High-Quality Immunohistochemical Stains through Computational Assay Parameter Optimization

Author(s):
Arar, Nuri M.; Pati, Pushpak; Kashyap, Aditya; Fomitcheva Khartchenko, Anna; Goksel, Orcun; Kaigala, Govind V.; Gabrani, Maria

Publication date:
2019-10

Permanent link:
https://doi.org/10.3929/ethz-b-000331801

Originally published in:
IEEE Transactions on Biomedical Engineering 66(10), https://doi.org/10.1109/tbme.2019.2899156

High-Quality Immunohistochemical Stains through Computational Assay Parameter Optimization

Nuri Murat Arar†, Pushpak Pati†, Aditya Kashyap, Anna Fomitcheva Khartchenko, Orcun Goksel, Govind V. Kaigala, Maria Gabrani

***Abstract*—Accurate profiling of tumors using immunohistochemistry (IHC) is essential in cancer diagnosis. The inferences drawn from IHC-stained images depend to a great extent on the quality of immunostaining, which is in turn affected strongly by assay parameters. To optimize assay parameters, the available tissue sample is often limited. Moreover, with current practices in pathology, exploring the entire assay parameter space is not feasible. Thus, the evaluation of IHC stained slides is conventionally a subjective task, in which diagnoses are commonly drawn on images that are suboptimal. In this work, we introduce a framework to analyze IHC staining quality and its sensitivity to process parameters. To that extent, first histopathological sections are segmented automatically. Then, machine learning techniques are employed to extract disease-specific staining quality metrics (SQMs) targeting a quantitative assessment of staining quality. Lastly, an approach to efficiently analyze the parameter space is introduced to infer sensitivity to process parameters. We present results on microscale IHC tissue samples of five breast tumor classes, based on disease state and protein expression. A disease-type classification F1-score of 0.82 and a contrast-level classification F1-score of 0.95 were achieved. With the proposed SQMs an area under the curve of 0.85 was achieved on average over different disease types. Our methodology provides a promising step in automatically evaluating and quantifying staining quality of IHC stained tissue sections, and it can potentially standardize immunostaining across diagnostic laboratories.**

***Index Terms*—automated standardization protocol, HER2, immunostaining, quality assessment, quantitative evaluation**

I. Introduction

MALIGNANCIES are often studied and detected by acquiring a protein expression profile on a tissue section. Such a protein expression map on a tissue is obtained by immunohistochemical (IHC) staining thereby generating a visual signal while retaining the tissue structure of tissues (histology). IHC has been an invaluable tool in the field of both cancer diagnostics and research, owing to a rapidly obtainable snapshot of status of cells within tissue samples. In this paper, we focus on a new methodology for realizing high-quality immunostaining both at the micrometer-length scale and for conventional whole-tissue staining for tumor stratification.

Manuscript received February 19, 2018; revised August 24, 2018 and February 7, 2019; accepted January 25, 2019. Date of current version February 7, 2019. A. Kashyap, A. Fomitcheva Khartchenko and G. Kaigala acknowledge funding from the European Research Council (Project No. 311122). All authors acknowledge support from E. Delamarche, W. Riess, C. Bekas and P. Buhler. (N.M. Arar† and P. Pati† contributed equally to this work.) (Corresponding author: M. Gabrani)

N.M. Arar, P. Pati, A. Kashyap, A. Fomitcheva Khartchenko, G. Kaigala and M. Gabrani are with IBM Research-Zurich, Saeumerstrasse 4, 8803 Rüschlikon, Switzerland (e-mail: mga@zurich.ibm.com). P. Pati and O. Goksel are with ETH Zurich, Switzerland.

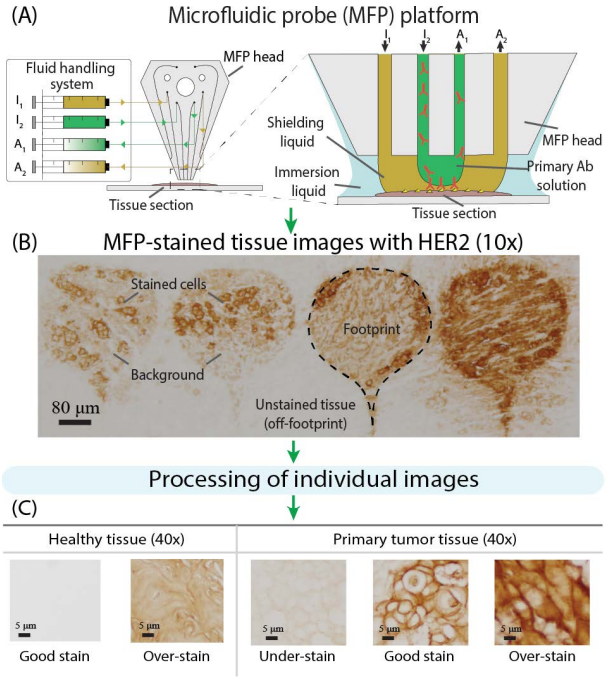

Fig. 1. Immunostaining process and staining quality variability due to process parameter variation of IHC. (A) illustration of microfluidic probe platform for microscale IHC, (B) sample HER2-stained tissue images using an MFP, (C) staining quality variability on healthy and primary tumor tissues. For HER2 non-expressing healthy tissue, low and high HER2 expression indicates high-quality staining (true negative) and over-staining (false positive), respectively. For HER2 overexpressing primary tumor tissue, low HER2 expression indicates under-staining (false negative). HER2 expression solely on membranes implies high-quality staining (true positive), whereas expression in cytoplasm and stroma indicates over-staining (false positive).

IHC is implemented by exposing a tissue to antibodies which bind to a specific protein, thus identifying prognostic and treatment-related biomarkers. The commonly used IHC protocol is a multi-step and multi-parametric process [1] and involves binding of a primary antibody specific to a protein of interest on the tissue, followed by a secondary antibody that binds to the primary. The colored signal on the tissue is obtained using a chromogenic moiety coupled with the secondary antibody, where the chromatic signal strength is a function of the density of the proteins of interest in the tissue, their accessibility, and the concentration of the antibody that is exposed to the antigen, among several other parameters. The IHC signal can provide vital information in a diagnosis workflow. However, when the process parameters are not optimal, this may lead to difficult-to-interpret images and potential misdiagnosis, e.g., false positive and false negative staining as demonstrated in Fig. 1.

Although IHC has been used now for decades, standardization and reproducibility remain two major concerns. Pathology laboratories manually determine the parameters leading to a good staining quality. Such a manual process comprising trial-and-error is cumbersome and tissue exhaustive. Besides, it is characterized by high inter- and intra-laboratory variability, leading to poor reproducibility. Nordic Immunohistochemical Quality Control (NordiQC), an international external quality-assurance organization, found that about 20% of the staining results in a breast-cancer IHC cross-lab examination were insufficient for diagnostic use [2]. Inaccurate and/or equivocal results are mostly obtained because of inappropriate parameters used in the staining process (protocol), less specific antibodies, insufficiently calibrated antibody dilutions, variable fixation processes and erroneous epitope retrieval methods. To improve the standardization in immunostaining, efforts have been made by ad-hoc committees on pathology [3]–[8], by external quality-assurance schemes [9]–[12], and by field researchers [13], [14] through addressing one or more of the factors affecting the staining results. The effect of specific process parameters on the quality is hard to deconvolve owing to limited tools that allow for the scanning of a range of process parameters on the same tissue. Thus, strategies that perform automated analysis of process parameter sensitivity and contextual quantitative analysis are crucial in improving the IHC standardization, and thus reproducibility.

More recently, the advent of digital pathology has prioritized the extraction of quantitative information from scanned histopathological sections to aid pathologists in the diagnostic process, while attempting to reduce or eliminate observer biases [15], [16]. Furthermore, computational pathology aims at automating the analysis of stained sections, as manually analyzing numerous biopsy slides can be tedious and labor intensive. Recent advances enabled the automated recognition of pathological patterns, which has the potential to provide valuable assistance to a pathologist. There exist several studies which demonstrate the agreement between digital image analysis- based methods and pathologistsvisual examination. For instance, Dobson et al. [17] and Brugmann et al. [18] demonstrated that HER2 antibody protein expression can be classified with a high accuracy by analyzing the staining intensity and membrane connectivity on IHC images with optimal staining quality. Differently from the previous work, this work deals with IHC-stained tissue images with both optimal and sub-optimal staining quality. The combination of such quantification-aided diagnosis with quantified grading has the potential to improve diagnostic accuracy.

