	ACC	*0,72*	*0,56*	*0,85*	*0,94*	***κ = 0.76***
RetinaNet-2 on nuclei - high quality nuclei						
		Pathologist				
		normal	low	high	unid.	REL
RetinaNet-2	normal	**77**	8	0	10	*0,81*
	low	1	**73**	4	1	*0,92*
	high	3	7	**233**	7	*0,93*
	unid.	13	13	3	**188**	*0,87*
	ACC	*0,82*	*0,72*	*0,97*	*0,91*	***κ = 0.85***
RetinaNet-2 on nuclei - low quality nuclei						
		Pathologist				
		normal	low	high	unid.	REL
RetinaNet-2	normal	**25**	13	4	6	*0,52*
	low	0	**7**	20	1	*0,25*
	high	0	7	**54**	11	*0,75*
	unid.	22	16	18	**331**	*0,86*
	ACC	*0,53*	*0,16*	*0,56*	*0,95*	***κ = 0.55***

both RetinaNets (**Fig. 3C**). In image 35 nuclei are clearly distinguishable and show massive amplification of the HER2 gene, which can be easily and clearly detected by both RetinaNets. Therefore, no differences in the nuclei classification were detected. In image 47 the performance of the two RetinaNets is equally low due to the general low quality of many nuclei occurring in the image. In addition, the overall number of nuclei in the image is low so that the influence of the "low quality" nuclei on the image-wide classification is higher. Reasons for different classification between RetinaNet-1 and RetinaNet-2 in images 9 and 34 may be due to weak and blurring FISH signals not seen by RetinaNet-2 and/or the interpretation of very adjacent located HER2 gene signals as HER2 cluster by RetinaNet-2 leading to false classification of the corresponding nucleus. More precisely, **Figure 4** shows selected and representative examples on three cases for a same and three cases for a different classification between both RetinaNets. The three nuclei in the left column (**Fig. 4A-C**) were classified identically by both networks and the classification corresponds to those of the pathologist, providing stronger confidence in the correct classification. The three nuclei in the right column, however, were classified differently (**Fig. 4D-F**). In the first case (**Fig. 4D**) RetinaNet-2 detected three HER2 signals in close vicinity as a single HER2 cluster, leading to a misclassification as high-grade nucleus while the RetinaNet-1 correctly classified this nucleus as low-grade.



B

	Image	9	34	35	47
Normal	RetinaNet-1	0	0	0	1
	RetinaNet-2	0	1	0	1
	Pathologist	0	0	0	4
Low-grade	RetinaNet-1	2	1	0	2
	RetinaNet-2	0	3	0	1
	Pathologist	0	0	0	4
High-grade	RetinaNet-1	10	18	8	0
	RetinaNet-2	11	13	8	2
	Pathologist	12	18	8	0
Indefinable	RetinaNet-1	3	4	2	6
	RetinaNet-2	4	6	2	5
	Pathologist	4	2	2	12
Correct	RetinaNet-1	13	20	10	9
	RetinaNet-2	15	15	10	7
	Pathologist	16	20	10	20
	RetinaNet-1	81%	100%	100%	45%
	RetinaNet-2	94%	75%	100%	35%
Difference		**-13%**	**25%**	**0%**	**10%**

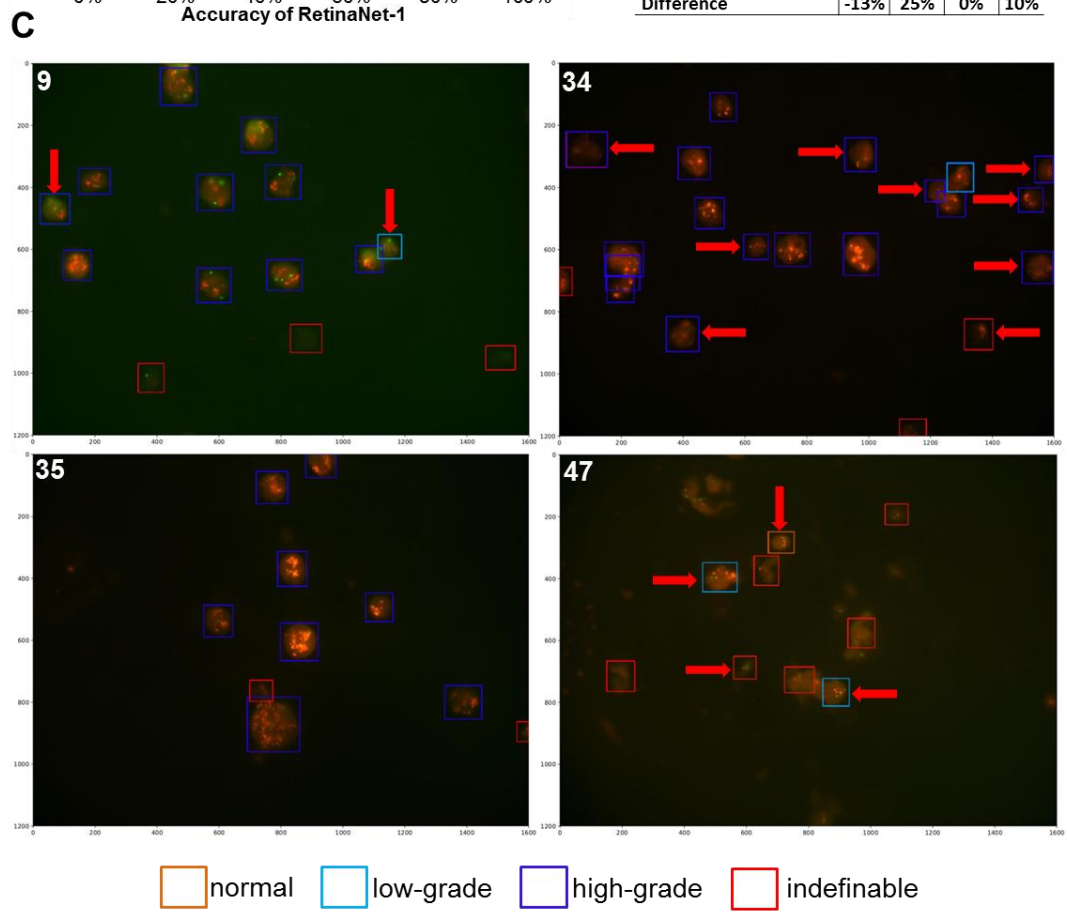


Figure 3 Comparison of the accuracy of RetinaNet-1 and RetinaNet-2 on the image-wide nucleus detection and classification.

