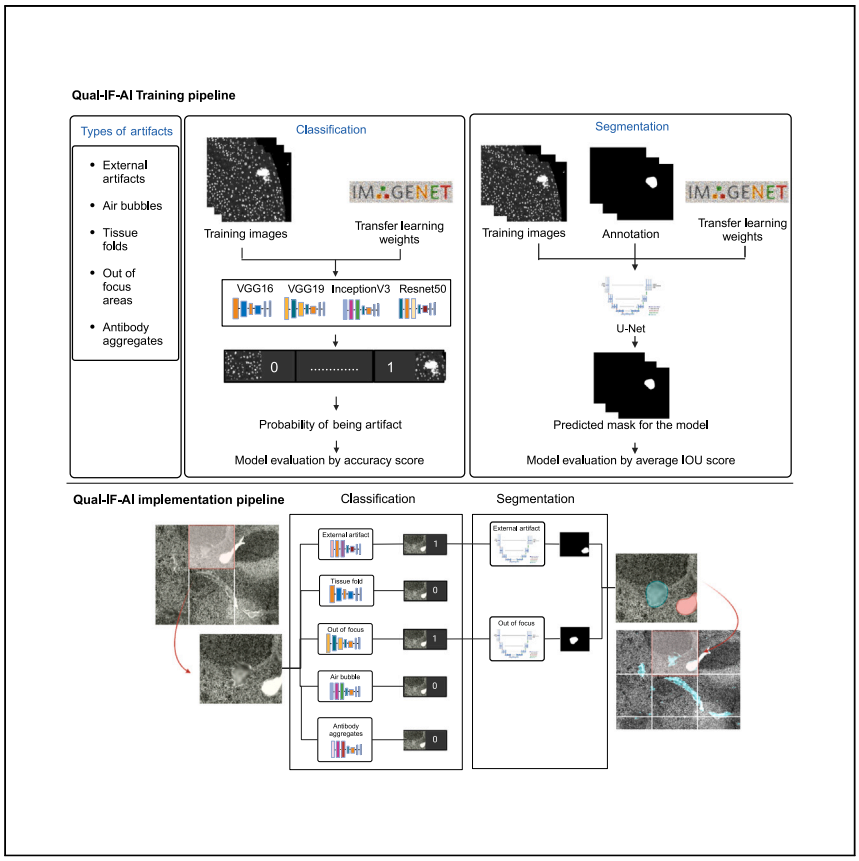
Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

Quality control of immunofluorescence images using artificial intelligence



Andhari et al. introduce QUALIFAI, an advanced AI-driven tool designed for the automated detection of common artifacts in fluorescent imaging across various multiplexing platforms. This innovative tool significantly enhances the accuracy and reliability of spatial proteomics analysis, offering a valuable asset for researchers in diverse scientific fields.

Madhavi Dipak Andhari, Giulia Rinaldi, Pouya Nazari, ..., Frederik De Smet, Francesca Bosisio, Asier Antoranz

asier.antoranzmartinez@kuleuven.be

Highlights

A deep-learning-based tool, QUALIFAI, is developed and demonstrated

QUALIFAI automates artifact detection in fluorescent imaging

QUALIFAI excels in classification and segmentation across multiplexing platforms

QUALIFAI improves reliability in spatial proteomics multiplexing analysis

Andhari et al., Cell Reports Physical Science 5, 102220
October 16, 2024 © 2024 The Author(s).
Published by Elsevier Inc.
https://doi.org/10.1016/j.xcrp.2024.102220

Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

Quality control of immunofluorescence images using artificial intelligence

Madhavi Dipak Andhari,[1] Giulia Rinaldi,[2] Pouya Nazari,[3] Johanna Vets,[1] Gautam Shankar,[1] Nikolina Dubroja,[1] Tessa Ostyn,[1] Maxime Vanmechelen,[3,4] Brecht Decraene,[3,5,6] Alexandre Arnould,[2] Willem Mestdagh,[2] Bart De Moor,[2] Frederik De Smet,[3] Francesca Bosisio,[1,7,8] and Asier Antoranz[1,8,9,*]

SUMMARY

Fluorescent imaging has revolutionized biomedical research, enabling the study of intricate cellular processes. Multiplex immunofluorescent imaging has extended this capability, permitting the simultaneous detection of multiple markers within a single tissue section. However, these images are susceptible to a myriad of undesired artifacts, which compromise the accuracy of downstream analyses. Manual artifact removal is impractical given the large number of images generated in these experiments, necessitating automated solutions. Here, we report the development of QUALIFAI (quality control of immunofluorescence images using artificial intelligence), a deep-learning-based tool designed to automate the identification of common artifacts in fluorescent imaging. We demonstrate the utility of QUALIFAI in detecting five of the most common types of artifacts in fluorescent imaging, achieving over 90% classification accuracy and a more than 0.6 intersection over union score across all artifact types in a variety of multiplexing platforms. Finally, we show how the implementation of QUALIFAI leads to more reliable results in downstream analysis.

INTRODUCTION

Fluorescent imaging has become an indispensable tool in various biological disciplines, enabling the study of protein dynamics, cellular functions, and neuronal activity, among others.[1] With the recent development of multiplex immunofluorescent imaging, the simultaneous detection and quantification of multiple markers in a single tissue section has become feasible.[2] The analysis of the images generated with multiplex immunofluorescent imaging requires an extensive pipeline of pre-processing steps, including illumination correction, image registration (in the case of cyclic multiplexing methods), cell segmentation, and autofluorescence subtraction, before extracting quantitative single-cell measurements from the images.[3–5] It is crucial to ensure that the input images are of high quality to obtain reliable single-cell measurements and meaningful biological insights. Undesired artifacts, such as external artifacts (dust particles, hair shafts, fibers), air bubbles, antibody aggregates, out-of-focus (OOF) areas, and tissue folds are common not only in hematoxylin and eosin (H&E) images but also in fluorescent microscopy.[6–9] These artifacts can originate from various sources. Dust particles, hair shafts, fibers, and air bubbles can get trapped during sample preparation or cover slipping. Denatured antibodies can lead to antibody aggregation artifacts in the images. Misalignment of the coverslip, inaccurate calculation of the focus depth, or tissue detachment during scanning can give rise to OOF areas. Suboptimal tissue embedding or tissue sectioning can lead to tissue folding or even complete tissue detachment from the slide.[9] These

[1]Translational Cell and Tissue Research Unit, Department of Imaging and Pathology, KU Leuven, Leuven, Belgium

[2]Dynamic Systems, Signal Processing and Data Analysis (STADIUS), KU Leuven, Leuven, Belgium

[3]Laboratory for Precision Cancer Medicine, Translational Cell and Tissue Research Unit, KU Leuven, Leuven, Belgium

[4]Department of Medical Oncology, University Hospitals Leuven, Leuven, Belgium

[5]Department of Neurosurgery, University Hospitals Leuven, Leuven, Belgium

[6]Department of Neurosciences, Experimental Neurosurgery and Neuroanatomy Research Group, Leuven Brain Institute (LBI), KU Leuven, Leuven, Belgium

[7]Department of Imaging and Pathology, University Hospitals Leuven, Leuven, Belgium

[8]These authors contributed equally

[9]Lead contact

*Correspondence: asier.antoranzmartinez@kuleuven.be

https://doi.org/10.1016/j.xcrp.2024.102220

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

artifacts can significantly impact quantitative image analysis by introducing false positives and false negatives.[10,11]

In traditional experiments, the limited number of markers that could be stained on a tissue sample resulted in a manageable number of image layers per experiment. Artifact detection in these experiments relied on manual annotation or bespoke classical algorithms, such as Laplacian variance for detecting blur[12] or manually tuned thresholds[13] for detecting saturation artifacts. These methods are time consuming and labor intensive and may not adapt well to unseen data. With the widespread adoption of multiplexing, a single tissue slide can now be stained for tens of markers. This exponential increase in the number of markers per sample has significantly increased the number of acquired image layers, making manual annotations unfeasible due to the sheer volume of data generated. Therefore, the rising popularity and adoption of multiplexing techniques are indeed exacerbating the challenges associated with manual annotation. Moreover, manual annotation is subjective and contingent on high intra- and inter-user variability.[14,15] To overcome these challenges, automated tools for detecting and managing artifacts in an unbiased and reproducible way are a necessity. Although some automated tools for detecting artifacts in H&E images exist,[16–19] limited tools are available for fluorescent microscopy. CellProfiler[7] can identify artifacts in fluorescent microscopy, but its sensitivity to input parameters makes it difficult to train robust models that generalize well to diverse datasets. Therefore, the set of tools currently available in the literature falls short in terms of the diversity of identifiable artifacts, ease of use, and overall performance in the context of fluorescent microscopy images.