Limited prior work exists on the quantitative analysis of the immunostaining quality. Pinard et al. [13] proposed a system that extracts quantitative quality indicators and compares them with the respective user-defined minimum acceptable quality thresholds. Failure of one or more of the indicators to meet its respective threshold suggests that the sample is unsuitable for a subsequent automated pathological evaluation. Similarly, Grunkin and Hansen [14] described a method for assessing the staining quality of specimens in a working laboratory. Their system compares the quality parameters, e.g., staining intensity, connectivity, number of cells, Allred-score, Nottingham index, obtained from a reference specimen prepared at a standardized laboratory according to a predetermined staining protocol with the quality parameters obtained from a specimen prepared at the working laboratory. The relative quality measure is computed using a distance metric between the quality parameters of the test and reference specimen. Both studies output relative quality estimates with respect to either a user-defined threshold or a reference specimen, which limits the standardization of the process and thus the reproducibility. Instead, we propose to use reference standards for quality labeling during the training phase and use the trained quality metrics during the testing phase, thereby removing the need of posterior standards. The proposed methodology does not completely remove the need of an external standard but reduces the dependency on it on a daily practice. In addition, neither of the aforementioned studies takes into account the diagnostical relevance of the signals on the stained images, which can potentially hamper the computed quality indicators. For an alternate perspective, our automated methodology first segments the diagnostically relevant and the contextually immaterial signals in an IHC-stained image, followed by machine learning models for estimating the quality indicators.

Addressing IHC assay limitations requires technologies that enable precise control of the various steps of the assay, including the ability to create multiple assay conditions on the same tissue section. Here we use a microfluidic probe (MFP) [19], a scanning microfluidic device that localizes nanoliter volumes of antibodies on micrometer scale areas of tissue sections. By leveraging the ability of the MFP to perform multiple microscale IHC tests on the same tissue section [20], [21], we not only can perform experimental parameter optimization of IHC by exposing adjacent areas on a sample to different experimental conditions (antibody concentration, incubation time), but can also be conservative of the tissue sample.

In this work, we introduce a complete methodology to quantify and analyze the staining quality and its sensitivity to IHC process parameters using well-established image processing and machine learning techniques. The proposed methodology first extracts quantitative information from scanned histopathological sections using an automated diagnostically relevant signal segmentation algorithm. It then learns multiple metrics for the quantitative assessment of the staining quality. Lastly, it performs an analysis of the sensitivity of staining quality to process parameters for the optimal parameter-space determination. Preliminary results of this work were presented in [22]. These have been extended herein with improvements on the methodology and validation. First, we refined our framework in order to account for different disease types. To achieve this, we conducted a comprehensive analysis of the impact of various image representation and classification-related parameters of the framework. We additionally explored alternative feature extraction and classification techniques, including deep learning strategies. We provide herein a comprehensive validation on a cohort of annotated breast cancer tissues from five different disease types. Moreover, we compared the proposed approach against the current clinical staining approach and demonstrate the superiority of the proposed staining approach.

Fig. 2. Overview of the proposed methodology. Images are segmented to extract different levels of information, which is used for generating various staining quality indicators. These are fed into a machine learning algorithm to learn multiple staining quality metrics. Lastly, an analysis of the quality sensitivity to the staining process parameters is performed for the identification of process parameter space resulting in optimal staining quality.

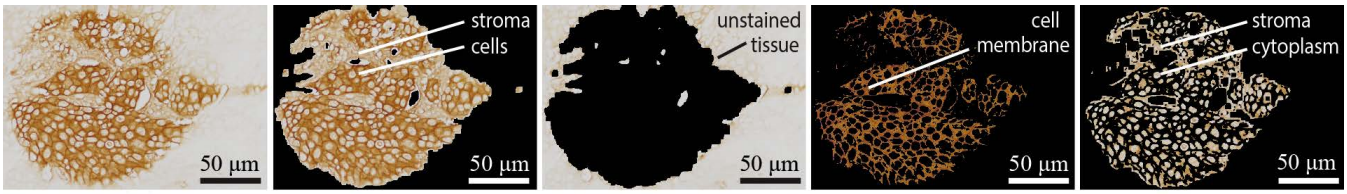

Fig. 3. Sample outputs of the segmentation algorithm: (a) input IHC-stained tissue, (b) stained region, where the probe is applied (*footprint*), (c) unstained region (*off-footprint*), (d) diagnostically relevant signal, e.g., staining of the cell membrane (*foreground*), (e) contextually immaterial signal (*background*).

II. Methods

The proposed methodology for staining quality and sensitivity assessment has 4 main components: a) separation of diagnostically relevant and contextually immaterial signals, b) staining quality metric learning, c) image and quality representation, and d) sensitivity analysis to staining process parameters. An overview of the methodology is illustrated in Fig. 2.

A. Diagnostically Relevant Signal Segmentation

Staining quality is directly proportional to the diagnostically relevant signal, i.e., the staining on interesting cell structures or regions (true positive staining), and is inversely proportional to the contextually immaterial signal, i.e., the staining on the remaining areas (false positive staining). An optimal staining quality is achieved when the ratio of the relevant signal to the immaterial signal is the highest. Therefore, our methodology essentially focuses on a good delineation of the two signals in IHC-stained tissue images prior to further analysis. Note that the definition of the two signals may vary depending on the choice of the biomarker used in the staining process, as each biomarker binds to a specific antigen present in stipulated cell structures. For instance, HER2 biomarker binds to the HER2 antigen in the cell membrane; developing diagnostically relevant signal on the cell membrane and arising immaterial signal on the remaining cell structures, i.e., the cytoplasm and stroma. Whereas, p53 biomarker produces the relevant signal on the nuclei of the tumor cells and develops the immaterial signal on the cell membrane and stroma.

Our methodology begins with an automatic segmentation algorithm based on a combination of well-known image processing techniques for separating aforementioned two signals in the images of μIHC-stained breast tissue. The algorithm firstly segments an image into two regions as *off-footprint* and *footprint*. The latter is further partitioned into two: *foreground* and *background*, as shown in Fig. 3 for a HER2-stained tissue. The segmentation process begins with finding and delineating the localized *footprint*, the tissue area where the MFP head is applied. To that end, we first estimate the *footprint* by Otsu binarization, followed by a morphological opening to generate highly confident masks for both the *footprint* and *off-footprint*; with the remaining regions considered as uncertain. Second, the obtained masks are fed into the Watershed algorithm to assign the uncertain areas into either *footprint* or *off-footprint* (Fig. 3B-C). Next, the *footprint* is subdivided into the relevant (*foreground*) and immaterial (*background*) regions as shown in Fig. 3D-E. Considering the intensity distribution difference between two regions, global thresholding with a robust threshold value is sufficient to extract the *foreground*. Here, we set the threshold value as the mean of the most frequent and maximum intensity values within the *footprint* region. To determine the threshold value more robustly, particularly in the presence of experimental or imaging artifacts (often resulting in a significantly high intensity), we calculate a 16-bin intensity histogram of the inverted gray-scale *footprint*, and extract the corresponding values from the histogram bins. Subsequently, we extract the *background*, i.e., false positive stain. Assuming that false positive staining highly occurs around the *foreground*, we subtract the binary *foreground* mask from the dilated *foreground* mask to extract the *background* within a close proximity of the *foreground*. We then ensure the connectivity of the *background* through a morphological closing operation and remove any remnants of the *foreground* pixel. Lastly, we derive the statistics on the amount of true positive and false positive staining within the segmented regions as part of the quality features and for an early assessment of the tissue sufficiency.