(A) The accuracy in classifying the detected nuclei was compared among the 57 validation FISH images. Exemplarily, the four most interesting images are depicted where (1) the accuracy was 100% in both RetinaNets (image 35), (2) the accuracy was low in both networks (image 47), (3) RetinaNet-1 had a much better accuracy than RetinaNet-2 (image 34) or (4) vice versa (image 9). **(B)** Detailed overview about the classification of the nuclei in these four FISH images and **(C)** visualization of the classification of RetinaNet-1. Nuclei with a differing classification by RetinaNet-2 are marked with a red arrow.

In the second case (**Fig. 4E**), RetinaNet-2 missed the detection of HER2 and CEN17 signals, presumably due to overexposure, and therefore a misclassification of the nucleus as normal was conducted. The RetinaNet-1 correctly classified the nucleus as low-grade. Finally, in **Figure 4F**, RetinaNet-2 correctly detected all signals but classified the nucleus as normal in contrast to RetinaNet-1 which conducted a classification as uncertain. However, the pathologist´s classification was low-grade.

Automated classification of high quality FISH images into normal, low- and high-grade

Our nuclei detection and classification system relies on the combination of two steps performed by the RetinaNet-1 and the RetinaNet-2 enabling a final decision on the HER2 gene amplification status with HER2 and CEN17 FISH signal counting of the whole FISH image. The decision relies on ratios being calculated in both RetinaNet steps. In the RetinaNet-1 the ratio-1 and the ratio-2 (ranging from 0 to 1, respectively) are calculated and indicate on the relative number of low grade nuclei (ratio-1) and high grade nuclei (ratio-2), respectively, compared to the overall occurrence of all classifiable nuclei. A low-grade or high-grade stage of is indicated by ratio-1 greater or equal to 0.2 and by ratio-2 greater than 0.4, respectively. Both thresholds are modifiable with respect to the pathologist's specified criteria. In the RetinaNet-2 an image-wide HER2/CEN17 ratio is calculated as average value among all nuclei-specific HER2/CEN17 ratios of classifiable nuclei. A HER2/CEN17 ratio greater than 1.5 and lower than 6.0 indicates a low-grade status of the FISH image. A value greater than 6.0 indicates a high-grade status of the FISH image. The maximum average value is 10.0 because the highest value a single nucleus can obtain is 10.0 due to the fact that as soon as a HER2 cluster is detected the value is automatically set to 10.0. The overall image-wide classification of the HER2 gene amplification status is mostly identical between our pipeline (RetinaNet-1 and RetinaNet-2) and the pathologist on the 57 test FISH images (**Tab. 6**). In two of the 57 cases a different classification of was denoted. In one of the two cases, the RetinaNet-1 classified a low-grade FISH image while the RetinaNet-2 had a tendency towards a high-grade image. In the second case, the RetinaNet-1 classified the image to be low-grade while the RetinaNet-2 classified it as high-grade due to a misclassification of one normal nucleus as high-grade nucleus.

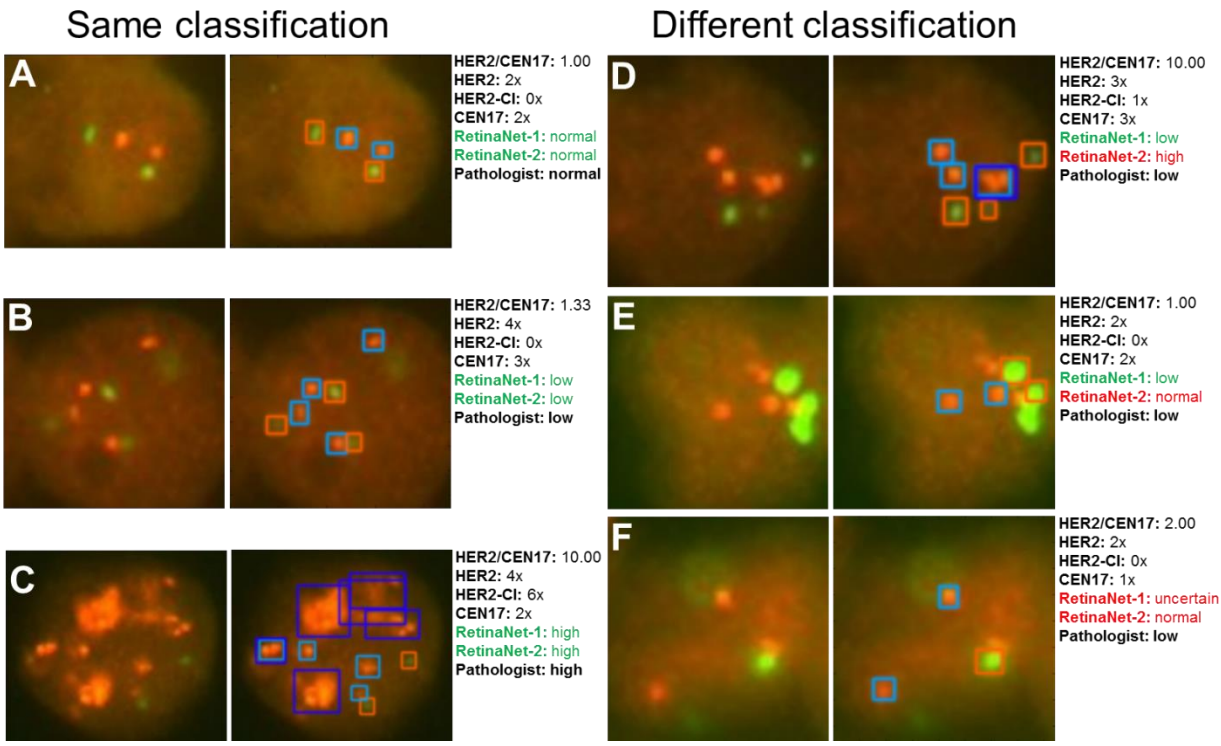


Figure 4. Application of the second RetinaNet pipeline on three interphase nuclei detection and re-classification. The second RetinaNet pipeline (RetinaNet-2) detects FISH signals in each of the separated and pre-classified nuclei from the first RetinaNet pipeline (RetinaNet-1) and classifies the signals into HER2 (light blue framed boxes), HER2 cluster (dark blue framed boxes) and CEN17 (red framed boxes) signals. Exemplarily three nuclei are shown: **(A-C)** The classification was confirmed. **(D)** The classification was not confirmed due to a misinterpretation of three very adjacent HER2 gene single signals as HER2 cluster. **(E)** The classification was not confirmed because two HER2 gene signals were not detected by the RetinaNet-2. **(F)** The classification was not confirmed although all FISH signals were detected by RetinaNet-2 because the classification was defined to be unclassifiable if only one CEN17 signals occurs.