The increase in popularity of deep learning (DL), and particularly convolutional neural networks (CNNs), for image analysis is starting to show their applicability for artifact detection in medical images.[20,21] One of the main bottlenecks of these approaches is that training models with high accuracy requires pixel-level annotations for thousands of images.[21] Transfer learning can partially address this issue by leveraging information learned on other, bigger datasets such as ImageNet,[22] a popular visual database containing over 14 million manually annotated images, which has been used to train a variety of DL architectures. Several models have been shown to perform better with transfer learning in classifying medical images for various tasks.[23–25] For example, a study conducted by Guan et al. demonstrated the effectiveness of transfer learning in differentiating papillary thyroid carcinoma from benign thyroid nodules using cytological images. By fine-tuning pre-trained VGG16 and Inception V3 models on relatively small datasets, they achieved a remarkable accuracy in identifying papillary thyroid carcinoma.[25]

Building upon these ideas, we have developed QUALIFAI (quality control of immunofluorescence images using artificial intelligence, pronounced "qualify"), a DL-based quality control (QC) tool for the detection and management of five distinct types of artifacts: OOF areas, external artifacts, tissue folds, antibody aggregates, and air bubbles. QUALIFAI is designed as a two-tier DL system incorporating both classification and segmentation tasks to identify and localize artifacts in fluorescence images. QUALIFAI first classifies tissue regions into having the presence or absence of given artifacts and then delineates (segments) artifacts if the initial prediction is positive. The tool consists of five different pre-trained models, each optimized for a specific type of artifact, ensuring high accuracy. To select the optimal DL architectures, we benchmarked several classification models (VGG16, VGG19, Inception V3, and ResNet50)[26] during the training phase for each artifact type. The best-performing models were then employed as the backbone for training the segmentation models using a U-Net architecture.[27]

Table 1. Overview of datasets used for training QUALIFAI models

Acquisition technology	MILAN	COMET (Lunaphore)	CODEX (Akoya)
Species	human	human	human
Tissue type	healthy lung, appendix, brain, liver, and pathologies including metastatic melanoma, pancreatic ductal adenocarcinoma, colorectal cancer, diffuse large B cell lymphoma, autoimmune hepatitis, kidney transplant, glioblastoma	healthy tonsil, pan cancer	metastatic melanoma
Slide type	WSI (whole-slide imaging) , TMA, CNB (core needle biopsy)	WSI, TMA	WSI
Acquisition instrument	Zeiss Axioscan Z1	Lunaphore COMET	Keyence BZ-X800
Resolution (μm/px)	0.65	0.22	0.37

Note: three different datasets from three different acquisition technologies were used to train and evaluate QUALIFAI. To increase robustness and reduce overfitting, these datasets included different types of tissues and slides. The images were also acquired with different instruments at different resolutions.

Overall, in this work, we present a comprehensive QC algorithm for artifact detection and management in fluorescent images. We demonstrate that QUALIFAI effectively detects five common artifact types—air bubbles, tissue folds, external artifacts, OOF areas, and antibody aggregates—achieving over 90% classification accuracy and a more than 0.6 intersection over union (IoU) score across various multiplexing platforms. QUALIFAI is available at https://github.com/TCWO/QualIFAI, where, for simplicity, a single wrapper function is provided, which sequentially implements all the models. Additionally, for enhanced flexibility, individual models are supplied.

RESULTS

Performance evaluation of QUALIFAI in the classification and segmentation of artifacts

The objective of QUALIFAI is to identify areas on fluorescence images affected by diverse types of artifacts, including OOF areas, air bubbles, tissue folds, antibody aggregates, and external artifacts. This goal is achieved through a two-tier DL model that first classifies tiles containing artifacts and, upon positive prediction, segments the area covered by the artifacts with pixel-level precision. QUALIFAI was trained and validated using datasets from three distinct fluorescence-based multiplexing technologies: multiple iterative labeling by antibody neodeposition (MILAN),[3] COMET,[28] and CODEX,[29] which have varying resolutions (0.65, 0.22, and 0.37 μm/px, respectively) and include a diverse range of slide and tissue types (Table 1). For each artifact type, the classification performance in four different DL architectures (VGG16, VGG19, Inception V3, and ResNet50) was compared. For each artifact type, the architecture that yielded the highest accuracy on the validation dataset was chosen. These architectures were also used as backbones for the segmentation of the respective artifacts.

Classification

The performance of various classification architectures was evaluated using accuracy values (Table 2). For all three multiplexing technologies, QUALIFAI achieved a classification accuracy above 90% across all artifact types. Specifically, air bubbles, tissue folds, OOF regions, and antibody aggregates obtained accuracy values of 94%, 96%, 96%, and 93%, respectively, for the VGG19 architecture. For external artifacts, the VGG16 architecture achieved the highest accuracy of 97%. This performance was consistent across images acquired with different technologies (Table 2). Additionally, we evaluated the robustness of the model by visualizing receiver operating characteristic (ROC) curves and t-distributed stochastic neighbor embedding (t-SNE) plots generated from the output features from the second-to-last layer of the best-performing classification models. The clear separation of the different artifact categories indicates that the models have learned meaningful features to effectively distinguish images with artifacts from those without (Figure 1A).[30]

Table 2. Performance of classification models

Classification	Total number of training images	Total number of testing images	Total number of validation images	VGG16	VGG19	Inception V3	ResNet50
Total number of layers	–	–	–	22	25	314	178
Total number of parameters (million)	–	–	–	31	36	73	90
Accuracy							
External artifacts							
MILAN	393	125	117	0.96	0.96	0.7	0.92
COMET	430	152	157	0.97	0.97	0.7	0.88
CODEX	60	24	31	1	0.90	0.7	0.9
TOTAL	883	301	305	0.97[a]	0.96	0.7	0.90
Air bubbles							
MILAN	316	104	99	0.96	0.98	0.87	0.90
COMET	70	26	37	0.84	0.86	0.7	0.68
CODEX	73	35	36	0.92	0.92	0.92	0.78
TOTAL	459	165	172	0.93	0.94[a]	0.84	0.83
Tissue folds							
MILAN	342	117	116	0.93	0.97	0.96	0.9
COMET	71	37	39	0.79	0.83	0.90	0.88
CODEX	274	93	90	0.93	1	0.92	0.91
TOTAL	687	247	245	0.91	0.96[a]	0.94	0.90
Out-of-focus areas							
MILAN	341	126	101	0.96	0.98	0.88	0.91
COMET	726	229	244	0.99	0.97	0.93	0.95
CODEX	1,403	469	479	0.95	0.95	0.92	0.94
TOTAL	2,470	824	824	0.96	0.96[a]	0.92	0.94
Antibody aggregates							
MILAN	285	109	99	0.76	0.91	0.77	0.84
COMET	68	29	32	0.69	1	0.8	0.57
CODEX	747	234	243	0.87	0.93	0.85	0.90
TOTAL	1,100	372	374	0.83	0.93[a]	0.82	0.86

Note: various architectures were evaluated for each artifact type across the various acquisition technologies. Results are reported as the classification accuracy achieved on the validation dataset.

[a]Accuracy achieved by the best-performing architecture of each category.