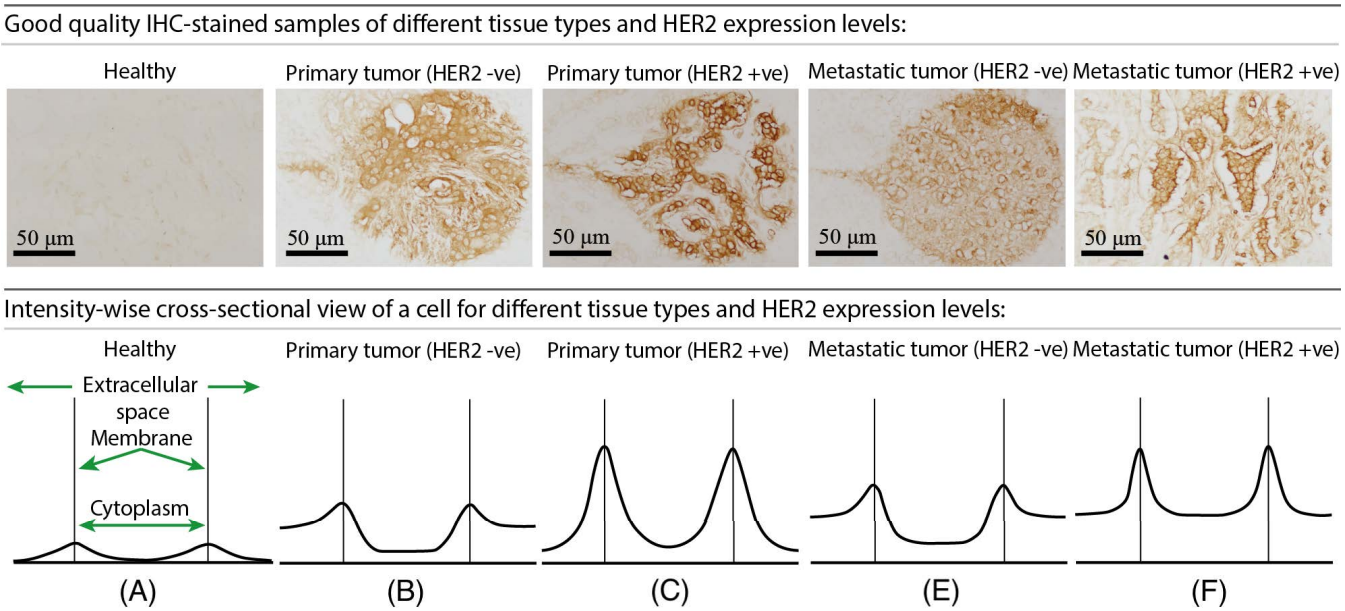

Fig. 4. Variability in immunostaining quality expectations with HER2 antibody for different disease types: (top row) Samples of high-quality immunostaining, and (bottom row) cross-sectional intensity profiles across a cell for each category. Note that for transmembrane HER2 antibody, high expression is expected only on the cell membrane for tumorous tissues.

B. Staining Quality Metric Learning

Optimal staining of cell structures reveal the disease type of an IHC-stained tissue, therefore, we consider that the definition of staining quality varies across disease types for a particular tissue type. Fig. 4 presents high-quality HER2-staining of 5 disease types. HER2 is a transmembrane receptor, thus the quantification of its overexpression can be modelled as detecting 'peaks' (cell membranes) versus 'valleys' (cell cytoplasm and stroma), as also depicted in Fig. 4. The 'peaks', and 'valleys' model represents the intensity profiles along the cross-sectional view of a cell for each disease type. The model indicates that a HER2+ tumorous tissue exhibits a high contrast, whereas a HER2- tumorous tissue or a healthy tissue exhibits low or no contrast between the 'peaks' and 'valleys'. Considering the variability in staining quality expectations, a unique SQM per disease type must be developed. Note that, previous works on staining quality assessment employed a reference-based staining quality estimation, e.g., [13], [14]. In contrast, herein we propose a machine learning-based *no-reference* SQM learning method, which enables to assess the staining quality of a tissue without the need of any reference specimen or user-defined quality threshold.

As per our experimental observations across various disease types, we hypothesize that an immunostaining can be of high-quality, a) if it contains sufficient information (signal) to reflect its disease type, and b) if the contrast level between diagnostically relevant and contextually immaterial signals aligns with the expected contrast level for the corresponding disease type. We develop our quality assessment metrics based on these two quality indicators (Fig. 5). For an IHC-stained tissue, we first capture the disease type information via a probability map indicating its likelihood of being a certain disease type. Secondly, we acquire the contrast information via another probability map indicating the relevant-to-immaterial signal contrast level irrespective of its disease type. We then learn disease type-specific SQMs based on these two pillars in our proposed staining quality assessment framework. Through further analysis, additional quality indicators may be included to improve the framework.

1) Disease Type Quality Indicator: Breast tissues can be categorized into 3 types, namely, healthy tissue adjacent to the tumor (HT), primary tumor tissue (PT), and lymph-node metastasis tissue (MT). On staining the tissues with HER2 biomarker, the latter two can present either an overexpression (HER2+) or a weak overexpression (HER2-) based on the aggressiveness of the cancer. Thereby, HER2-stained breast tissues can be categorized into 5 disease types, namely, HT, HER2+ PT (PT+), HER2- PT (PT-), HER2+ MT (MT+) and HER2- MT (MT-). We propose to train a 5-class supervised probabilistic classifier to identify the disease type of an IHC-stained tissue, and capture the first quality indicator.

2) Contrast Level Quality Indicator Extraction: HER2-stained breast tissues exhibits a certain degree of contrast between the cell membranes and, the cytoplasm and extra-cellular space depending on the aggressiveness of cancer, as presented by the 'peaks', and 'valleys' model in Fig. 4. To obtain the second quality indicator, we propose to train a binary-class supervised probabilistic classifier to identify the contrast level between the diagnostically relevant membrane and contextually immaterial background.

3) SQM Learning & Quality Assessment: We propose to learn disease type specific SQMs considering the unique expectation of staining quality per disease type. The quality indicators acquired for the samples of a particular disease type are used to train an individual SQM. An SQM is learned in a supervised manner using the quality labels for the respective samples obtained from a group of experts. In general, the experts evaluated each sample with various metrics, namely tissue type, antibody expression status, tissue

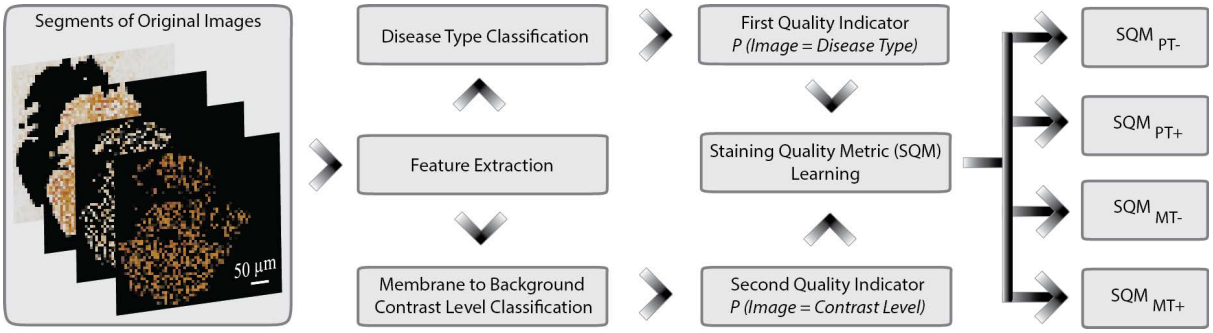

Fig. 5. Overview of the disease-specific staining quality metric (SQM) learning. It involves learning various quality indicators, which currently rely on the output probability maps of two classifiers such as a disease type classifier and a membrane-to-background contrast classifier.

sufficiency, membrane-to-background contrast etc., to assign a quality label $\in$ {Acceptable, NotAcceptable}. An SQM is a probabilistic classifier that learns to predict the quality labels of images using respective two quality indicators. The $P(Image=\text{Acceptable})$ is termed as the quality value (QV) of the image that indicates the acceptability of the image. For instance, a PT+ sample with low membrane-to-background contrast depicts a high-quality staining. Thus, a QV obtained from SQM_{PT+} will have a low value for that sample indicating its low acceptability. In summary, the SQM learning can be defined as:

$$
\begin{aligned}
QI_1 &= P(Image = DiseaseType_{i=1:5}),\\
DiseaseType_i &\in \{\text{HT}, \text{PT}-, \text{PT}+, \text{MT}-, \text{MT}+\},\\
QI_2 &= P(Image = ContrastLevel_{j=1:2}),\\
ContrastLevel_j &\in \{\text{High}, \text{Low}\},\\
SQM_i &= P(Image = QualityLabel_{k=1:2} | QI_1, QI_2,\\
&\qquad Image \in DiseaseType_i),\\
QualityLabel_k &\in \{\text{Acceptable}, \text{NotAcceptable}\},\\
QV_i &= P(Image = \text{Acceptable} | QI_1, QI_2, SQM_i)
\end{aligned}
$$

C. Image & Quality Representation

The supervised probabilistic classifiers for identifying disease type and contrast level in an immunostained tissue are trained using a set of quality relevant features extracted from HER2-stained training images. We propose a comprehensive feature extraction followed by feature selection to obtain a more efficient representation for individual classification task. We experiment with traditional machine learning and deep learning approaches for training the classifiers. The machine learning-based system relies on hand-crafted features, which are shown to be successful in the prior work, whereas the deep architecture is trained with features extracted from a pre-trained network. The individual feature sets are discussed in details in following sections.