Table 6. Detection of the HER2 gene amplification status in 57 validation FISH images

	RetinaNet-1[1]		RetinaNet-2[2]	Grade estimation	Pathologist
Image	**Ratio1[3]**	**Ratio2[4]**	**HER2/CEN17[5]**	**based on RetinaNet-1/-2**	**Grade**
1	0,08	0,83	9,42	HIGH/HIGH	HIGH
2	0,07	0,93	9,43	HIGH/HIGH	HIGH
3	0	1,00	10,00	HIGH/HIGH	HIGH
4	0,13	0,88	7,75	HIGH/HIGH	HIGH
5	0,33	0,67	10,00	HIGH/HIGH	HIGH
6	0,07	0,93	9,69	HIGH/HIGH	HIGH
7	0,22	0,78	9,11	HIGH/HIGH	HIGH
8	0	0,91	9,13	HIGH/HIGH	HIGH
9	0,17	0,83	10,00	HIGH/HIGH	HIGH
10	0	1,00	10,00	HIGH/HIGH	HIGH
11	0	1,00	10,00	HIGH/HIGH	HIGH
12	0	1,00	10,00	HIGH/HIGH	HIGH
13	0,17	0,75	9,46	HIGH/HIGH	HIGH
14	0,25	0,75	9,15	HIGH/HIGH	HIGH
15	0,26	0,74	8,08	HIGH/HIGH	HIGH
16	0,57	0	2,63	LOW/LOW	LOW

17	0,75	0	3,67	LOW/LOW	LOW
18	0,67	0,17	1,40	LOW/LOW	LOW
19	0,31	0	3,37	LOW/LOW	LOW
20	0,36	0	1,55	LOW/LOW	LOW
21	0,40	0	4,33	LOW/LOW	LOW
22	0,40	0	1,65	LOW/LOW	LOW
23	0,31	0	1,65	LOW/LOW	LOW
24	0,33	0	1,93	LOW/LOW	LOW
25	0,50	0,10	1,67	LOW/LOW	LOW
26	0,50	0	4,89	LOW/LOW	LOW
27	0,44	0,22	3,33	LOW/LOW	LOW
28	0,27	0	1,00	**LOW**/NORMAL	**NORMAL**
29	0,25	0	3,43	LOW/LOW	LOW
30	0,15	0,69	8,71	HIGH/HIGH	HIGH
31	0	1,00	8,20	HIGH/HIGH	HIGH
32	0	1,00	9,00	HIGH/HIGH	HIGH
33	0,09	0,91	8,82	HIGH/HIGH	HIGH
34	0,05	0,95	8,93	HIGH/HIGH	HIGH
35	0	1,00	10,00	HIGH/HIGH	HIGH
36	0,15	0,85	9,43	HIGH/HIGH	HIGH
37	0,17	0,83	10,00	HIGH/HIGH	HIGH
38	0	1,00	10,00	HIGH/HIGH	HIGH
39	0	1,00	10,00	HIGH/HIGH	HIGH
40	0,15	0,85	9,41	HIGH/HIGH	HIGH
41	0	1,00	10,00	HIGH/HIGH	HIGH
42	0	1,00	10,00	HIGH/HIGH	HIGH
43	0	1,00	10,00	HIGH/HIGH	HIGH
44	0,29	0	1,35	LOW/LOW	LOW
45	0,44	0	1,63	LOW/LOW	LOW
46	0,56	0	1,92	LOW/LOW	LOW
47	0,67	0	5,63	LOW/LOW	LOW
48	1,00	0	4,50	LOW/LOW	LOW
49	0,53	0	3,20	LOW/LOW	LOW
50	0,69	0,06	1,91	LOW/LOW	LOW
51	0,38	0	2,44	LOW/LOW	LOW
52	0,40	0	2,81	LOW/LOW	LOW
53	0,67	0,17	4,93	LOW/LOW	LOW
54	1,00	0	1,17	LOW/LOW	LOW
55	0,40	0	7,00	**LOW/HIGH**	**LOW**
56	0,29	0	1,39	LOW/LOW	LOW
57	0,43	0	2,90	LOW/LOW	LOW

1 The first RetinaNet pipeline detects and classifies nuclei image-wide in a FISH image
2 The second RetinaNet pipeline detects, classifies and counts FISH signals in each nucleus detected by the first RetinaNet pipeline
3 Ratio-1 represents number of low-grade nuclei divided by the number of all classified nuclei. A value greater than 0.2 indicates LOW stage of the FISH image.
4 Ratio-2 represents number of high-grade nuclei divided by the number of all classified nuclei. A value greater than 0.4 indicates HIGH stage of the FISH image.
5 The average of all detected HER2/CEN17 ratios among all nuclei in one FISH image. A value greater than 1.0 indicate LOW stage and greater than 6.0 HIGH stage of the FISH image.

Conclusions

In this study, we developed a deep learning pipeline for analyzing Fluorescence *in situ* hybridization (FISH) images regarding the image-wide detection of interphase nucleus and their classification depending on the HER2 gene amplification level. The pipeline can be useful in assisting pathologists in analyzing the HER2 gene amplification stage of a breast cancer samples by automatically analyzing high quality FISH images for control purposes. It can also be used for automatic investigation in retrospective

studies of large amounts of documented FISH images collected over several years at our Institute for re-evaluation. Another application could be the enhancement of the documentation quality of the images. Furthermore, an anonymized and human-independent evaluation of the HER2 gene amplification level is possible. Analyzing one FISH image including the generation of the annotated image data and the report files occurs in less than a second which is quite faster than comparable human evaluation so far. Therefore, we interpret our pipeline as a first step towards the automation of the HER2 gene amplification detection in FISH images.

Image-wide ratios representing the number of abnormal nuclei in relationship to all classified nuclei are calculated which serve as guideline for classifying the HER2 gene amplification status of the corresponding tumor sample from which the FISH image originated from. Our pipeline works on the basis of two CNNs for localization and classification, called RetinaNet. RetinaNet-1 detects and classifies nuclei in the FISH images and calculates two ratios. Ratio-1 (ranging from 0 to 1) represents how frequently nuclei with a low amplification status of HER2 genes occurred and ratio-2 (ranging from 0 to 1) indicates the same for nuclei with high amplification status of the HER2 gene. The second RetinaNet, RetinaNet-2, detects and classifies FISH signals for each nucleus into HER2 signals, CEN17 signals and HER2 cluster, which consists of multiple, non-distinguishable HER2 signals. An average HER2/CEN17 image-wide ratio is calculated on the basis of all nucleus-specific HER2/CEN17 ratios which serves for decision making on the image-wide HER2 gene amplification status of the FISH image. The reliability of our pipeline in making correct classification of the image-wide HER2 gene amplification status was demonstrated to be comparable to the pathologist with an accuracy of 96% based on 57 annotated FISH images (55 out of 57 images).