Creating a comprehensive dataset encompassing all possible artifact types, particularly in the first stages of training, is unfeasible. Therefore, on top of transfer learning from ImageNet, QUALIFAI was trained sequentially following an active learning framework (Figure S1A; see the experimental procedures). Considered a hyperparameter, the number of neighbors on the template projection was optimized, and we found that using 20 neighbors (K) yielded significantly better performance compared to using 5 or 15 neighbors (Wilcoxon, $p < 0.05$) (Figure S1B; see the experimental procedures). This training strategy is increasingly favored over random selection of training images due to several compelling advantages, such as training efficiency[31] and class imbalance mitigation,[32] by focusing on ambiguous instances that are most informative for the model.[33] This targeted approach to data selection led to a more rapid and effective convergence of the model by achieving 95% accuracy after 3 cycles, whereas random learning achieved 91% accuracy with the same number of cycles (Wilcoxon $p < 0.01$ after three independent runs). In fact, random learning required five iterations to reach the same accuracy level (Figure S1C). These findings highlight the efficacy of active learning in optimizing model performance and efficiency in the training of DL models.

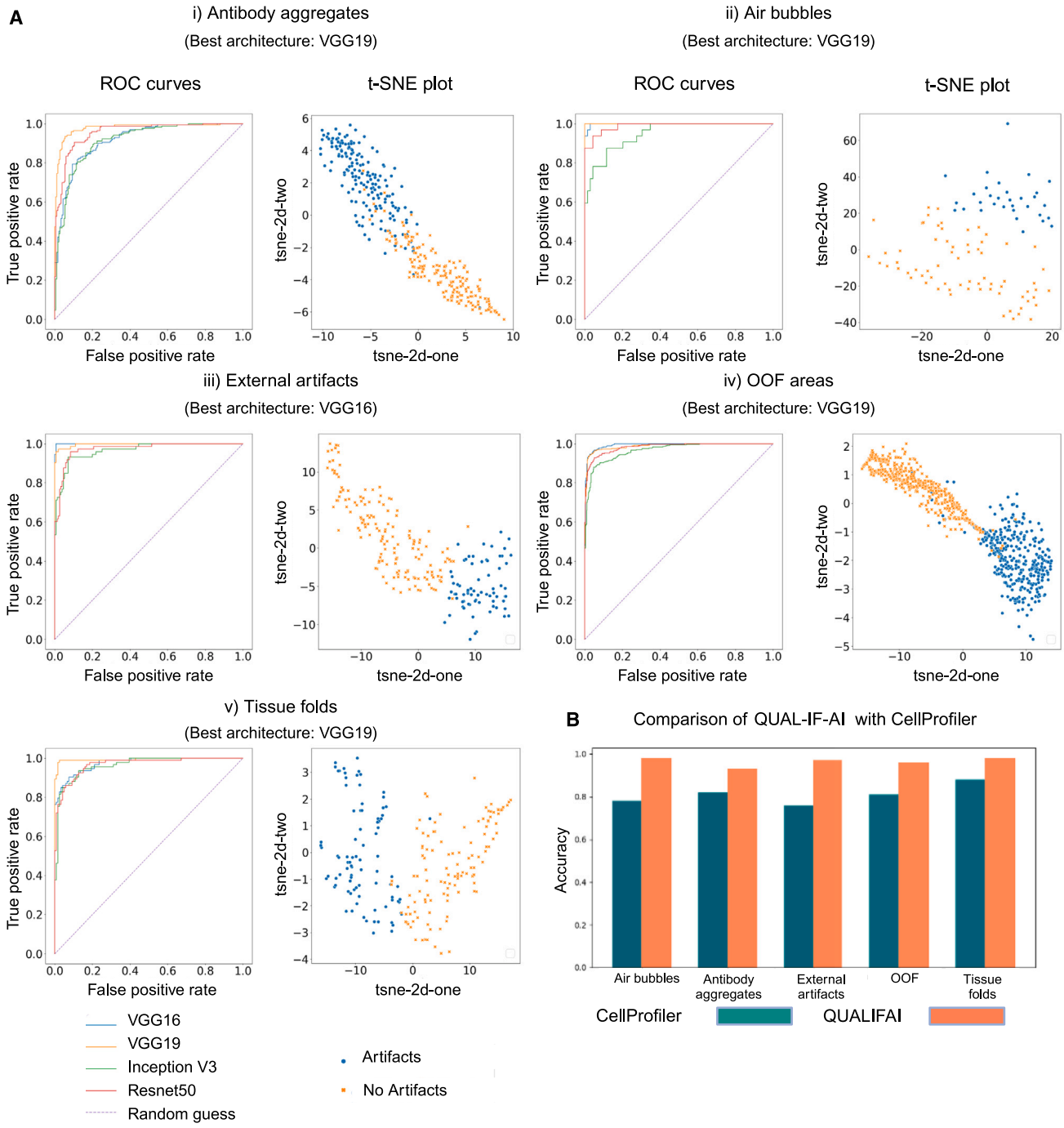


Figure 1. QUALIFAI classification model performance

(A) QUALIFAI classification model performance. ROC curves (left) display the performance of classification models on the validation dataset of distinct artifacts: (i) antibody aggregates, (ii) air bubbles (iii) external artifacts, (iv) OOF areas, and (v) tissue folds. The ROC curves (left) delineate the model performance, with the VGG16 architecture represented in blue, VGG19 in orange, Inception V3 in green, and ResNet50 in red. The purple dashed line denotes the performance of random guessing. The t-SNE plots (right) display the separation of images with artifacts and without artifacts. These plots were generated for visualization by applying dimensionality reduction on the tensor layer generated before the final classification output layer. In the t-SNE plot, each dot corresponds to an image from the validation dataset. Images categorized as artifacts are depicted in blue, while those classified as artifact free are represented in orange.

(B) Comparison of QUALIFAI with CellProfiler. Bar plots showing the classification accuracy obtained by CellProfiler (green) and QUALIFAI (orange) for different artifact types.

Table 3. Performance of segmentation models

Segmentation					
	External artifacts	Air bubbles	Tissue folds	Out-of-focus areas	Antibody aggregates
U-Net encoding backbone architecture	U-Net + VGG16	U-Net + VGG19	U-Net + VGG19	U-Net + VGG19	U-Net + VGG19
Total number of layers	66	69	69	69	69
Total number of parameters (million)	23	29	29	29	29
IoU (mean ± SD)					
MILAN	0.83 ± 0.15	0.83 ± 0.11	0.71 ± 0.20	0.57 ± 0.25	0.67 ± 0.2
COMET	0.67 ± 0.14	0.77 ± 0.2	0.72 ± 0.12	0.75 ± 0.21	0.75 ± 0.03
CODEX	0.56 ± 0.14	0.87 ± 0.08	0.75 ± 0.16	0.74 ± 0.19	0.63 ± 0.14
TOTAL	0.75 ± 0.17	0.82 ± 0.16	0.74 ± 0.18	0.73 ± 0.21	0.65 ± 0.17

Note: for each artifact type, IoU (mean ± SD) scores on the validation dataset are reported for each acquisition technology. SD, standard deviation.

Next, we compared the performance of QUALIFAI with CellProfiler, the leading automated QC tool available. CellProfiler is an open-source software designed for generating image-based measurements and training machine learning models to classify images into various phenotypes. Previous research has demonstrated the potential of CellProfiler in identifying fluorescent images with OOF areas and saturation debris.[7,8] Our results showed a clear superiority of QUALIFAI over CellProfiler across all artifact categories. The largest difference was observed in the classification of external artifacts, where QUALIFAI achieved an accuracy of 97%, compared to CellProfiler's 76% accuracy, on the validation dataset. Similarly, for the classification of air bubbles, antibody aggregates, OOF areas, and tissue folds, CellProfiler achieved accuracies of 78%, 82%, 81%, and 88%, respectively. In contrast, QUALIFAI achieved accuracies of 94%, 93%, 96%, and 96% for these categories, respectively (Figure 1B).