1) Hand-crafted Features: Hand-crafted features are extracted both holistically, features from individual segmented regions to capture information about relevant and immaterial signals in the whole image, and locally, patch-wise features from relevant regions to capture local structural and morphological information. Local features are extracted from patches containing a sufficient amount of *foreground*, as segmenting each cell for the analysis is not feasible.

Holistic features: Intensity-based features directly relate to the amounts of relevant and immaterial signal in an image. We extract mean *foreground* intensity (relevant signal strength), mean *footprint* intensity (immaterial signal strength), mean *footprint* intensity from segmented regions, and relative intensity of relevant to immaterial signal. Note that relative intensity feature is used a major quality indicator in [13] and [14].

Percentile features, namely % of *foreground* in the image, % of *foreground* within the *footprint* and % of *footprint* within the image are included to encode the amount of relevant area in the whole image.

Difference of Gaussians is used to detect keypoints on the *foreground* of an image, and SIFT features [27] are extracted around the keypoints. We combine the SIFT features using K-means clustering, K decided by Bayesian information criterion, with bag-of-words to define a fixed-dimensional feature representation for the image.

Local features: We extract texture, spatial and frequency domain features in a patch-wise manner. Texture features in terms of contrast and entropy statistics are obtained from Gray-level co-occurrence matrices with two distance values and four orientations. Additional image-gradient-based sharpness features namely, mean gradient magnitude, mean and standard deviation of blur difference, sharpness and Tenengrad response are extracted as suggested in [28].

We acquire morphological and topological clues in the neighborhood of cells using spatial and frequency domain wavelet features at multi-scale resolution. Gabor wavelet based rotation- and scale-invariant features are extracted using complete and non-orthogonal basis set of Gabor filters with eight rotations and five scaling factors, as in [29]. Discrete Haar wavelet transformation at 3 levels is performed per patch to extract mean, variance, rotation-invariant energy and anisotropy of energy features along horizontal, vertical and diagonal sub-bands, as in [30] and [31]. Visual perceptual directionality, contrast and coarseness features are extracted using [32]. More comprehensive shift invariant Haralick features are extracted from individual sub-images obtained via Dual tree complex wavelet transform of patches [33].

The patch-wise features are extracted from patches with sufficient amount of *foreground*, thereby, making them homogeneous in nature over an image. Hence firstly, we exclude the outlier patches for an image, based on the distance between the per patch feature representation and the mean feature representation, computed using all patches from the

image. Secondly, we compute the mean feature representation across all the remaining patches to define the final feature representation for the complete image.

Feature Selection: Classification-task specific feature selection is performed on the extracted set of features to remove irrelevant and redundant attributes. We use *recursive feature elimination* with Random forest feature importance to select the optimal set of features.

2) Deep Learning-Based Features: Popular pre-trained networks on ImageNet dataset, such as, VGG19 [35] and ResNet50 [36], are used to extract feature representations for the images. Considering the difference between HER2-stained image dataset and ImageNet, we extract more generalizable lower level features using the pre-trained networks. Subsequently, the extracted features are used to train supervised convolutional neural networks (CNNs) to generate the desired staining quality indicators.

D. Sensitivity to Process Parameters

Sensitivity analysis of the staining quality to the staining process parameters benefits in obtaining optimal range for the process parameters for high-quality immunistaining. As the staining expectations differ across disease types, the optimal parameter space depends on the disease type. Therefore, we utilize the quality values (QV) of samples per disease type to perform the sensitivity analysis to parameters, namely antibody concentration/dilution (C) and residence time (RT).

First, QVs are interpolated for all C and RT configurations over the entire parametric space to have a dense and smooth distribution of QV. We triangulate the input data (C, RT, QV), available at specific configurations, with Quickhull algorithm [37] and construct a piece-wise cubic interpolating Bezier polynomial on each triangle [38] for interpolating at desired C and RT configurations using a Clough–Tocher scheme [39]. Second, a smooth 3D manifold is fitted to the QVs on a 3D coordinate system with C, RT and QV as the major axes. The 3D manifold enables a better visualization and more comprehensive statistical analysis of the sensitivity of QV with respect to the staining process parameters.

We perform sensitivity analysis at every configuration using variation quantification, similar to [24]. At a point p_i=(C_i, RT_i, QV_i) on the surface, we calculate the difference vector, v_i, between p_i and its 8-connected neighbors. Then, covariance matrix is computed for v_i, as $\mathcal{C}_i = v_i v_i^H$, where v_i and $\mathcal{C}_i$ signify the degree of change in QV in the neighboring configurations. Eigenvalue decomposition of $\mathcal{C}_i$ quantifies the degree of variation at p_i in different directions. The maximum eigenvalue indicates the degree of maximum variation at p_i and the corresponding eigenvector indicates the direction of maximum deviation. The higher the eigenvalue at a point, the higher the degree of variation, implying a higher sensitivity of staining quality to slight variations in corresponding process parameters at the point. Subsequent to obtaining the disease type specific sensitivity information at all parametric configurations, we can select the operational parameter bounds that produce a high staining quality with a low sensitivity of the quality to variations in the process parameters.

III. Materials

Tissue microarrays (TMAs) (Novusbio, USA) of HT, PT, and MT from different patients were obtained to perform HER2-staining. HER2 is a clinically relevant protein, since it is related to an aggressive tumor progression and is target of the immunoherapeutic agent trastuzumab. TMA cores were graded as HER2+ or HER2- by the vendor depending on their expression levels of protein. TMAs were dried at 60°C for 15 min, dewaxed, rehydrated, and processed with heat induced - antigen retrieval. Peroxidase and protein blocks were applied to the TMA prior to staining as recommended by the vendor. Monoclonal HER2 antibodies (Thermo Fisher Scientific, USA) with concentrations of 6.25, 12.5 or 25 μg/mL were exposed on to the core using an MFP head, that stained the tissue section in a diameter of 300 μm. Each TMA core was patterned with 8 footprints of increasing incubation time between 12 and 289 seconds to generate a gradient. Images of each stained regions were acquired at 40x magnification using a bright field microscope. Exposure time was set to 24 ms with a lamp voltage of 6V, field stop is set to 30.5 mm, and aperture stop to 30.5 mm. The neutral density filter was adjusted for 5.8% transmittance. White balance was automatically adjusted with a region clear of cells on the tissue as a reference prior to imaging. Several tissue specimens were collected per TMA core and each specimen was stained for only a particular antibody concentration and residence time configuration.

IV. Evaluation and Results

The image dataset used for empirical evaluation of the proposed staining quality assessment and sensitivity analysis to process parameters methodology consisted of 488 IHC-stained images from 61 TMA cores across five disease types, namely, HT, PT-, PT+, MT-, MT+. Each image is annotated as Acceptable and NotAcceptable regarding the quality of immunostaining. Our methodology starts with a segmentation of each image into *footprint*, *off-footprint*, *foreground* and *background* regions. Subsequently, hand-crafted and deep learning-based features were extracted to train disease type and membrane-to-background contrast level identifying supervised probabilistic classifiers that returns the two quality indicators. The conducted experiments and the impact of experimental hyperparameters on the extraction of individual quality indicators are explained in detail in the following subsections.