However, there are limitations of our pipeline: An important issue is the difficulty in predicting the HER2 gene amplification status in routine FISH images of relatively low quality characterized by high background noise, low signal-to-noise ratio, a large number of artifacts, strong differences in nuclei shape, weak signals and truncated or overlapping nuclei. To overcome these limitations we would need to either (1) greatly enlarge the manually annotated data set for training by incorporating many thousands of examples for each of the previously mentioned cases causing the low quality of FISH images or (2) better standardize the image acquisition practice in clinical routine to obtain higher quality images and (3) increase the number of pathologists in annotating the data. However, due to performing the FISH diagnostics on slides originating from FFPE material it might be difficult to obtain higher quality images. Training the pipelines on these samples will largely enhance its performance on future cases of a similar reduced overall quality. Nevertheless, even now our pipeline (trained on high quality FISH images) makes predictions on the HER2 amplification status of the tumor on the basis of these low quality FISH images demonstrating the general potential of deep learning on this task (**Suppl. Fig. 1**). It should be noticed that, in clinical practice, pathologists do not analyze every nucleus in a FISH image. Instead, a certain number of nuclei (at least 20) are selected and, in this process nuclei are excluded that are difficult to analyze, e.g. due to low image quality. Additionally, variations in the experimental setup among different pathology labs might result in different shape, structure and nuclei composition of the FISH images (e.g. used antibodies and fluorophores, tissue type, tissue preparation protocol, consideration of DAPI staining, fluorescence microscope type and parameters). Therefore, a customization of our pipeline, e.g. setting different thresholds, and additional training of both networks will be necessary to adapt the detection and classification pipeline to lab-specific conditions and lab-specific investigated tissue types in

order to automatize the HER2 amplification detection of tumors in other pathology labs.

Pathologists normally analyze the FISH slides directly under the fluorescence microscope. Due to shifting in z-dimension HER2 and CEN17 signals of one nucleus can be located which are not detectable in a single 2D position only. Since, however, a FISH image is only a 2D representation of the 3D space of the FISH slide only limited information is available for the nuclei classification for our pipeline potentially leading to false estimations of the HER2 gene amplification status of the corresponding tumor sample. Therefore, a deep learning application based on nuclei detection and classification on at least a stack of images representing the 3D space of the FISH slide will be largely superior compared to the 2D solution used in our study. Our pipeline is in principle able to make nuclei detections and classifications on videos which is under investigation in our lab. Future solutions should directly implement one-stage detectors or similar CNN architectures into the fluorescence microscope for instantly classifying the nuclei while the pathologist is observing it. A comparable solution was recently developed by Google Inc. for marking tumor areas in Hematoxylin and Eosin stained slides[27]. Alternatively, a fully automated software solution recording all layers and positions of a FISH slide as large data input for the deep learning-based nuclei detection and classification might be used.

Competing Interests

No

Authors' contributions

FZ, WdB and PH wrote the manuscript. FZ, WdB and PH designed the study. FZ, WdB, MW, RM and TW planed and conducted the bioinformatics data preparation and analysis. SZ and DEA generated the FISH data. KF, SZ, DEA and GB performed the pathological analysis of the data. KF, DEA, IR and GB supervised the project and assisted in the writing of the manuscript.

Acknowledgments

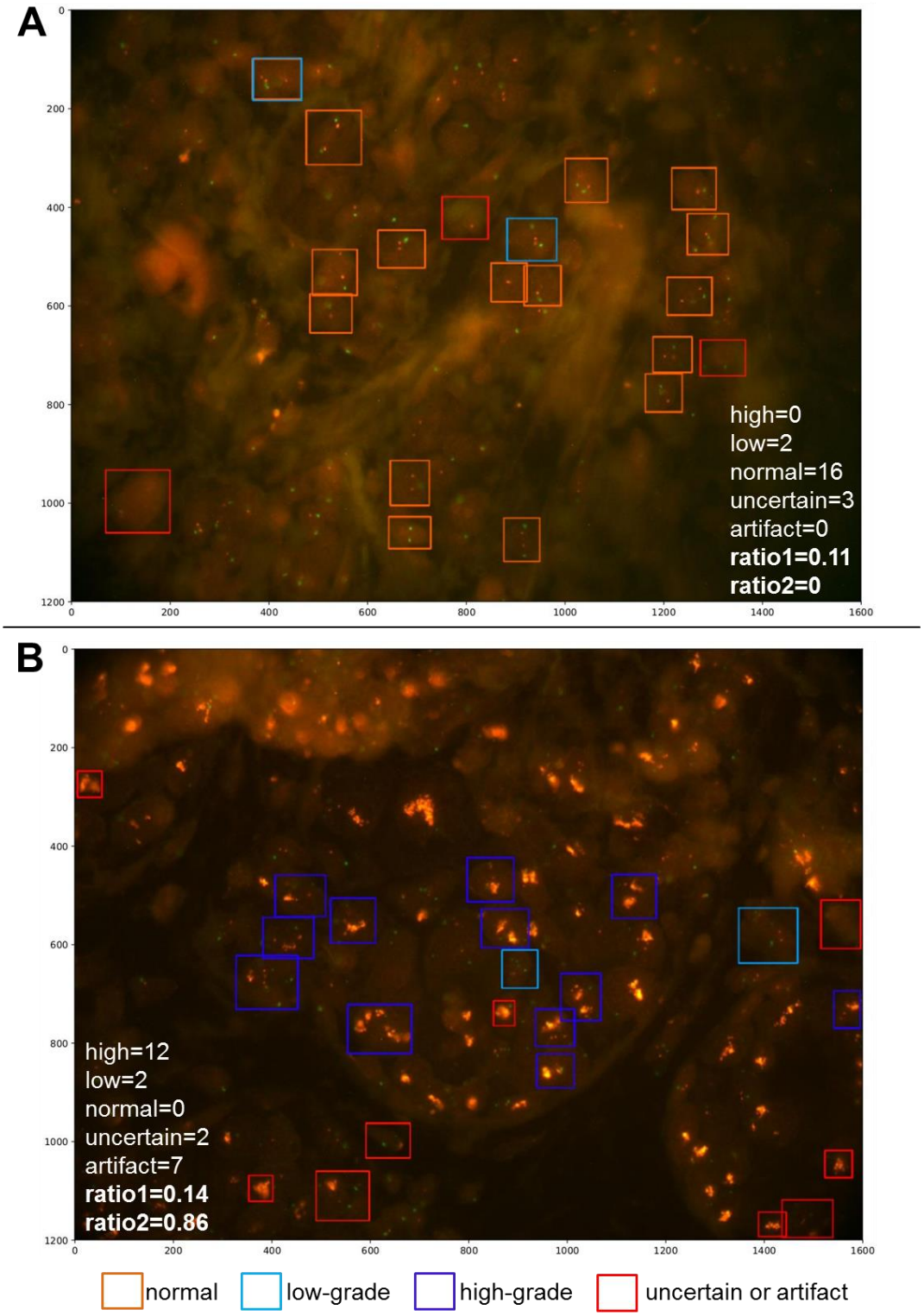
We are thankful to the Machine Learning Community (MLC) Dresden for helpful input, comments and inspiration. We also thank Regina Pohlers for technical assistance in conducting the FISH experiments and image recording.