Segmentation

The top-performing architectures in the classification module, VGG16 for external artifacts and VGG19 for air bubbles, tissue folds, antibody aggregates, and OOF areas, were utilized as the U-Net backbone for training the respective segmentation models. Generally, an IoU score of 0.5 is acceptable for artifact detection in the field of biomedical imaging.[18,34–36] We achieved a mean IoU > 0.6 on the validation cohort for each artifact type. Specifically, the model achieved mean IoU scores of 0.82 for air bubbles, 0.75 for external artifacts, 0.74 for tissue folds, 0.65 for antibody aggregates, and 0.73 for OOF areas (Table 3). A diverse set of examples illustrating the performance of QUALIFAI for different artifact categories is included in Figure 2. These examples illustrate the correspondence between the annotated and predicted masks. Furthermore, we examined the distribution of IoU scores across all images within different artifact categories in the validation cohort to evaluate the robustness of QUALIFAI. We observed that most cases have an IoU score above 0.6, with very few cases having an IoU below 0.5 (Figure 2F). Figure S2 shows the image with the low IoU for each artifact category. In these images, it can be observed that the model has identified the artifact; however, the contour is incorrectly delineated. Besides poor model performance, these results can also be explained by suboptimal human annotations in especially challenging cases, representing an imperfect gold standard for model training.

In summary, the translatability of QUALIFAI across different datasets was demonstrated by its consistent performance metrics when applied to various fluorescence-based multiplexing technologies and resolutions. Figure S3 illustrates an

CellPress
OPEN ACCESS

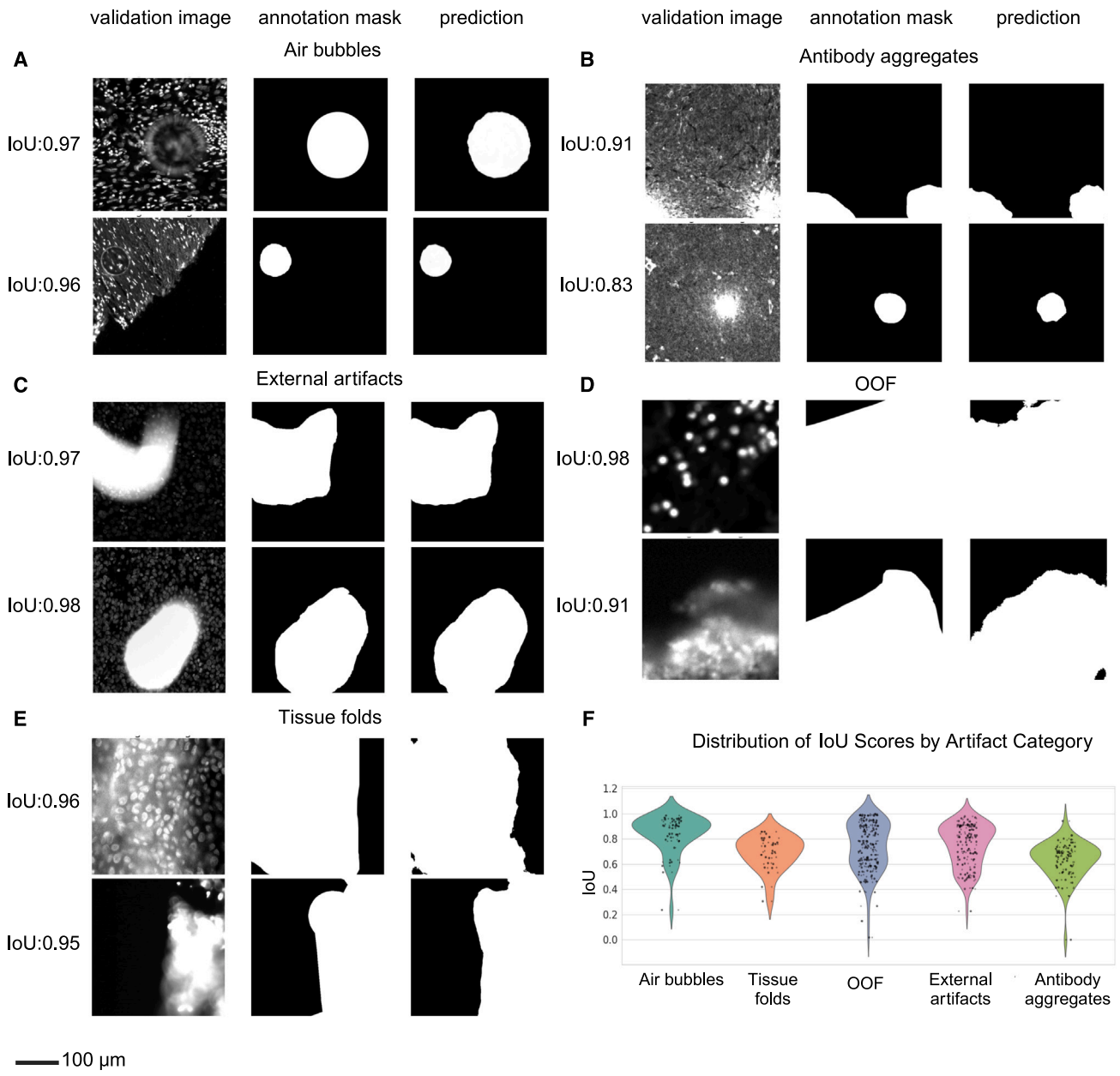


Figure 2. QUALIFAI segmentation model performance

(A–E) Representative segmentation masks as predicted by QUALIFAI for each artifact category: (A) air bubbles, (B) antibody aggregates, (C) external artifacts, (D) OOF areas, and (E) tissue folds. Each image depicts, from left to right, the image containing the artifact, the respective annotated mask, and the mask predicted by the segmentation model.

(F) Violin plot depicting the distribution of IoU scores across all images in the validation cohort for each artifact category. The x axis represents the different artifact categories, while the y axis represents the IoU score. Each dot indicates the IoU score achieved by an individual image in the validation cohort. The scale bar represents 100 µm

example of the final output of QUALIFAI on representative images. The classification and segmentation models showed robustness and adaptability, achieving high classification accuracy and segmentation IoU scores regardless of variation in input data due to different acquisition technologies. This indicates that QUALIFAI can effectively generalize to distinct types of fluorescence imaging data, making it a versatile tool for identifying and segmenting artifacts across diverse experimental setups.

