A whole slide imaging (WSI) scanner scans pathological specimens to produce digital images for monitor-based diagnosis and analysis. However, the image quality is sometimes insufficient due to focus-error or noise, in which case the slide needs to be rescanned. In previous work, a referenceless quality evaluation technique was proposed, but some artifacts (i.e. tissue-fold, air-bubble) were detected as false positives. Those artifacts need to be ignored in determining whether rescanning is necessary or not, because they are not caused in the scanning but slide preparation stage. This paper proposes a method for a more practical system to assess WSI quality by distinguishing the origins of quality degradation; the focus-error or noise caused by the scanner and the artifact occurred in the slide preparation. In the method, a support vector machine detects artifacts first, and then quality is evaluated excluding artifact regions. The effectiveness of the proposed system has been experimentally demonstrated.

Keywords: digital pathology, whole slide imaging scanner, image quality evaluation, tissue artifact detection, pathology specimens.

1. Introduction

A whole slide imaging (WSI) scanner opened the door to digital technology for pathological works 1)~3). The scanner produces high-resolution digital images from the glass slide specimens to use in pathology practice, research, and computational pathology. It allows us to perform image analysis and diagnosis on a computer screen using automated algorithms or human visualization in the hospitals and pathology departments. It extends the limits of light microscopy and holds tremendous promises not only for education and research but also for primary diagnosis and secondary opinions 4). However, for the adoption of a WSI scanner as the routine medical device, it is crucial to consider a number of factors like image quality, color variability, insufficient standardization of slide preparation, and image format 3),5). In this research, we focused on image quality.

In the hospitals and pathology departments, a WSI scanner would be used to scan many slides daily 6). WSI scanners sometimes fail to produce sufficient quality images. Images with poor quality can cause serious complications for diagnosis and unproductiveness for analysis. Currently, the WSI system has to rely on the operator for assuring the quality of the scanned image. However, such manual evaluation is costly, and also not reliable as it is vulnerable to fatigue and could be inconsistent and biased. Therefore, it is necessary to evaluate the quality of scanned images automatically to ensure access to good quality images.

One of the main causes of image quality degradation is a scanning failure, which introduces focus blur or noise in the WSI scanner. Focus points are selected automatically or manually for scanning the glass slides. If focus points are selected from a region of different focus depth than normal tissue area, it makes the neighboring region out of focus. Noise is the independent random value different from neighboring image pixels. Figure 1 shows the example of focus blur, noise, and undegraded tissue area.

If the scanned image is considerably affected by the focus error or noise, it is necessary to rescan the slide. Previously, a referenceless quality evaluation method (RQM) was proposed by Hashimoto et al. to evaluate the quality of the scanned image 7) for the purpose of...
detecting the slides that should be rescanned. This method evaluates the quality of images based on sharpness and noise measurement, which are the main factors for the quality failure in the WSI scanner.

However, there is another cause of image quality failure, which generates the tissue artifacts in the glass slide preparation stage. Tissue fold and air bubble are the major artifacts found in WSI, shown in fig. 2. Tissue artifacts hide or alter significant information and mislead analysis and diagnosis. Moreover, such artifacts result in focusing error. They pose different focus depth, and if focus points are selected from these artifacts, it makes the neighboring region out of focus. In the previous RQM method, tissue artifacts such as air-bubble and tissue-fold would also be detected. Rescanning a slide can recover the quality of an image if it is affected by scanning error but not the tissue artifacts. If a WSI quality is poor due to tissue artifacts only, it is worthless to rescan the slide. Artifacts should be detected and ignored for quality assessment as well as for analysis. Therefore, an efficient system must discriminate the cause of quality failure, so as to utilize the result of the judgment in the workflow to deal with the quality failure.

In this paper, we propose the introduction of the detection step of tissue artifacts and the quality evaluation on focus error and noise with excluding the tissue artifacts, as shown in fig. 3. The proposed method discriminates the cause of quality failure; if the degradation is focus error or noise, the failure is caused in the image scanning process and the slide is rescanned. If the quality failure is the tissue artifacts, rescanning is not needed, and the region will be excluded for the diagnosis or image analysis.

The proposed system utilizes the machine learning technique for artifact detection in the low-resolution WSI. On the other hand, the RQM is done using the high-resolution image, requiring a higher computational cost. The computation for artifact detection is fast because it is done in a low-resolution image. Thus, the computational cost is saved when the artifact is detected since the quality estimation is not required in the artifact region. This paper shows that a practical and efficient system is realized by the proposed approach for the quality evaluation of WSI. Previously, we reported the concept of the proposed system with a result of a preliminary experiment. In this paper, we report the detailed algorithms, results, and discussions on the proposed system with the extension of the method.