A. Extraction of First Quality Indicator

The first quality indicator conveys the disease type information for an image, which is obtained via a 5-class supervised classifier. A balanced subset of 267 images across all disease types was selected that contained sufficient cell materials and represented the respective disease types for both poor (over- and insufficient staining) and high-quality staining. Both statistical machine learning-based and deep learning-based disease type classifiers were trained to maximize the 10-fold cross-validation F1 score. Details on the training and tuning of individual classifiers are presented below.

TABLE I
DISEASE TYPE CONFUSION MATRIX FOR THE BEST CLASSIFIER TRAINED WITH HAND-CRAFTED FEATURES.

	HT	PT-	PT+	MT-	MT+
HT	52	2	1	0	3
PT–	0	41	2	6	1
PT+	0	2	51	1	6
MT–	1	5	1	51	1
MT+	1	2	9	3	25

1) Traditional Machine Learning-Based Classifier: We extracted 584 hand-crafted features for each image, namely intensity (5), segmentation statistics (3), SIFT (128), texture (22), Gabor (26), discrete wavelet transform (100) and dual-tree complex wavelet transform (300) based features. The acquired features and disease type labels, obtained from the vendor, were used to train a Support Vector Machine (SVM) classifier. To obtain the optimal classifier, several hyperparameters, such as patch size for extracting local features, feature categories, and feature combinations, kernel types and hyperparameters, were fine-tuned as described below in order.

Most of the features are extracted from local patches, hence the choice of patch size has a significant impact on the overall classification performance. We examined with different patch sizes, i.e., 48×48, 64×64, 96×96, 128×128 and 160×160 pixels. The best F1 score was achieved for the classifier trained with a patch size of 64×64 pixels.

We evaluated the impact of individual feature categories on the disease type classification by training separate classifiers for each feature group. Rotation- and scale-invariant Gabor wavelet features performed the best as individual feature types, followed by texture and dual-tree complex wavelet features. Combination of different feature groups significantly improved the classification performance. The combination of intensity, texture, Gabor wavelet and dual-tree complex wavelet feature groups (353 features in total) achieved the best F1 score. A further improvement was attained by ranking and selecting the top features. After feature selection, we obtained the best accuracy by including 70 features, which reduced the total number of features by 5-folds and provided an increment of 5% in overall accuracy, with the optimal set of hyperparameters.

Different types of kernels, namely linear, polynomial with degrees of 3, 5, 7 and 10, sigmoid, radial-basis function (RBF), Hellinger, Jensen-Shanon, with appropriate fine-tuning of hyperparameters were examined with SVM classifier. The best F1 score of **0.823** was achieved for an SVM classifier trained with RBF kernel and optimal feature set. Table I presents the confusion matrix obtained for 5-class disease type classification for the best-trained classifier. The confusion matrix indicates the efficacy of the trained classifier in identifying the tissue-types and HER2-expression status. As expected, most of the confusion occurs between PT- and MT- and between PT+ and MT+, which corresponds to a high similarity in the staining behaviors of the HER2- and HER2+ disease types.

2) Deep Learning-Based Classifier: We augmented the dataset for training a deep network by extracting 50 random patches per image, which were of size 224×224 pixels and hold more than 70% overlapping with the *foreground*. The patches extracted from an image were annotated with the disease type label of the complete image. We employed a ConvNet, pre-trained on ImageNet, to process the patches. Considering the dissimilarity between the HER2-stained IHC patches and ImageNet, we extracted more generalizable features from a lower layer of the network. Subsequently, another shallow CNN was trained using the per-patch extracted representations and disease type labels.

We experimented with two pre-trained ConvNets, VGG19 and ResNet50, in Keras version 2. The features were extracted after the third block in both the architectures that resulted in outputs of size 28×28×256 and 28×28×512 for VGG19 and ResNet50 respectively. The subsequent CNN architecture and network training parameters are presented in Table II. In the testing phase, the trained network predicted disease types for 50 extracted patches from a test image, and majority voting was performed to assign the final disease type to the image. The trained networks with features from VGG19 and ResNet50 pre-trained models achieved **0.758** and **0.834** F1 scores respectively.

The results in Table III indicate that the hand-crafted feature-based method performs similarly to the CNN in the disease type, HER2 expression and tissue type identification tasks. The computed F1 score and Cohen's kappa coefficient indicates very good agreement between the original labels and the predicted labels. The kappa coefficient conveys that the trained classifiers perform well in spite of class-imbalance in

TABLE II
CNN ARCHITECTURE AND NETWORK HYPERPARAMETERS FOR DISEASE TYPE CLASSIFICATION.

Type	Output Size	Block	Strides
convolution	28x28x16	[1x1, 16]	1
max pool	14x14x16	-	2
convolution	14x14x8	[3x3, 8]	1
max pool	7x7x8	-	2
convolution	7x7x4	[3x3, 4]	1
flatten	196	-	-
dropout (50%)	196	-	-
linear	196	-	-
softmax	5	-	-
convolution layer = (convolution + ReLU + batch normalization), batch size = 128, He uniform initialization, Adam optimizer, cross-entropy loss, learning rate=0.01			

TABLE III
COMPARISON OF MACHINE LEARNING-BASED AND DEEP LEARNING-BASED DISEASE TYPE CLASSIFICATION RESULTS.

Approach	Task	F1	Kappa
Hand-crafted + SVM	5-class disease type	0.823	0.779
	3-class HER2 expression	0.921	0.878
	3-class tissue type	0.854	0.773
VGG19 + CNN	5-class disease type	0.761	0.699
	3-class HER2 expression	0.884	0.820
	3-class tissue type	0.802	0.692
ResNet50 + CNN	5-class disease type	0.835	0.793
	3-class HER2 expression	0.925	0.884
	3-class tissue type	0.865	0.791
5-class disease type = (HT, PT-, PT+, MT-, MT+) 3-class HER2 expression = (HT, HER2+, HER2-) 3-class tissue type = (HT, PT, MT)			

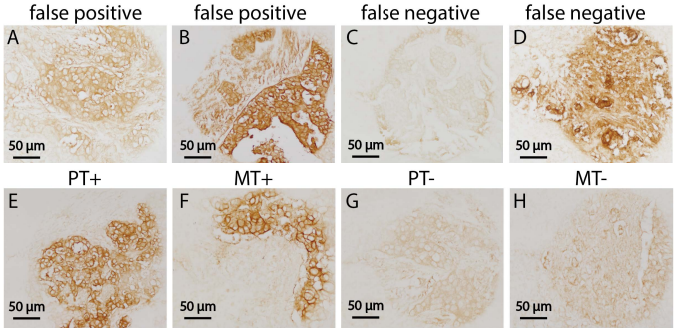

Fig. 6. Misclassified samples by the Hand-crafted + SVM classifier. (top row) False positives, where PT- and MT- samples were interpreted as PT+ (A, B) and false negatives, where PT+ samples were interpreted as PT- and MT- respectively (C, D). (bottom row) Ambiguous PT+ (E), MT+ (F) and PT- (G), MT- (H) samples.

the training dataset. SVM is faster and easier to train, easier to tune hyperparameters and possess easier explainability compared to a deep network. As the results from the SVM and the (ResNet50 + CNN) were comparable, we proceeded with SVM-based development in subsequent tasks. With the inclusion of more images in the dataset in future, the deep networks may then become the method of choice. Fig. 6 presents the misclassified images by the SVM. Overstained PT- and MT- samples (A, B respectively) displayed comparable membrane staining to HER2+, thus being misclassified as PT+ (false positive). Understained and overstained PT+ (C, D respectively) lack sufficient staining contrast between the foreground and the background, similar to HER2- images, thus being misclassified as PT- and MT- (false negative). The bottom row presents the ambiguous PT+ (E), MT+ (F) and PT- (G), MT- (H) samples. Owing to the similarity in the staining behaviour between HER2 overexpressing tissues (PT+ and MT+) and HER2 no over-expressing tissues (PT- and MT-), a few images were misclassified by the classifier.

B. Extraction of Second Quality Indicator

The second quality indicator conveys the membrane-to-background contrast level information of a stain, which is obtained via a binary-class supervised probabilistic classifier. A balanced subset of 77 images were selected that clearly represented high and low contrast levels irrespective of the disease type. We used the 584 features extracted per image to train the contrast level classifier and tuned the same hyperparameters similar to the first quality indicator. The best trained classifier achieved a 5-fold cross validated F1 score of **0.947**. The best classifier was an SVM trained with RBF kernel and with 63 features, which predominantly included features from intensity, Gabor wavelet and dual-tree complex wavelet transform categories.