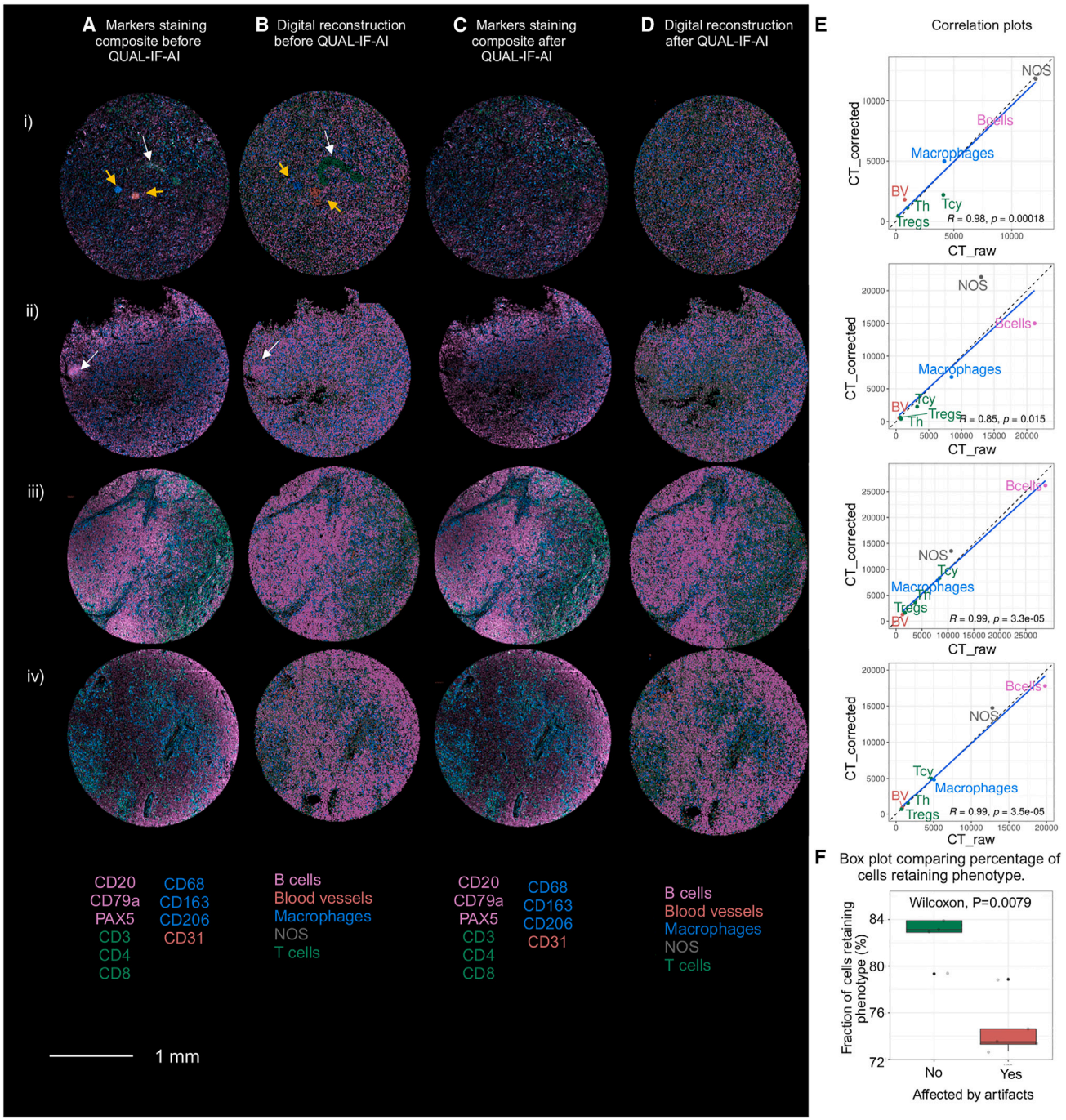


Figure 3. Impact of QUALIFAI in cell phenotyping

(A and B) Composite of marker stainings before (A) and after (C) artifact removal with QUALIFAI. Markers are colored based on the cell types they identify: CD20, CD79A, and PAX5 (B cells, magenta); CD3, CD4, and CD8 (T cells, green); CD68, CD163, and CD206 (macrophages, blue); and CD31 (blood vessels, red). The scale bar represents 1 mm.

(C and D) Digital reconstructions projecting the identity of every cell to its original location in the tissue before (B) and after (D) artifact removal with QUALIFAI. The color scheme is the same as described for the stainings in (A) and (B): B cells (magenta), T cells (green), macrophages (blue), and blood vessels (red). Cells not characterized by these markers are shown in gray (NOS, not otherwise specified). Yellow arrows indicate antibody aggregates, and white arrows indicate external artifacts. The scale bar represents 1 mm.

Figure 3. *Continued*
(E) Scatterplots showing the number of cells identified for each type before (x axis) and after (y axis) QUALIFAI. The blue line indicates the fitted linear model. The black dashed line indicates the perfect fit (x = y). Pearson correlation coefficients as well as *p* values from correlation tests are reported.
(F) Boxplot comparing the percentages of cells retaining their phenotype after QUALIFAI in the cores affected by artifacts (red) versus those not affected by artifacts (green). The *p* value from a Wilcoxon rank-sum test is reported (*p* value = 0.0079).

QUALIFAI improves the accuracy of multiplexed fluorescent image analysis

In this section, we demonstrate the importance of using QUALIFAI for artifact removal to improve downstream analysis of multiplexed immunofluorescence images. To this end, we analyzed a 10-core tissue microarray (TMA) from patients with diffuse large B cell lymphoma.[37] Five of the TMA cores contained artifacts (Figures 3, i and ii, and S4, v, vi, and vii), while the remaining 5 did not include artifacts (Figures 3, iii and iv, and S4, viii, ix, and x). The tissue slides included markers to identify B cells (CD20, CD79A, PAX5), T cells (CD3, CD4, CD8), macrophages (CD68, CD163, CD206), and blood vessels (CD31). Analyzing these images with state-of-the-art image analysis pipelines[3] but without artifact management results in the misclassification of cells (Figure 3B). However, coupling the same imaging pipeline with QUALIFAI provides a more accurate representation of tissue architecture. When comparing the fraction of cells preserving the same label between artifact-affected and non-affected cores, we observed significantly greater concordance in cores not affected by artifacts ($p = 0.0079$, Figures 3A and 3B). For example, in tissue id i (Figure 3, i), artifacts in the CD3 and CD8 markers caused the erroneous identification of cytotoxic T cells. After artifact removal, the relative fraction of identified cytotoxic T cells decreased by 46.28%. Additionally, artifacts in the CD31 channel led to the misclassification of cells as blood vessels while missing out on the real blood vessels (Figure 3A, tissue [i]). In contrast, TMA cores unaffected by artifacts showed a maximum relative change of 10.12% across all cell types, which we deemed reasonable considering the impact of artifact-containing cores during downstream analysis. Furthermore, when examining the set of cells that changed their predicted phenotype upon the implementation of QUALIFAI, we observed that while areas affected with artifacts had a higher density of cells changing predicted phenotype, cells outside artifactual areas were also affected (Figure S5). This may be due to downstream steps such as normalization and clustering. Cells in areas with artifacts may exhibit significantly higher expression of the affected markers, skewing normalization. This skewed normalization alters the expression space, affecting clustering and, thus, the assigned cell labels. This phenomenon can also explain why cores not affected by artifacts show differences in cell phenotype assignment.

DISCUSSION

In the context of artifact detection in fluorescence microscopy, several approaches have been explored. Besides manual annotation, classical algorithms, such as Laplacian variance[12] for detecting blur and manually tuned thresholds for detecting saturation artifacts,[13] have been employed. These bespoke solutions, although customized, can be resource intensive and may struggle to adapt to varying experimental conditions or novel data.

In contrast, a comprehensive AI-based tool like QUALIFAI offers a superior solution by automating the detection process. QUALIFAI leverages DL techniques capable of learning complex patterns from large datasets, thereby reducing the need for manual algorithm development and providing a more scalable and adaptable solution. QUALIFAI employs a two-tier approach for automated artifact detection and management in fluorescent images, utilizing binary classification and segmentation models, each fine-tuned to specifically delineate specific artifact types.

In this manuscript, we favor binary models over multiclass models due to the unique challenges presented by each artifact type. For example, external artifacts are easier to segment due to their distinct intensity, while OOF regions are more subjective and complex, particularly at varying magnifications. While specialized binary models offer a more time-consuming implementation, they are generally more robust than multiclass models.[38] Despite these challenges, we do not dismiss the potential for multiclass models to achieve comparable performance with more extensive training data and advanced optimization techniques. Future research should focus on augmenting the training datasets and experimenting with novel architectures or hybrid approaches that could better balance the trade-offs between simplicity and comprehensive artifact management.

To the best of our knowledge, QUALIFAI represents a first-in-class methodology due to its ability to identify five different common artifacts present in fluorescent images. QUALIFAI focuses on the most common types of artifacts, including air bubbles, tissue folds, external artifacts, antibody aggregates, and OOF areas, utilizing data acquired from three different multiplexing technologies: MILAN, COMET, and CODEX. However, the diversity of artifact categories present in immunofluorescence images is larger, and some less frequent but potentially relevant artifacts may have not been included in the method. Additionally, acquisition technologies not included in our study might require fine-tuning before the tool provides satisfactory results.

In summary, our study presents an artifact identification method for immunofluorescence images that achieves a classification accuracy of over 90% and a segmentation IoU of more than 0.6 for any of the five artifact types included in the study. We have shown that adopting an active learning framework accelerates the improvement in accuracy, requiring less data annotation. Collectively, our approach advances artifact detection, enabling more accurate image analysis and laying the groundwork for enhanced immunofluorescence staining applications.