References

1. Wolff, A. C. *et al.* American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *J. Clin. Oncol.* **25,** 118–45 (2007).
2. Rüschoff, J. *et al.* HER2 diagnostics in gastric cancer—guideline validation and development of standardized immunohistochemical testing. *Virchows Arch.* **457,** 299–307 (2010).
3. Simon, R. *et al.* Patterns of her-2/neu amplification and overexpression in primary and metastatic breast cancer. *J. Natl. Cancer Inst.* **93,** 1141–6 (2001).
4. Vincent-Salomon, A. *et al.* HER2 status of bone marrow micrometastasis and their corresponding primary tumours in a pilot study of 27 cases: a possible tool for anti-HER2 therapy management? *Br. J. Cancer* **96,** 654–659 (2007).
5. Slamon, D. J. *et al.* Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science* **235,** 177–82 (1987).
6. Slamon, D. J. *et al.* Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science* **244,** 707–12 (1989).
7. Ross & Fletcher. The HER-2/neu Oncogene in Breast Cancer: Prognostic Factor, Predictive Factor, and Target for Therapy. *Oncologist* **3,** 237–252 (1998).
8. Romond, E. H. *et al.* Trastuzumab plus Adjuvant Chemotherapy for Operable HER2 positive Breast Cancer. *N. Engl. J. Med.* **353,** 1673–1684 (2005).
9. Marty, M. *et al.* Randomized Phase II Trial of the Efficacy and Safety of Trastuzumab Combined With Docetaxel in Patients With Human Epidermal Growth Factor Receptor 2–Positive Metastatic Breast Cancer Administered

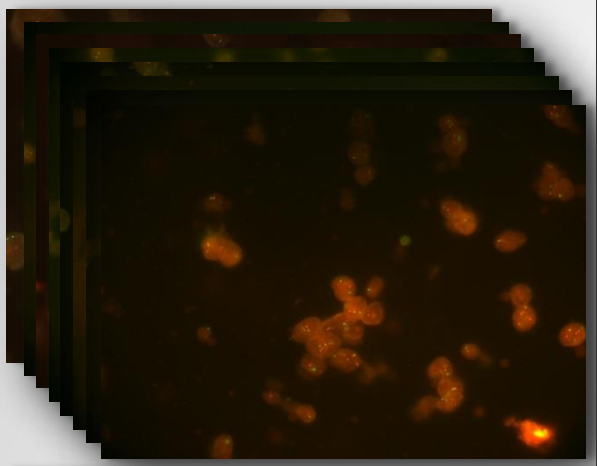
As First-Line Treatment: The M77001 Study Group. *J. Clin. Oncol.* **23,** 4265–4274 (2005).

10. Slamon, D. J. *et al.* Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2. *N. Engl. J. Med.* **344,** 783–792 (2001).
11. Albarello, L., Pecciarini, L. & Doglioni, C. HER2 Testing in Gastric Cancer. *Adv. Anat. Pathol.* **18,** 53–59 (2011).
12. Gutierrez, C. & Schiff, R. HER2: biology, detection, and clinical implications. *Arch. Pathol. Lab. Med.* **135,** 55–62 (2011).
13. Wolff, A. C. *et al.* American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Arch. Pathol. Lab. Med.* **131,** 18–43 (2007).
14. Lin, T.-Y. *et al.* Feature Pyramid Networks for Object Detection. (2016).
15. Bright, D. S. & Steel, E. B. Two-dimensional top hat filter for extracting spots and spheres from digital images. *J. Microsc.* **146,** 191–200 (1987).
16. Olivo-Marin, J.-C. Extraction of spots in biological images using multiscale products. *Pattern Recognit.* **35,** 1989–1996 (2002).
17. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).
18. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. 1097–1105 (2012).
19. Xu, Y. *et al.* Deep learning of feature representation with multiple instance learning for medical image analysis. in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 1626–1630 (IEEE, 2014). doi:10.1109/ICASSP.2014.6853873
20. Gudla, P. R., Nakayama, K., Pegoraro, G. & Misteli, T. SpotLearn: Convolutional Neural Network for Detection of Fluorescence In Situ Hybridization (FISH) Signals in High-Throughput Imaging Approaches. *Cold Spring Harb. Symp. Quant. Biol.* **82,** 57–70 (2017).
21. Pardo, E., Morgado, J. M. T. & Malpica, N. Semantic segmentation of mFISH images using convolutional networks. (2018). doi:10.1002/cyto.a.23375
22. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal Loss for Dense Object Detection. (2017).
23. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. (2015).
24. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal Loss for Dense Object Detection. (2017).
25. Keras-RetinaNet. Available at: https://github.com/fizyr/keras-retinanet.
26. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33,** 159–74 (1977).
27. Google AI Blog: An Augmented Reality Microscope for Cancer Detection. Available at: https://ai.googleblog.com/2018/04/an-augmented-reality-microscope.html. (Accessed: 18th May 2018)
28. Chollet, F. Keras. (2015). Available at: https://github.com/keras-team/keras.
29. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. (2016).
30. labelImg. Available at: https://github.com/tzutalin/labelImg.