C. Staining Quality Assessment

The disease type specific SQMs were trained with the two quality indicators extracted from the two supervised probabilistic classifiers. We selected sets of 91, 106, 96 and 60 images from PT-, PT+, MT- and MT+ disease categories respectively with balanced sets of images from both Acceptable and NotAcceptable staining qualities. Individual SQMs were trained in supervised manner using quality labels acquired from the consensus of three experts. Individual SQMs were analyzed using the area-under-the-curve (AUC) measure of the respective receiver operating characteristic (ROC) curves. The optimal SQMs were obtained for SVM with RBF kernel, which achieved AUC scores of **0.84**, **0.83**, **0.82** and **0.90** for the PT-, PT+, MT- and MT+ disease types respectively. An average AUC score of **0.85** is achieved for the proposed methodology with individual SQMs. For comparison, we trained an SQM that learns the staining quality labels for samples directly using their respective feature representations. After tuning all the hyperparameters of the direct SQM, we achieved an overall AUC score of **0.63**. The ROC curves for individual SQMs, the aggregated SQM and the direct SQM are presented in Fig. 7.

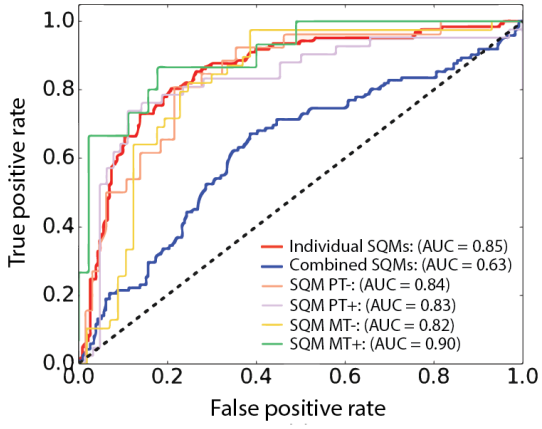

Fig. 7. ROC curves and computed AUC scores for individual SQMs, aggregated behavior of individual SQMs and direct SQM, trained directly with feature sample features and sample quality labels.

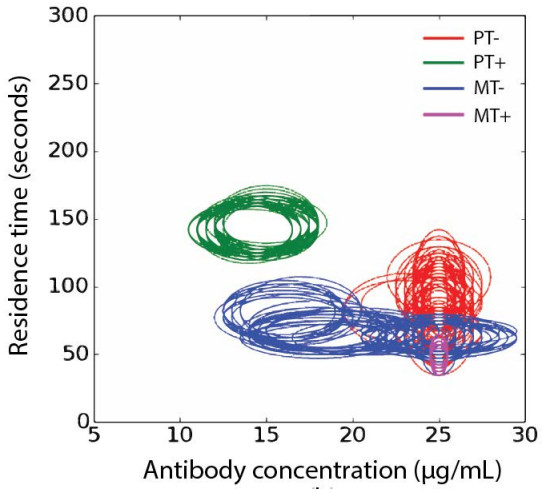

Fig. 8. Disease type specific 95% confidence ellipses fitted to quality values in the process parameter space of antibody concentration and residence time.

The QVs for the samples were acquired from the individual SQMs and the QVs were interpolated over the entire parameter space of C and RT. Subsequently, the disease type specific 3D manifolds were fitted to the QVs, C and RT configurations, as shown in Fig. 9[A-D]. Fig. 9E displays the 2D contour plot of fitted 3D manifold for SQM$_{PT+}$, where the yellow region corresponds to the parameter space with $\geq$95%

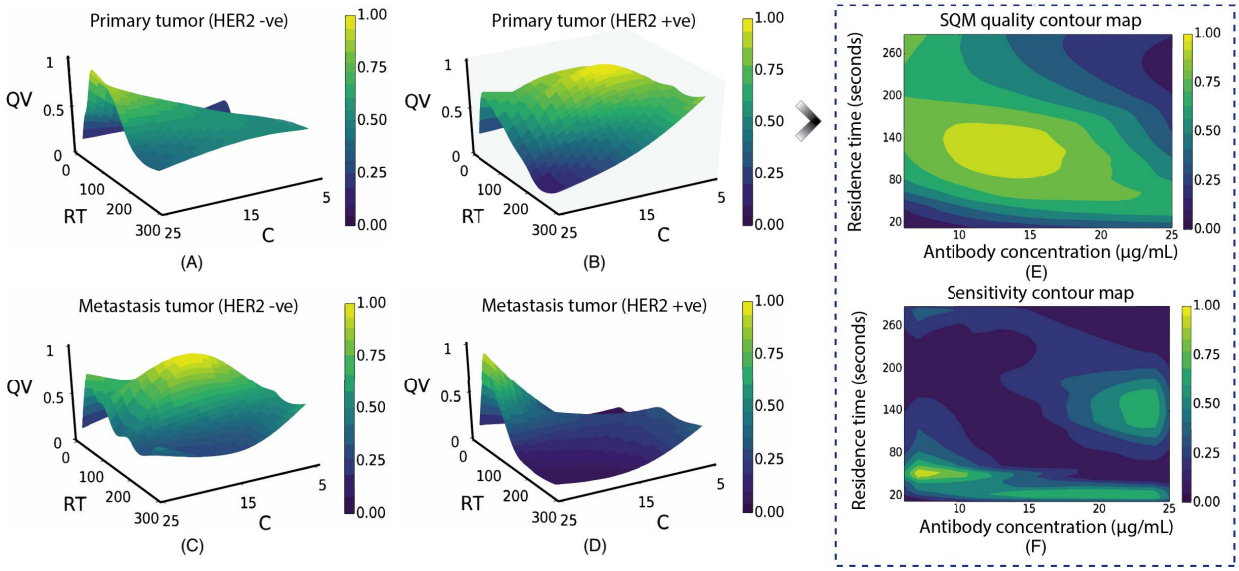

Fig. 9. Manifold fitting to quality values (QVs) acquired using disease type specific SQMs, **(a)** PT-, **(b)** PT+, **(c)** MT-, and **(d)** MT+. The manifolds are fitted using QVs of samples across all configurations of antibody concentrations (C) and residence times (RT). Also shown is, the sensitivity analysis for PT+: **(e)** 2D contour map of QVs and **(f)** 2D contour map of the staining sensitivity to process parameters (C and RT).

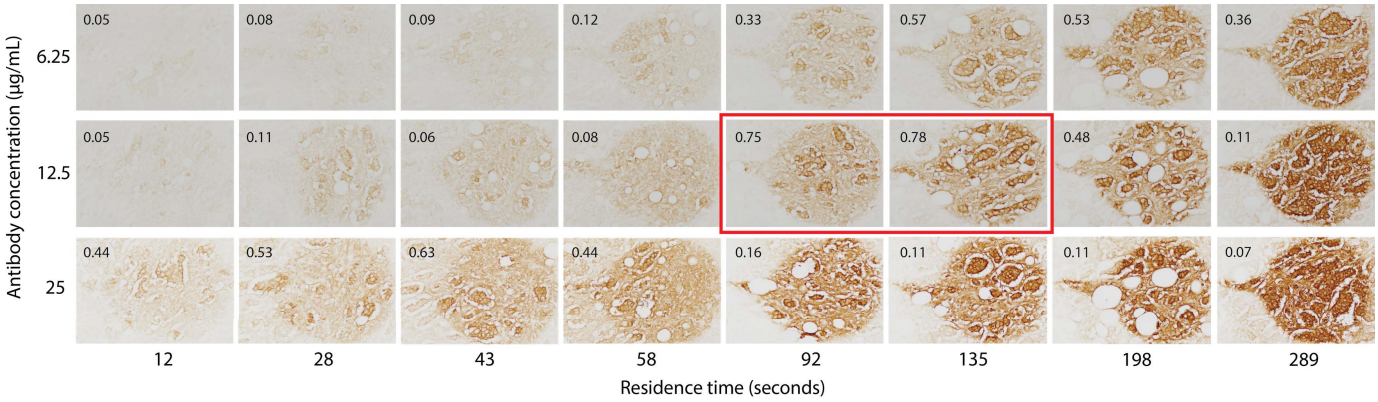

Fig. 10. IHC stained images of a sample PT+ TMA core for the entire parametric configurations of antibody concentration and residence time. The PT+ specific quality values are noted at the top left corner of each image. The best QVs, corresponding to the optimal staining region, are highlighted in red.

staining quality. Fitting an ellipse to this region corresponds to the 95% confidence interval of QV. We evaluated the robustness of our SQM algorithm by generating confidence ellipses for individual SQMs using 200 over-sampled bootstrap datasets for different disease type categories. Fig. 8 displays the confidence ellipses for all the trained disease type specific SQMs. For PT+ and MT+, the confidence ellipses are concentrated in a specific parameter region, whereas for PT- and MT- they are more dispersed. The consistency of the confidence ellipses for PT+ and MT+ indicate that the staining process parameters can be confidently confined to a specific parameters space to achieve high-quality staining, whereas this sort of confinement is not possible for PT- and MT- disease types. Usually in HER2+ tissue sections, the overexpression of the HER2 protein is consistent, resulting in consistent stained-expressions. For HER2- tissue samples, the HER2 protein has weak HER2-overexpression, which can have a high variability in staining expressions. The degree of variability in HER2 overexpression can explain the resulting behavior of the confidence ellipses.