EXPERIMENTAL PROCEDURES

Data collection

A diverse and heterogeneous dataset was built including fully anonymized images from different acquisition technologies: MILAN,[39] Lunaphore (COMET),[28] and Akoya (CODEX).[29] To increase the generalizability of our models, the collected datasets included a variety of tissue types representing different pathologies, assembled in several types of slides, and acquired at different resolutions. The complete set of parameters considered when collecting these datasets is detailed in Table 1. The MILAN dataset[3,37,40–42] and the CODEX[43] dataset were built by acquiring images from previously published studies. To construct the Lunaphore dataset, the sequential immunofluorescence (seqIF) technique was employed using Lunaphore's automated staining and imaging COMET platform. Two formalin-fixed and paraffin-embedded (FFPE) slides were acquired: a whole resection biopsy of a human tonsil and a TMA containing a variety of cancer tissues (2 breast cancer, 2 glioblastoma, 2 hepatocellular carcinoma, 2 melanoma, and 1 lymph node). These were stained with a set of 10 markers, including CD68, FoxP3, CD8, PD-1, Ki67, PD-L1, CK, CD4, CD20, and CD3. The stainings were performed following the recommended protocol reported by Lunaphore in the literature.[28]

Image pre-processing and artifact annotations

The images were processed with zero padding and normalized using q99 normalization. Subsequently, images were split into tiles of 512 pixels on each side with a

20 pixel overlap between consecutive slides. For each of the above-mentioned artifacts, 50 randomly selected tiles were visually evaluated and manually annotated by image analysis experts (M.D.A., T.O., J.V., M.V., B.D., N.D.) using our in-house-built fast labeling tool (see below). The image tiles obtained were first categorized based on the presence or absence of artifacts. Annotations for air bubbles, external artifacts, tissue folds, and OOF areas were done only on the DAPI channel, as these artifacts affect all channels equally and appear on the same regions (channel-agnostic artifacts), which represents a common practice in the field.[7] Later, training was performed using all the channels. In the case of antibody aggregates, these are channel-specific artifacts, and therefore the annotations and training of the models were performed only in the affected channels. For images where artifacts were detected, manual artifact segmentation was performed to precisely delineate the areas affected by these artifacts. Further, following an active learning approach, incrementally more tiles were annotated up to a total of 7,508 (Table 2). To evaluate the performance of the trained models, a validation dataset, consisting of a total of 1,920 images belonging to one of the mentioned 5 artifact categories, was built separately and not used during active learning or model training.

QUALIFAI fast labeling tool

We have developed a labeling tool to perform a fast annotation of artifacts in the images. This tool is divided into two modules. The first module is the classification module (also called fast-flagging mode), where images from a pre-defined input directory are sequentially displayed to the user, who can flag the images as having an artifact by clicking "yes" or "no." If the user flags an image as having an artifact, then the tool moves forward to the segmentation module (also called labeling mode), where the user can identify the type of artifact and annotate/draw the exact location of the artifact. ImageJ plugins are incorporated in the tool to modify the images for better visualization[44] (Figure S6).

Model training

Model training was performed independently for each artifact type, and it concatenates a classification and a segmentation module to generate the final output. Tiles were split into training and testing datasets. Splitting was performed in a stratified manner, keeping a balanced representation of each artifact in each cohort.

Classification

For each artifact category, a separate binary classification model was trained to recognize tiles containing a specific artifact. Given that the same tile can have multiple artifacts at the same time, a multilevel classification approach was discarded. Several popular classification architectures were tested for this task. The model architecture and initial weights were imported from the ImageNet dataset as implemented in the Keras library in Python.[26] Using models pre-trained on the ImageNet dataset reduces the number of images required for model training.[45] In DL models, first-layer features are general, and they are applicable to many datasets and tasks. Features eventually transition from general to dataset-specific tasks by the last layer of the network. The choice to unfreeze the last four layers while retaining the transfer learning weights for other layers was based on the need to fine-tune the model to our specific task while preserving the general feature extraction capabilities of the pre-trained layers. This approach is widely recommended in the literature for balancing generalization and task-specific adaptation.[45] The "ImageDataGenerator" function from Keras was used for data augmentation with horizontal flip, vertical flip, and random rotation of maximum 90°. The "Adam" optimizer with a

learning rate of 0.001 was used, and "binary cross-entropy" was selected as the loss function. Adam is known for its efficiency and effectiveness in training deep neural networks, as it adapts the learning rate for each layer.[46] We selected a batch size of 16 considering the relatively small data size of our dataset. Additionally, we augmented the data. Each model was trained for a maximum of 50 cycles (epochs). To enhance performance and avoid overfitting, we monitored the test accuracy during training. The model achieving the highest test accuracy was saved as the best model.

Active learning framework

After the initial training including 50 images, a template was constructed by applying principal-component analysis (PCA) on the 128-tensor generated by the dense layer preceding the output layer of the classification models (Figure S1A). Then, new data were projected onto this PCA and using a k-nearest neighbors (KNN) approach. Considered a hyperparameter, different numbers of neighbors were tested: 5, 15, and 20. The projected images were classified into artifacts (if more than 75% neighbors belong to the artifact category), no artifacts (if more than 75% neighbors belong to the no-artifact category), and uncertain (if less than 75% artifacts belong to either category). Then, 35 tiles from the uncertain category were randomly selected and annotated as described in the experimental procedures. Finally, the model was retrained, and the process repeated until convergence was reached. We defined convergence as no improvement in the validation accuracy in three consecutive iterations.

Comparison with CellProfiler

We broadly assigned the external artifacts and antibody aggregates mentioned in our work to the saturation debris category, while tissue folds and air bubbles were assigned to the OOF areas category. We measured various parameters, including blur metrics and saturation metrics, as described by Bray et al.[7,8] We then trained random forest classifiers separately for each artifact category.

Segmentation

For each artifact category, a separate segmentation model was trained to delineate the tissue area covered by the artifact with pixel-level resolution. The U-Net architecture from the "segmentation models" Python library was used for this task.[27] The U-Net model consists of two parts: an encoder and a decoder. The function of the encoder is to extract features and encode the original input image to obtain a fixed-length vector. The function of the decoder is to decode the vector to generate a new output image. The segmentation models library allows users to choose an encoder architecture with a varying number of parameters. In this study, the model with the highest accuracy in the classification task was used as the encoding backbone to build the segmentation model. Weights from the ImageNet dataset were used as a starting point to train the segmentation models. The last four layers of the encoding backbone and the complete decoder were unfrozen. The ImageDataGenerator function from Keras was used for data augmentation with vertical flip, horizontal flip, and random rotation of maximum 90°. The Adam optimizer with a learning rate of 0.001 was used, and binary cross-entropy was selected as the loss function. Models were run for a maximum of 50 epochs, with the model showing the highest Jaccard index on the test dataset being saved. Running for up to 50 epochs provides a balance between sufficient training time and preventing overfitting. A batch size of 16 was also used for segmentation models to maintain consistency and due to its effectiveness observed during classification tasks.

CellPress
OPEN ACCESS

Evaluation metrics

Classification models were evaluated using accuracy, which judges the overall correctness of predictions across all classes and is convenient in a balanced dataset.[47] Accuracy scores were represented together with ROC curves to compare different models for different artifact types. ROC curves provide a graphical representation of the trade-off between the true positive rate and the false positive rate at various threshold values.[48] Segmentation models were evaluated using the IoU metric. IoU is a metric between 0 and 1 that quantifies the overlap between the predicted mask and its corresponding annotation.[49]

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Asier Antoranz (asier.antoranzmartinez@kuleuven.be).

Materials availability

This study did not generate new unique materials.

Data and code availability

The Lunaphore dataset generated for this study is available at Zenodo: https://zenodo.org/records/12699470. All original codes have been deposited at Zenodo: https://doi.org/10.5281/zenodo.13358778 and are publicly available from the date of publication.