2. Related Work

Image quality evaluation methods can be broadly divided into two groups: 1) evaluation using a reference image and 2) without using a reference, though reduced-reference methods are available in some applications. Nevertheless, reference-based evaluation is not suitable for the detection of a quality failure in the WSI scanner system.
as an ideal reference image is not available. The referenceless (or no reference) quality evaluation methods (RQMs) have been developed for the quality assessment in a camera, display, video communication, and codec applications. RQM is essential for evaluating and comparing the performance of different imaging devices. Early RQMs involve specific distortion depending on application requirements, e.g., blurriness, noise, and compression artifacts. Then the metrics that correlate with the subjective evaluation score, such as the mean opinion score (MOS). Multiple distortion factors were also considered, for example, focus blur and noise, by employing the regression to the MOS. Later, the RQMs that consider nonspecific or general distortion types have been developed. Various local and nonlocal image features are computed, and the MOS is predicted by linear or nonlinear regressions, such as neural network or support vector regression. Those methods supposed natural images, and the machine learning was used to learn the statistics of natural images and the features of distortion. Recently the convolutional neural network (CNN) has been applied instead of the handcrafted features. The use of a generative adversarial network has also been proposed. Recently another CNN based method is proposed for evaluating quality considering both synthetic (i.e. compression artifacts) and authentic distortion (i.e. exposure problems) of images.

In WSI applications, the main issue is focus error and noise for quality management of digital pathology. Thus, the RQMs considering specific distortion have been applied. Since computational image analysis is an important issue in digital pathology as well as visual diagnosis, an objective score such as a root mean square error is predicted in addition to subjective quality score in the method proposed by Hashimoto et al. based on sharpness and noise measurement. Shrestha et al. presented a method to evaluate the quality of WSI with the intension to measure the scanner reproducibility. However, this method relies on a reference image for evaluation and not suitable for practical applications. Another method detected out-of-focus regions using deep learning, but the distortion due to noise is not considered. The system proposed in this paper is based on Hashimoto et al.’s method, as it is efficient and straightforward. However, considering the workflow of the WSI scanning process in digital pathology, it is necessary to consider the source of degradation to differentiate between the tissue artifacts and scanning errors, as already shown in fig. 3. In previous methods, in addition to the low-quality blocks caused in the scanning step, tissue artifacts such as air-bubble and tissue-fold would also be detected as false positives in the quality evaluation. The system that automatically analyzes the source of quality degradation and utilizes the results in the subsequent workflow is an important technology, and there has been no previous report on the instance of such technology in practical quality evaluation for WSI.

Object detection or classification in an image is performed in many fields of computer vision. There are supervised and unsupervised approaches. For air-bubble and tissue-fold detection in the proposed system, the supervised method is suitable as the unsupervised method may confuse artifacts with other abnormalities, which may be clinically meaningful. As a supervised technique, template matching is a popular and traditional technique for object detection. Also, statistical methods such as SVM and neural network classifiers have been widely employed for object recognition from a given image. In such supervised classification, handcrafted features have been used, such as local binary pattern, gray level co-occurrence matrix (GLCM), and scale-invariant feature transform. To define the features for classifier, it is known that the performance can be improved by selecting the features using sequential feature selection method (SFS).

More recently, the evolution of deep learning technology enabled non-handcrafted features extracted by CNN from the data. It is also possible to apply CNN directly to an image. It has been shown that the deep-learning-based approach like CNN achieves higher performance than a handcrafted-feature based approach, while it requires a large dataset for training. It tends to be overfitted when used with limited dataset. Transfer learning is often used in such cases to solve the problem of data limitation. Dropout and data augmentation are helpful to prevent a network from overfitting. In this work, we compared the method based on SVM with a VGG16 model pretrained on the ImageNet dataset to investigate whether the current CNN approach is applicable or not for our system in which a limited amount of training data is available.

In histopathology, object detection and classification have been applied to image analysis, and the artifact detection as well. The unsupervised method proposed by Palokangas et al. detects tissue folds by applying k-means clustering on the saturation-intensity
difference image \cite{37}. This method resulted in false positives in the absence of artifacts in a slide. The color enhancement method by Bautista and Yagi detects tissue fold by shifting the saturation \cite{37}. Another method for detecting tissue folds based on the saturation is proposed by Kothari et al. \cite{39} All of these methods are based on the assumption that tissue fold poses higher saturation and lower luminance, but still have room for improvement. For example, a tissue fold in a moderately stained slide has similar color properties like the tissue areas of a highly stained slide. The use of a supervised machine learning technique is simple and effective, for we can employ the same approach for both air-bubbles and tissue-folds. It is also expected to apply CNN-based classification for artifact detection, but it usually needs a larger number of training data. If the target objects of classification are specific, the use of handcrafted features is still effective as it achieves stable classification results while trained with a comparatively smaller dataset.