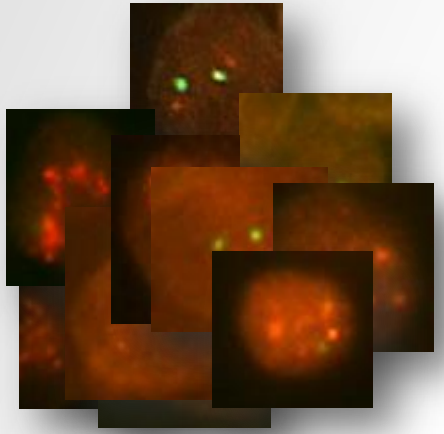


Supplemental Figure 1. Two examples of the application of our pipeline on FISH images of low quality. In **(A)** a normal stage was detected and corresponds to the decision of a pathologist. In **(B)** a high-grade stage was detected which also corresponds to the pathologists decision on the FISH image. Numerous nuclei have not been detected in both images indicating the limitations of our pipeline (RetinaNet-1) on FISH images of low quality. Training on a large set of FISH images of low quality would enhance the accuracy in detecting most nuclei.

TRAINING DATA



Manually annotated FISH images (n~300)



Manually annotated nuclei from FISH images (n~300)

TRAINING PIPELINE

Two RetinaNets for nucleus and FISH signal detection and classification

RetinaNet-1
Learns nucleus localization and classification

Data was split into 90% training data and 10% validation data and testing was done using 57 FISH images

RetinaNet-2
Learns FISH signal detection and classification

Training on 1xGPU
Duration = 2-3 days

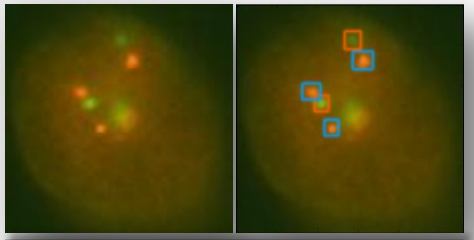
PREDICITION

Nucleus detection and classification via RetinaNet-1

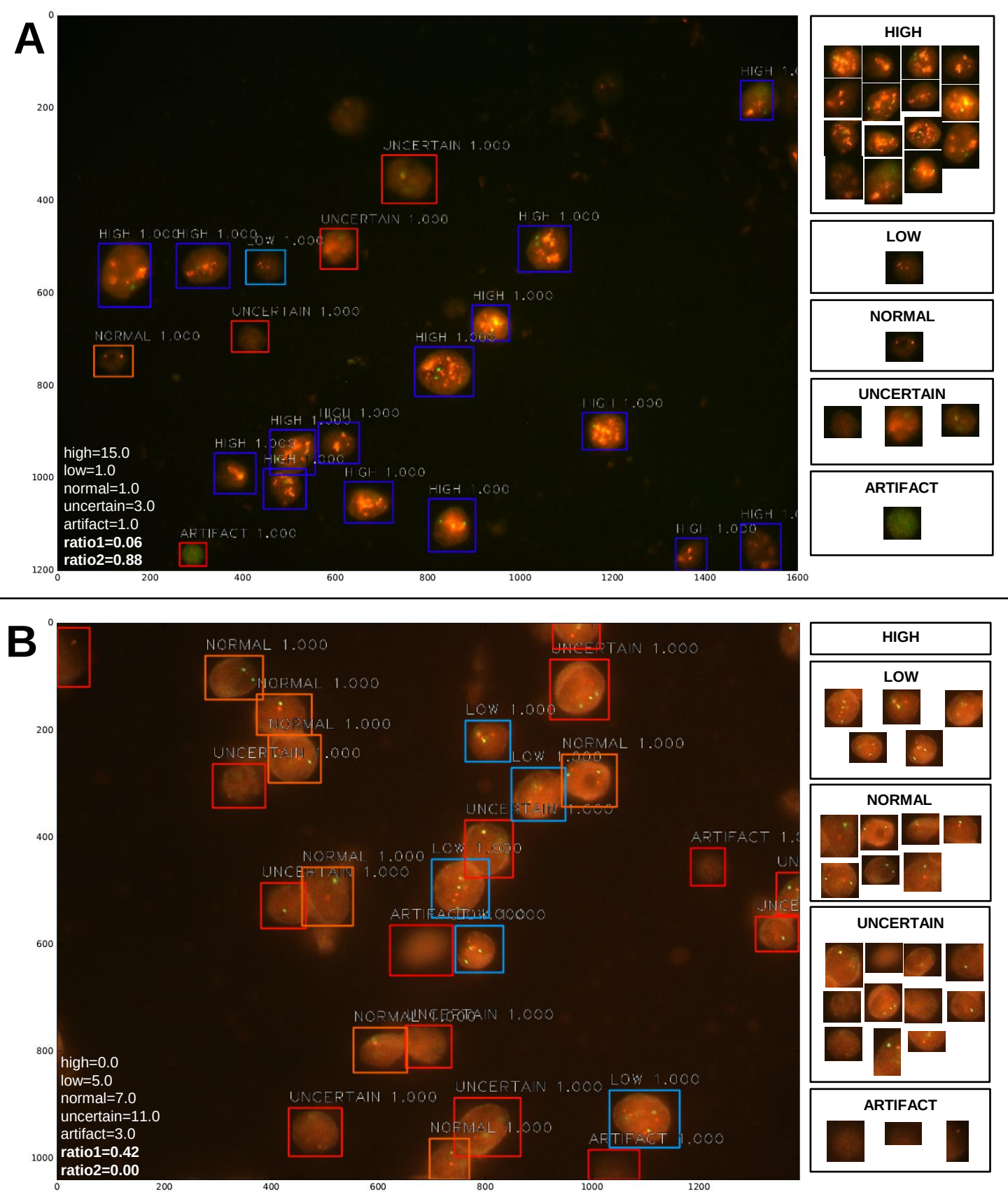
RetinaNet-1 report

+

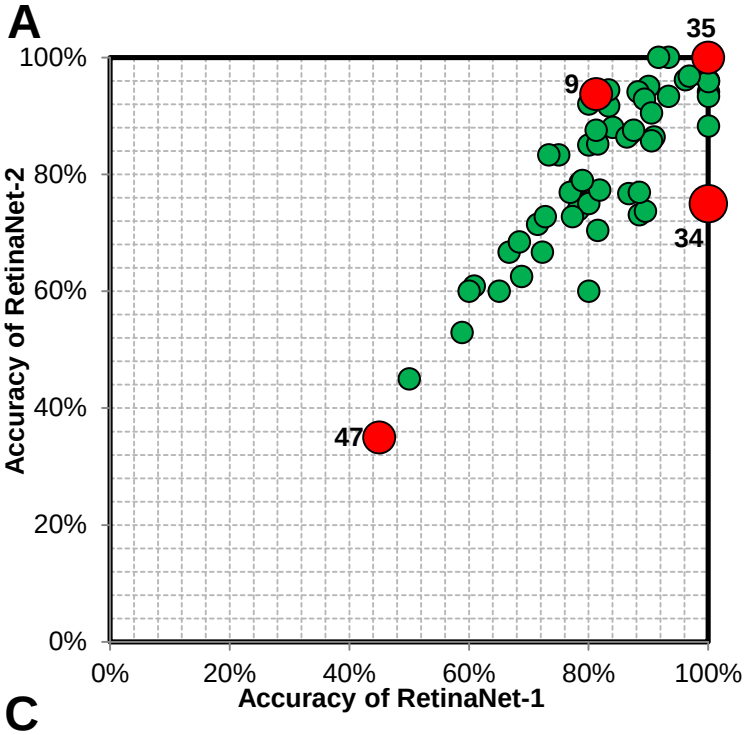
Signal classification and counting in each nucleus via RetinaNet-2