D. Sensitivity to Process Parameters

The disease type specific SQM manifolds were used to evaluate the variability of the staining quality scores with respect to variations in the process parameters. Using the eigenvalue based variational quantification approach, we inspected the sensitivity of the staining quality for all possible process parameter configurations. For instance, Fig. 9E and Fig. 9F present the 2D contour plot of the staining quality and the 2D contour plot of the sensitivity analysis for SQM_{PT+}. Fig. 9E shows that high-quality staining can be obtained when operating in the range of $9 < C < 17\,\mu$g/mL and $90 < RT < 160\,$s and that the best quality is obtained for $C = 14\,\mu$g/mL and $RT = 120\,$s. It also illustrates that the staining quality is low for low-end and high-end C and RT values, which is consistent with the concepts of under-staining and over-staining, respectively. These observations can help reduce false negative and false positive staining, but this map does not convey the stability in operating with these parameter settings. Fig. 9F presents the sensitivity information by plotting the degree of variation in the staining quality for each C and RT configuration. It shows that the staining quality

is slightly sensitive towards the lower-end and the upper-end of the aforementioned range of C values. Combining the knowledge from both the maps, an operational range of $11 < C < 15\,\mu$g/mL and $90 < RT < 130\,$s can be selected for generating stable and high-quality stains for PT+. Fig. 10 presents stained images of a PT+ TMA core for the entire parametric configurations of antibody concentration and residence time showing the staining and quality value variability within a core. The PT+ specific quality values per image are indicated at the top left corner of each image. The best QVs, corresponding to the optimal staining region, are highlighted in red. Similarly, optimal parameter configurations and best practices for other tissue categories and biomarkers can be inferred from their sensitivity analyses.

E. Comparison with a clinical staining protocol

To gauge the value of the proposed method, we aimed to understand whether the methodology can be transferred to a clinical setting. Thus, we performed staining M1, using a protocol used currently in a hospital, and staining M2, the proposed optimized parametrization. These evaluations were performed using in vitro diagnostics antibodied (Herceptest, Dako) with an on-bench approach, i.e. without the use of an MFP for primary antibody deposition. Since the antibody for these tests differs from the antibody used for developing the proposed methodology (anti-Her2 antibody, ThermoFisher), the trained classifiers, SQMs and derived optimum staining configurations could potentially present variations, due to differences in antibody kinetic parameters. However, we expected that the results can be transferable between different antibodies, showing the robustness of the method. Therefore, we analyzed M1 and M2 using the parametric space from our development dataset. To remove bias to antibody selection, we extracted feature representations for the stained cell blocks and scaled the features to the same range as of the features used during the training phase.

For M1, SKBR3 cell blocks were stained using an antibody concentration of 2 μg/mL for 30 minutes, as recommended by the provider (Fig. 11A). The proposed disease type classifier categorized SKBR3 to be MT+, as expected since SKBR3 cells were acquired from a metastatic site. Then for M2, we applied the MT+ specific optimal staining condition, 25 μg/mL for 58 s, as derived by the proposed method (Fig. 11B). The QVs were computed using SQM_{MT+}, and resulted to QVs of 0.12 and 0.88 for M1 and M2 respectively. Both visual inspection and qualitative assessment demonstrate that M2 produced clearer diagnostically relevant information, namely sharper stained membrane signal compared to M1.

Despite the change in antibody, the optimized staining approach delivered a better QV. Hence, we proceeded to stain PT+ tissues with the PT+ specific optimized configuration (antibody concentration of 12.5 μg/mL for 135 s) using clinically validated antibodies on-bench. Fig. 12[A−C] depicts three stained PT+ images. Their QVs were computed to be 0.83, 0.51 and 0.87 respectively using SQM_{PT+}. Despite the much shorter residence time than recommended by the provider (135 s vs 30 minutes), these samples present appropriate staining

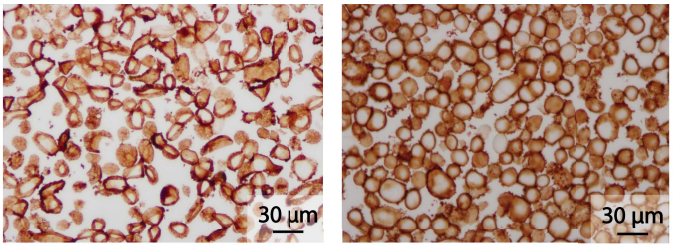

Fig. 11. IHC images of SKBR3 cell blocks stained with clinical protocol and proposed optimized staining method.

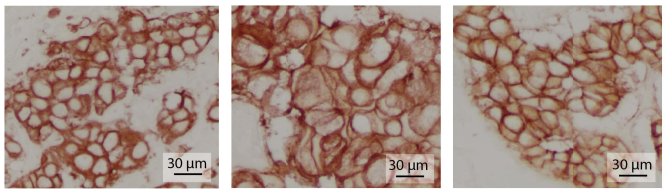

Fig. 12. IHC images of PT+ tissue samples stained with the proposed optimized staining method.

on the cell membranes and possess good relative intensity between the membrane and the cytoplasm, as confirmed by our experts. The estimated QVs classified the images to be in the good staining range.

V. Conclusions

In this paper, we introduce a methodology to analyze the immunostaining quality sensitivity with respect to the staining process parameters, i.e., antibody dilution and residence time. The proposed methodology initially delineates the diagnostically relevant and contextually immaterial signals in a given immunostained tissue. It then learns machine learning based disease type specific staining quality metrics using extracted comprehensive features relevant to staining quality. Subsequently, it performs statistical sensitivity analysis of the staining quality with respect to the process parameters. The proposed quantitative quality metric and the sensitivity analysis contribute to the process parameter optimization to achieve high-quality staining for various disease types. As a model system, the proposed methodology is validated on a cohort of HER2-stained breast cancer tissues from five different disease types, stained using μIHC under various parameter configurations of the MFP. Utilization of the MFP allowed to stain a small fraction of a tissue section for the analysis and extrapolation of suitable process parameters.

We believe that the entire methodology can be extended for other types of staining, disease types, and tumor types as it does not involve any prior assumptions about the staining method. It can easily be applied to the conventional whole-slide-staining and staining with other biomarkers. This allows the comparison of different antibodies leading to the choice of the best and finding the corresponding optimal staining protocols.

With the proposed method, the number of false positives and false negatives produced by incorrect parametrization can be reduced substantially. For instance, when the disease state is unknown, the optimal configuration of PT+ could be applied

as a first approximate set of parameters. In case of a low-quality result for PT+, one of the parameters can be kept fixed, while the other is modified to the closest optimal value from MT+. Performing this sequentially with the aim to maximize the quality metric, a set of optimal parameters can easily be scanned on a tissue. Comparing the information known about the disease type and the one obtained with the algorithm can provide potential information on the developmental status of the tumor.

Digital and computational pathology have received increasing attention from the medical community as they can aid the accuracy of decisions made by pathologists, while also reducing workloads and removing subjective artifacts. However, the underlying aspects of staining quality still remain only partially solved. The suite of methods outlined in this paper can assist pathologists and improve the reproducibility since it establishes an objective metric and reduces the human factor.