ACKNOWLEDGMENTS

M.D.A. is supported by KU Leuven grant EHU-D6472-C16/19/006. M.V. is supported by FWO PhD fellowship fundamental research (11L0824N). A. Antoranz is supported by a postdoctoral fellowship senior grant from FWO (12ATN24N). F.B. is supported by FWO Fundamenteel Klinisch Mandaat EMH-D8972-FKM/20. B.D.M. is funded by the KU Leuven Research Fund (projects iBOF/23/064, C16/15/059, C3/19/053, C24/18/022, C3/20/117, and C3I-21-00316); the KU Leuven Industrial Research Fund: EOS project no. G0F6718N (SeLMA), SBO project S005319N, infrastructure project I013218N, and TBM project T001919N; a PhD grant (SB/1SA1319N); EWI: the Flanders AI Research Program, VLAIO: CSBO (HBC.2021.0076) Baekeland PhD (HBC.20192204); the European Commission: European Research Council under the European Union's Horizon 2020 research and innovation program; and the Foundation "Kom op tegen Kanker," CM (Christelijke Mutualiteit).

AUTHOR CONTRIBUTIONS

Methodology, M.D.A.; conceptualization, B.D.M., F.D.S., F.B., and A. Antoranz; validation, M.D.A.; formal analysis, M.D.A., P.N., G.S., A. Arnould, and A. Antoranz; investigation, M.D.A., N.D., M.V., and B.D.; writing – original draft, M.D.A.; writing – review & editing, M.D.A., G.R., P.N., G.S., N.D., M.V., B.D., A. Arnould, W.M., B.D.M., F.D.S., F.B., and A. Antoranz; software, G.R. and W.M.; supervision, B.D.M., F.D.S., F.B., and A. Antoranz; funding acquisition, F.B.;

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT[50] in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrp.2024.102220.

Received: March 13, 2024
Revised: July 24, 2024
Accepted: September 2, 2024
Published: September 24, 2024

REFERENCES

1. R. Šachl, and M. Amaro, eds. (2023). Fluorescence Spectroscopy and Microscopy in Biology (Springer International Publishing). https://doi.org/10.1007/978-3-031-30362-3.

2. Bosisio, F.M., Van Herck, Y., Messiaen, J., Bolognesi, M.M., Marcelis, L., Van Haele, M., Cattoretti, G., Antoranz, A., and De Smet, F. (2022). Next-Generation Pathology Using Multiplexed Immunohistochemistry: Mapping Tissue Architecture at Single-Cell Level. Front. Oncol. *12*, 918900.

3. Antoranz, A., Van Herck, Y., Bolognesi, M.M., Lynch, S.M., Rahman, A., Gallagher, W.M., Boecxstaens, V., Marine, J.-C., Cattoretti, G., van den Oord, J.J., et al. (2022). Mapping the Immune Landscape in Metastatic Melanoma Reveals Localized Cell–Cell Interactions That Predict Immunotherapy Response. Cancer Res. *82*, 3275–3290. https://doi.org/10.1158/0008-5472.CAN-22-0363.

4. Schapiro, D., Sokolov, A., Yapp, C., Chen, Y.-A., Muhlich, J.L., Hess, J., Creason, A.L., Nirmal, A.J., Baker, G.J., Nariya, M.K., et al. (2022). MCMICRO: a scalable, modular image-processing pipeline for multiplexed tissue imaging. Nat. Methods *19*, 311–315. https://doi.org/10.1038/s41592-021-01308-y.

5. Eng, J., Bucher, E., Hu, Z., Zheng, T., Gibbs, S.L., Chin, K., and Gray, J.W. (2022). A framework for multiplex imaging optimization and reproducible analysis. Commun. Biol. *5*, 438. https://doi.org/10.1038/s42003-022-03368-y.

6. Jiang, J., Moore, R., Jordan, C.E., Guo, R., Maus, R.L., Liu, H., Goode, E., Markovic, S.N., and Wang, C. (2023). Multiplex Immunofluorescence Image Quality Checking Using DAPI Channel-referenced Evaluation. J. Histochem. Cytochem. *71*, 121–130. https://doi.org/10.1369/00221554231161693.

7. Bray, M.-A., and Carpenter, A.E. (2018). Quality Control for High-Throughput Imaging Experiments Using Machine Learning in Cellprofiler. In High Content Screening Methods in Molecular Biology, P.A. Johnston and O.J. Trask, eds. (Springer), pp. 89–112. https://doi.org/10.1007/978-1-4939-7357-6_7.

8. Bray, M.A., Fraser, A.N., Hasaka, T.P., and Carpenter, A.E. (2012). Workflow and metrics for image quality control in large-scale high-content screens. J. Biomol. Screen *17*, 266–274. https://doi.org/10.1177/1087057111420292.

9. Taqi, S.A., Sami, S.A., Sami, L.B., and Zaki, S.A. (2018). A review of artifacts in histopathology. J. Oral Maxillofac. Pathol. *22*, 279. https://doi.org/10.4103/jomfp.JOMFP_125_15.

10. Wright, A.I., Dunn, C.M., Hale, M., Hutchins, G.G.A., and Treanor, D.E. (2021). The Effect of Quality Control on Accuracy of Digital Pathology Image Analysis. IEEE J. Biomed. Health Inform. *25*, 307–314. https://doi.org/10.1109/JBHI.2020.3046094.

11. Dimitriou, N., Arandjelović, O., and Caie, P.D. (2019). Deep Learning for Whole Slide Image Analysis: An Overview. Front. Med. *6*, 264.

12. Bansal, R., Raj, G., and Choudhury, T. (2016). Blur image detection using Laplacian operator and Open-CV. In 2016 International Conference System Modeling & Advancement in Research Trends (SMART) (IEEE), pp. 63–67. https://doi.org/10.1109/SYSMART.2016.7894491.

13. Kothari, S., Phan, J.H., and Wang, M.D. (2013). Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade. J. Pathol. Inform. *4*, 22. https://doi.org/10.4103/2153-3539.117448.

14. van der Laak, J., Litjens, G., and Ciompi, F. (2021). Deep learning in histopathology: the path to the clinic. Nat. Med. *27*, 775–784. https://doi.org/10.1038/s41591-021-01343-4.

15. Brunt, E.M., Clouston, A.D., Goodman, Z., Guy, C., Kleiner, D.E., Lackner, C., Tiniakos, D.G., Wee, A., Yeh, M., Leow, W.Q., et al. (2022). Complexity of ballooned hepatocyte feature recognition: Defining a training atlas for artificial intelligence-based imaging in NAFLD. J. Hepatol. *76*, 1030–1041. https://doi.org/10.1016/j.jhep.2022.01.011.

16. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M., and Madabhushi, A. (2019). HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides. JCO Clin. Cancer Inform. *3*, 1–7. https://doi.org/10.1200/CCI.18.00157.

17. Ke, J., Liu, K., Sun, Y., Xue, Y., Huang, J., Lu, Y., Dai, J., Chen, Y., Han, X., Shen, Y., and Shen, D. (2023). Artifact Detection and Restoration in Histology Images with Stain-Style and Structural Preservation. IEEE Trans. Med. Imaging *42*, 3487–3500. https://doi.org/10.1109/TMI.2023.3288940.

18. Rodrigues, D., Reinhard, S., Waldburger, T., Martin, D., Couto, S., Zlobec, I., Caie, P., and Burlingame, E. (2023). Abstract 5442: SlideQC: An AI-based tool for automated quality control of whole-slide digital pathology images. Cancer Res. *83*, 5442. https://doi.org/10.1158/1538-7445.AM2023-5442.

19. Hosseini, M.S., Brawley-Hayes, J.A.Z., Zhang, Y., Chan, L., Plataniotis, K., and Damaskinos, S. (2020). Focus Quality Assessment of High-Throughput Whole Slide Imaging in Digital Pathology. IEEE Trans. Med. Imaging *39*, 62–74. https://doi.org/10.1109/TMI.2019.2919722.