RetinaNet-2 report


A
high=15.0
low=1.0
normal=1.0
uncertain=3.0
artifact=1.0
ratio1=0.06
ratio2=0.88
HIGH
LOW
NORMAL
UNCERTAIN
ARTIFACT
B
high=0.0
low=5.0
normal=7.0
uncertain=11.0
artifact=3.0
ratio1=0.42
ratio2=0.00
HIGH
LOW
NORMAL
UNCERTAIN
ARTIFACT

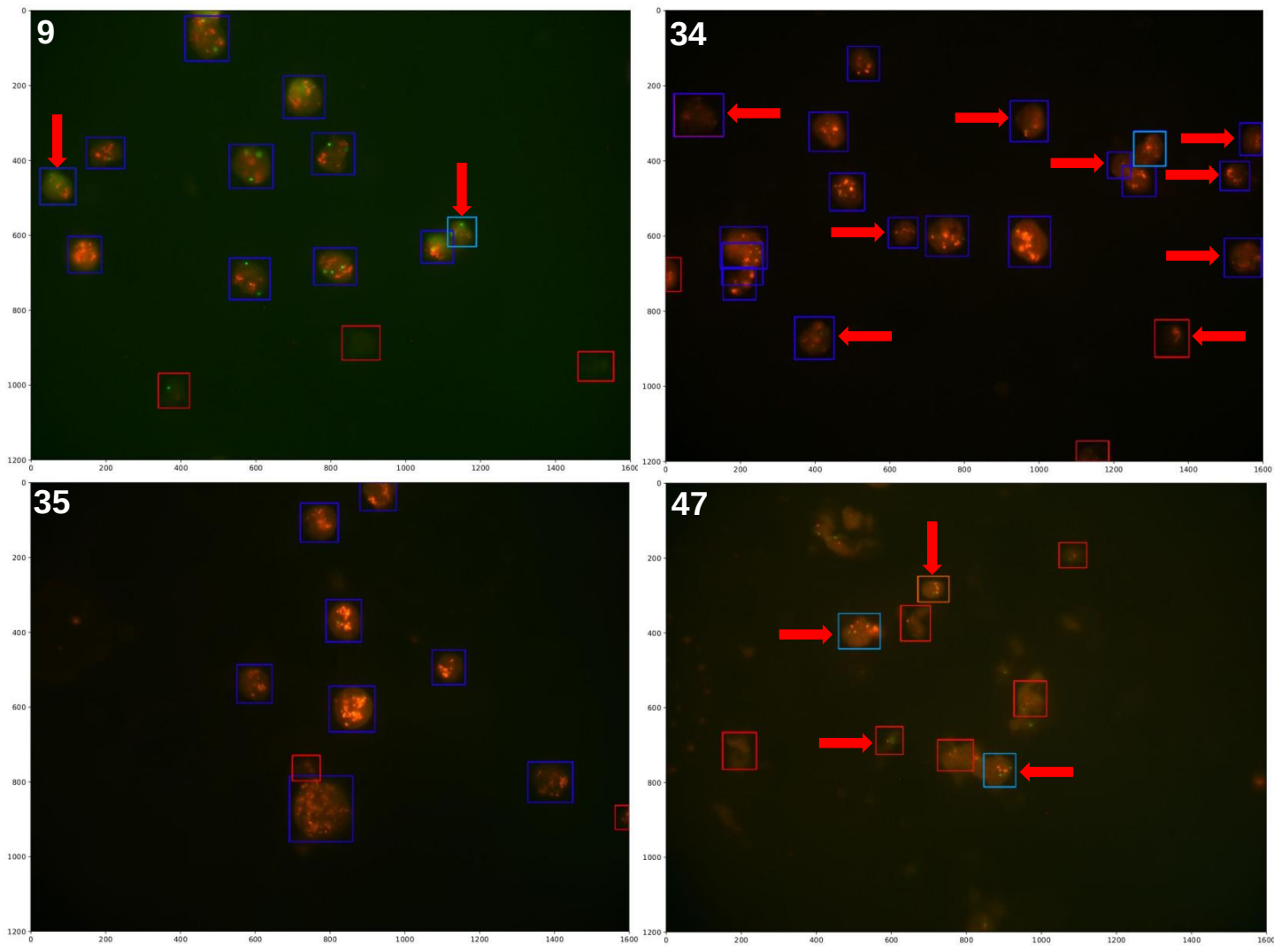
A



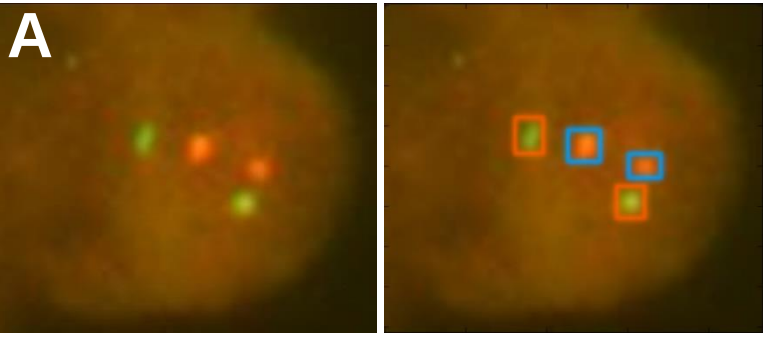
B

	Image	9	34	35	47
Normal	RetinaNet-1	0	0	0	1
	RetinaNet-2	0	1	0	1
	Pathologist	0	0	0	4
Low-grade	RetinaNet-1	2	1	0	2
	RetinaNet-2	0	3	0	1
	Pathologist	0	0	0	4
High-grade	RetinaNet-1	10	18	8	0
	RetinaNet-2	11	13	8	2
	Pathologist	12	18	8	0
Indefinable	RetinaNet-1	3	4	2	6
	RetinaNet-2	4	6	2	5
	Pathologist	4	2	2	12
Correct	RetinaNet-1	13	20	10	9
	RetinaNet-2	15	15	10	7
	Pathologist	16	20	10	20
	RetinaNet-1	81%	100%	100%	45%
	RetinaNet-2	94%	75%	100%	35%
Difference		**-13%**	**25%**	**0%**	**10%**