REFERENCES

[1] C.R. Taylor. (2000, Jul.). The total test approach to standardization of immunohistochemistry. *Arch Pathol. Lab Med.* 124(7), pp. 945–951.

[2] M. Vyberg and S. Nielsen. (2015, Aug.). Proficiency testing in immunohistochemistry-experiences from Nordic Immunohistochemical Quality Control (NordiQC). *Virchows Arch.* 468, pp. 19–29.

[3] N.S. Goldstein et al. (2007, Jun.). Recommendations for improved standardization of immunohistochemistry. *Appl Immunohistochem Mol Morphol.* 15(2), pp. 124–133.

[4] H. Yaziji et al. (2008, Dec.). Consensus recommendations on estrogen receptor testing in breast cancer by immunohistochemistry. *Appl Immunohistochem Mol Morphol.* 16(6), pp. 513–520.

[5] M.E. Hammond et al. (2010, Jun.). American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J. Clin. Oncol.* 28(16), pp. 2784–2795.

[6] A.C. Wolff et al. (2013, Nov.). Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update. *J. Clin Oncol.* 138(2), pp. 241–256.

[7] E.E. Torlakovic et al. (2014, Apr.). Standardization of negative controls in diagnostic immunohistochemistry: Recommendations from the International Ad Hoc Expert Panel. *Appl. Immunohistochem. Mol. Morphol.* 22(4), pp. 241–252.

[8] E.E. Torlakovic et al. (2015, Jan.). Standardization of positive controls in diagnostic immunohistochemistry: Recommendations from the International Ad Hoc Expert Committee. *Appl Immunohistochem Mol Morphol.* 23(1), pp. 1–18.

[9] R.V. Wasielewski et al. (2008, Dec.). Proficiency testing of immunohistochemical biomarker assays in breast cancer. *Virchows Arch.* 453(6), pp. 537–543.

[10] M. Copete et al. (2011, Jan.). Inappropriate calibration and optimisation of pan-keratin (pan-CK) and low molecular weight keratin (LMWCK) immunohistochemistry tests: Canadian Immunohistochemistry Quality Control (CIQC) experience. *J Clin Pathol.* 64(3), pp. 220–225.

[11] W.J. Howat et al. (2014, Nov.). Antibody validation of immunohistochemistry for biomarker discovery: Recommendations of a consortium of academic and pharmaceutical based histopathology researchers. *Methods.* 70(1), pp. 34-38.

[12] G. O'Hurley et al. (2014, Jun.). Garbage in, garbage out: A critical evaluation of strategies used for validation of immunohistochemical biomarkers. *Mol. Oncol.* 8(4), pp. 783–798.

[13] R. Pinard et al., "Methods and system for validating sample images for quantitative immunoassays," U.S. Patent 8 417 015 B2, Apr. 9, 2013.

[14] M. Grunkin and J.D. Hansen, "Assessment of staining quality," International Patent WO/2015/135550, Sep. 17, 2015.

[15] H. Masmoudi et al. (2009, Jun.). Automated quantitative assessment of HER-2/neu immunohistochemical expression in breast cancer. *IEEE Trans. Med. Imag.* 28(6), pp. 916–925.

[16] A.E. Rizzardi et al. (2012, Apr.). Quantitative comparison of immunohistochemical staining measured by digital image analysis versus pathologist visual scoring. *Diagn. Pathol.* 7, pp. 42.

[17] L. Dobson et al. (2010, Jul.). Image analysis as an adjunct to manual HER-2 immunohistochemical review: A diagnostic tool to standardize interpretation. *Histopathology.* 57(1), pp. 27–38.

[18] A. Brügmann et al. (2012, Feb.). Digital image analysis of membrane connectivity is a robust measure of HER2 immunostains. *Breast Cancer Res. Treat.* 132(1), pp. 41–49.

[19] G.V. Kaigala et al. (2011, May). A vertical microfluidic probe. *Langmuir.* 27(9), pp. 5686–5693.

[20] R.D. Lovchik et al. (2012, Mar.). Micro-immunohistochemistry using a microfluidic probe. *Lab Chip.* 12(6), pp. 1040–1043.

[21] D. Taylor et al. (2016, Sep.). Centimeter-scale surface interactions using hydrodynamic flow confinements. *Langmuir.* 32(41), pp. 10537–10544.

[22] N.M. Arar et al. (2017, Sep.). Computational immunohistochemistry: recipes for standardization of immunostaining. *Proc. MICCAI.* 10434, pp.48–55.

[23] A. Saltelli. (2010, Feb.) Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2), pp. 259–270.

[24] B. Seguin et al. (2014, Apr.) Estimating pattern sensitivity to the printing process for varying dose/focus conditions for RET development in the sub-22nm era. *Proc. SPIE 9050, Metrology, Inspection, and Process Control for Microlithography XXVIII.* 90500P.

[25] E. Zerhouni et al. (2016, Apr.) A computational framework for disease grading using protein signatures. *Proc. ISBI*, pp. 1401–1404.

[26] N. Codella et al. (2016, Mar.) Lymphoma diagnosis in histopathology using a multi-stage visual learning approach. *Proc. SPIE 9791, Medical Imaging 2016: Digital Pathology.* 97910H.

[27] D.G. Lowe. (2004, Nov.) Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision.* 60(2), pp. 91–110.

[28] X. Lopez et al. (2013, Dec.) An automated blur detection method for histological whole slide imaging. *PLoS ONE.* 8(12).

[29] J. Han and K.K. Ma. (2007, Sep.) Rotation-invariant and scale-invariant Gabor features for texture image retrieval. *Journal of Image and Vision Computing.* 25, pp. 1474–1481.

[30] S. Livens et al. (1996, Apr.) A Texture analysis approach to corrosion image classication. *Microscopy, Microanalysis, Microstructures.* 7(2), pp. 143–152.

[31] H. Shan et al. (2014, Jan.) Texture feature extraction based on wavelet transform and gray-level co-occurrence matrices applied to osteosarcoma diagnosis. *Bio-Medical Materials and Engineering.* 24, pp. 129–143.

[32] M. Jian et al. (2009, Aug.) Texture image classification using visual perceptual texture and Gabor wavelet features. *J. of Computers.* 4(8), pp. 763–770.

[33] P. Yang and G. Yang. (2016, Jul.) Feature extraction using dual-tree complex wavelet transform and gray level co-occurrence matrix. *Neurocomputing.* 197, pp. 212–220.

[34] Y.W. Chen and C.J. Lin. (2006) Combining SVMs with Various Feature Selection Strategies. *Feature Extraction. Studies in Fuzziness and Soft Computing.* 207, pp. 315–324.

[35] K. Simonyan and A. Zisserman. (2014) Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.

[36] K. He, X. Zhang, S. Ren and J. Sun. (2015) Deep residual llearning for image recognition. *arXiv:1512.03385*.

[37] C.B. Barber et al. (1996, Dec.) The Quickhull algorithm for convex hulls. *ACM Trans. on Mathematical Software.* 22(4), pp. 469–483.

[38] G. Farin. (1986, Aug.) Triangular Bernstein-Bezier patches. *Computer Aided Geometric Design.* 3(2), pp. 83–127.

[39] P. Alfeld. (1984, Nov.) A trivariate Clough–Tocher scheme for tetrahedral data. *Computer Aided Geometric Design.* 1(2), pp. 169–181.

[40] Z. Mitri et al. (2012, Nov.) The HER2 receptor in breast cancer: pathophysiology, clinical use, and new advances in therapy. *Chemother. Res. Pract.* 2012(2012).

[41] R. Kaufmann et al. (2011, Apr.) Analysis of Her2/neu membrane protein clusters in different types of breast cancer cells using localization microscopy. *J. Microsc.* 242(1) pp. 46–54.

[42] E.A. Perez et al. (2014, Nov.) Trastuzumab plus adjuvant chemotherapy for human epidermal growth factor receptor 2âĂŞpositive breast cancer: Planned Joint Analysis of Overall Survival From NSABP B-31 and NCCTG N9831. *J. Clin. Oncol.* 32(33), pp. 3744–3752.

[43] K.A.B. Goddard et al. (2011, Nov.) HER2 evaluation and its impact on breast cancer treatment decisions. *Public Health Genomics.* 15(1), pp. 1–10.