20. Samani, Z.R., Alappatt, J.A., Parker, D., Ismail, A.A.O., and Verma, R. (2019). QC-Automator: Deep Learning-Based Automated Quality Control for Diffusion MR Images. Front. Neurosci. *13*, 1456.

21. Ali, M.A.S., Hollo, K., Laasfeld, T., Torp, J., Tahk, M.-J., Rinken, A., Palo, K., Parts, L., and Fishman, D. (2022). ArtSeg—Artifact segmentation and removal in brightfield cell microscopy images without manual pixel-level annotations. Sci. Rep. *12*, 11404. https://doi.org/10.1038/s41598-022-14703-y.

22. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

23. Kim, H.E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M.E., and Ganslandt, T. (2022). Transfer learning for medical image classification: a literature review. BMC Med. Imaging *22*, 69. https://doi.org/10.1186/s12880-022-00793-7.

24. Talo, M. (2019). Automated classification of histopathology images using transfer learning. Artif. Intell. Med. *101*, 101743. https://doi.org/10.1016/j.artmed.2019.101743.

25. Guan, Q., Wang, Y., Ping, B., Li, D., Du, J., Qin, Y., Lu, H., Wan, X., and Xiang, J. (2019). Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study. J. Cancer *10*, 4876–4882. https://doi.org/10.7150/jca.28769.

26. Chollet, F. (2015). keras. https://keras.io/.

27. Iakubovskii, P. (2023). qubvel/segmentation_models. https://github.com/qubvel/segmentation_models.

28. Rivest, F., Eroglu, D., Pelz, B., Kowal, J., Kehren, A., Navikas, V., Procopio, M.G., Bordignon, P., Pérès, E., Ammann, M., et al. (2023). Fully automated sequential immunofluorescence (seqIF) for hyperplex spatial proteomics. Sci. Rep. *13*, 16994. https://doi.org/10.1038/s41598-023-43435-w.

29. Black, S., Phillips, D., Hickey, J.W., Kennedy-Darling, J., Venkataraaman, V.G., Samusik, N., Goltsev, Y., Schürch, C.M., and Nolan, G.P. (2021). CODEX multiplexed tissue imaging with DNA-conjugated antibodies. Nat. Protoc. *16*, 3802–3835. https://doi.org/10.1038/s41596-021-00556-8.

30. Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

31. Settles, B., and Craven, M. (2008). An Analysis of Active Learning Strategies for Sequence Labeling Tasks. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, M. Lapata and H.T. Ng, eds. (Association for Computational Linguistics), pp. 1070–1079.

32. Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B.B., Chen, X., and Wang, X. (2021). A Survey of Deep Active Learning. ACM Comput. Surv. *54*, 1–40. https://doi.org/10.1145/3472291.

33. Holub, A., Perona, P., and Burl, M.C. (2008). Entropy-based active learning for object recognition. In 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (IEEE) (IEEE), pp. 1–8. https://doi.org/10.1109/CVPRW.2008.4563068.

34. Zhao, Y., Ossowski, J., Wang, X., Li, S., Devinsky, O., Martin, S.P., and Pardoe, H.R. (2021). Localized Motion Artifact Reduction on Brain MRI Using Deep Learning with Effective Data Augmentation Techniques. In 2021 International Joint Conference on Neural Networks (IJCNN) (IEEE) (IEEE), pp. 1–9. https://doi.org/10.1109/IJCNN52387.2021.9534191.

35. Hliboký, M., Magyar, J., Bundzel, M., Malík, M., Števík, M., Vetešková, Š., Dzian, A., Szabóová, M., and Babič, F. (2023). Artifact Detection in

Lung Ultrasound: An Analytical Approach. Electronics *12*, 1551. https://doi.org/10.3390/electronics12071551.

36. Kozaczko, K., Szpot, R., Brodzicki, A., and Wójcicka, A. (2023). Automatic Artifact Removal from Dermoscopic Images. In Advanced, Contemporary Control, M. Pawelczyk, D. Bismor, S. Ogonowski, and J. Kacprzyk, eds. (Springer Nature Switzerland), pp. 193–202. https://doi.org/10.1007/978-3-031-35173-0_19.

37. Marcelis, L., Antoranz, A., Delsupehe, A.-M., Biesemans, P., Ferreiro, J.F., Debackere, K., Vandenberghe, P., Verhoef, G., Gheysens, O., Cattoretti, G., et al. (2020). In-depth characterization of the tumor microenvironment in central nervous system lymphoma reveals implications for immune-checkpoint therapy. Cancer Immunol *69*, 1751–1766. https://doi.org/10.1007/s00262-020-02575-y.

38. Berstad, T.J.D., Riegler, M., Espeland, H., De Lange, T., Smedsrud, P.H., Pogorelov, K., Kvale Stensland, H., and Halvorsen, P. (2018). Tradeoffs Using Binary and Multiclass Neural Network Classification for Medical Multidisease Detection. In 2018 IEEE International Symposium on Multimedia (ISM) (IEEE), pp. 1–8. https://doi.org/10.1109/ISM.2018.00009.

39. Cattoretti, G., Bosisio, F.M., Marcelis, L., and Bolognesi, M.M. (2019). Multiple Iterative Labeling by Antibody Neodeposition (MILAN). Protocol Exchange. https://doi.org/10.21203/rs.2.1646/v5.

40. Cortesi, E.E., Antoranz, A., Geudens, V., Mc Donough, J.E., Vanaudenaerde, B., Gayan-Ramirez, G., Carlon, M.S., Janssens, V., Wuyts, W.A., Bosisio, F.M., and Roskams, T. (2023). Single-cell spatial proteomics depicts novel dynamics involving senescent progenitors in fibrotic lesions in COPD and IPF. Eur. Respir. J. *62*, OA1572. https://doi.org/10.1183/13993003.congress-2023.OA1572.

41. Lamarthée, B., Callemeyn, J., Van Herck, Y., Antoranz, A., Anglicheau, D., Boada, P., Becker, J.U., Debyser, T., De Smet, F., De Vusser, K., et al. (2023). Transcriptional and spatial profiling of the kidney allograft unravels a central role for FcγRIII+ innate immune cells in rejection. Nat. Commun. *14*, 4359. https://doi.org/10.1038/s41467-023-39859-7.

42. Naulaerts, S., Datsi, A., Borras, D.M., Antoranz Martinez, A., Messiaen, J., Vanmeerbeek, I., Sprooten, J., Laureano, R.S., Govaerts, J., Panovska, D., et al. (2023). Multiomics and spatial mapping characterizes human CD8+ T cell states in cancer. Sci. Transl. Med. *15*, eadd1016. https://doi.org/10.1126/scitranslmed.add1016.

43. Pozniak, J., Pedri, D., Landeloos, E., Van Herck, Y., Antoranz, A., Vanwynsberghe, L., Nowosad, A., Roda, N., Makhzami, S., Bervoets, G., et al. (2024). A TCF4-dependent gene regulatory network confers resistance to immunotherapy in melanoma. Cell *187*, 166–183.e25. https://doi.org/10.1016/j.cell.2023.11.037.

44. Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671–675. https://doi.org/10.1038/nmeth.2089.

45. Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? arXiv. https://doi.org/10.48550/ARXIV.1411.1792.

46. Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. Preprint at arXiv. https://doi.org/10.48550/ARXIV.1412.6980.

47. Hicks, S.A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M.A., Halvorsen, P., and Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. Sci. Rep. *12*, 5979. https://doi.org/10.1038/s41598-022-09954-8.

48. Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J. Intern. Med. *4*, 627–635.

49. Padilla, R., Netto, S.L., and da Silva, E.A.B. (2020). A Survey on Performance Metrics for Object-Detection Algorithms. In 2020 International Conference on Systems, Signals and Image Processing (IWSSIP) (IEEE), pp. 237–242. https://doi.org/10.1109/IWSSIP48289.2020.9145130.

50. OpenAI (2024). ChatGPT (4o) [Large language model]. https://chatgpt.com.