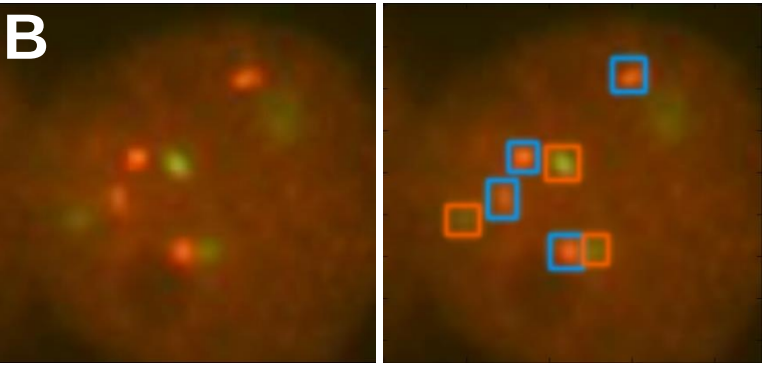
C



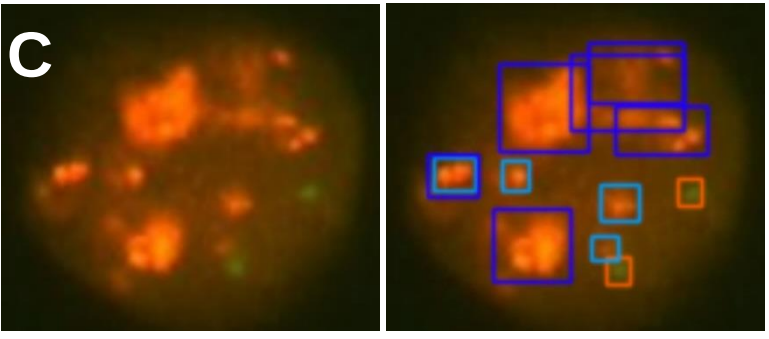
Same classification



HER2/CEN17: 1.00
HER2: 2x
HER2-Cl: 0x
CEN17: 2x
RetinaNet-1: normal
RetinaNet-2: normal
Pathologist: normal

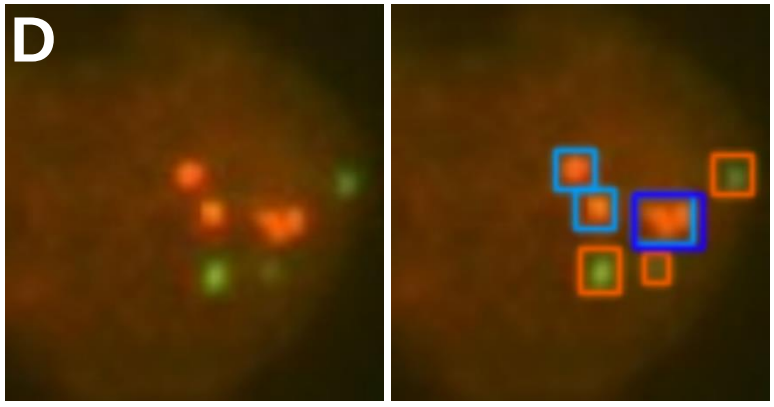


HER2/CEN17: 1.33
HER2: 4x
HER2-Cl: 0x
CEN17: 3x
RetinaNet-1: low
RetinaNet-2: low
Pathologist: low

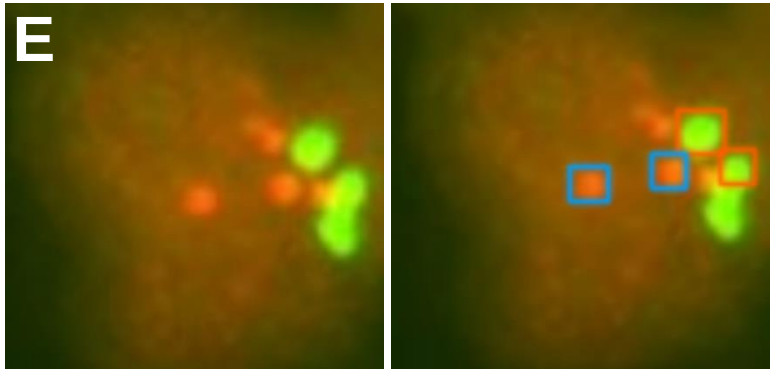


HER2/CEN17: 10.00
HER2: 4x
HER2-Cl: 6x
CEN17: 2x
RetinaNet-1: high
RetinaNet-2: high
Pathologist: high

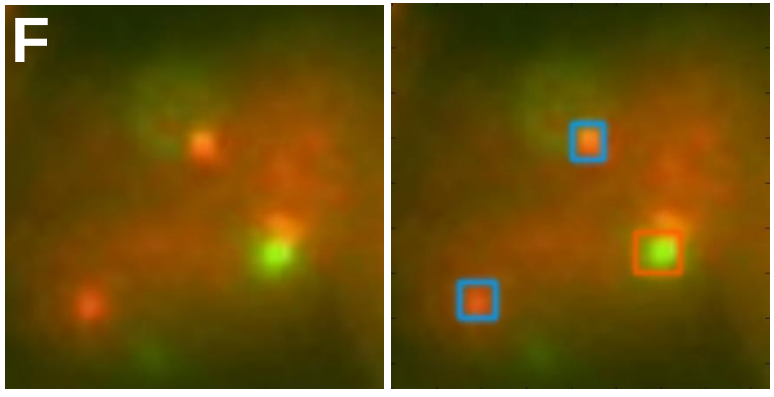
Different classification



HER2/CEN17: 10.00
HER2: 3x
HER2-Cl: 1x
CEN17: 3x
RetinaNet-1: low
RetinaNet-2: high
Pathologist: low



HER2/CEN17: 1.00
HER2: 2x
HER2-Cl: 0x
CEN17: 2x
RetinaNet-1: low
RetinaNet-2: normal
Pathologist: low



HER2/CEN17: 2.00
HER2: 2x
HER2-Cl: 0x
CEN17: 1x
RetinaNet-1: uncertain
RetinaNet-2: normal
Pathologist: low