The proposed method utilizes the RQM for estimating the quality of the image. We combined the artifact detection method with RQM to evaluate slide quality by removing artifacts for the practical system. The results of our experiments show that the proposed method can evaluate scanned image quality by eliminating the false positives due to tissue artifacts and detect slides necessary to rescan. Further, we have compared the proposed method with RQM in terms of its ability to select the appropriate slide to rescan and the time requirements for practical implementation. In the future, the proposed quality evaluation method can be integrated with the automatic image analysis system, such as automatic HER2 quantification \cite{40}.

3. Method

3.1 Overview of the Proposed Method

The proposed method evaluates the quality of WSI by considering the source of quality degradation. If the source of poor quality is artifacts, this method detects them in the first step. Then, it evaluated the quality of the WSI except for detected regions. The decision to use WSI for analysis or rescanning slide is taken based on the quality evaluated regions and artifacts areas are eliminated from this consideration as they are useless for analysis, and rescanning cannot fix the issue. This way, the proposed method achieves the right selection of slides for analysis and rescanning. The proposed method utilizes the machine learning technique in the first stage to detect artifacts and then estimates quality for the rest of the slide based on its sharpness and noise. The algorithm of the proposed evaluation method is illustrated in fig. 4. The previous method did not have the artifact detection step, while this paper proposes the introduction of artifact detection step for a more practical system, where efficient implementation is important.

The size of a WSI data is very large, e.g., 10,000²–100,000² pixels, and the WSI systems utilize a multi-resolution pyramid model where higher magnification gives a higher-resolution image and lower magnification for a low-resolution image. High-magnification images are large in size and contain finer details, which are useful for observing tissue structures.
and anomalies such as focus problems or noise. Lower magnification images are smaller in size and fast to handle. Since the features of the artifacts are not very finely structured, they can be detected well at low resolutions. In the proposed method, low-magnification images (i.e., 1x or 2x) are used in tissue artifact detection, whereas high-magnification images (i.e., 20x or 40x) for quality estimation because the focusing error cannot be detected from the low-magnification image.

The high-resolution WSI is divided into fixed-size non-overlapped image blocks. Image blocks contained more than 75% white pixels are detected as glass area and eliminated for quality estimation. A pixel with an intensity value higher than a certain threshold (Tw=200 in the following experiment) in the grayscale image is considered as the white pixel. Next, the detection of tissue artifacts is carried out for each block, as explained in section 3.2. Then, the quality is estimated for the blocks excluding the tissue artifact affected blocks and mostly white blocks. The details of the quality estimation step are explained in section 3.3. As the artifact detection is performed in the low-magnification image, this architecture ensures efficient, practical implementation.

Finally, the whole slide quality is estimated based on the quality of the evaluated blocks. For example, we represent the sets of mostly white, artifact affected, and tissue blocks as $W, R,$ and $T$ accordingly. The quality is estimated for the blocks that belong to $T$ where $T = 1 - (R \cup W)$. Then the percentage of poor regions in a slide is derived as the output of the system.

3.2 Artifact Detection Method

For artifact detection, the WSI image is divided into fixed-size and non-overlapped image blocks of $100 \times 100$ pixels at 1x magnification. Then, artifact detection is performed on these image blocks. Non-overlapped blocks were utilized to save time for computation and were satisfactory for this application. The proposed method estimates feature values for the blocks and then detected the artifacts using the SVM classifier based on the feature values. Image blocks contained artifacts are eliminated from the quality evaluation. Tissue artifacts pose different texture compared to the regular tissue areas. The proposed method uses two separate SVM binary classifiers to detect the artifacts: one for air bubble detection and the other for tissue fold, using the texture information and other physical properties. In this paper, we investigate two major issues for the artifact detection: (1) feature selection and (2) rotation-invariant classification.

The feature selection $^{41-45}$ is effective in SVM-based classification, especially when the training data is not so large. The features for detecting the artifacts are selected using the SFS method $^{41}$. SFS method was chosen because it is simple to implement and computationally fast. We prepare a set of image features suitable for representing the physical properties of the artifacts, as well as the textural features based on the Haralick’s GLCM features $^{46}$. Then a small number of features are selected by SFS, as explained in the experiment in section 4.2.

If an image signifies the artifact such as air bubble or tissue fold, the image rotated at an arbitrary angle is still the artifact. As GLCM features are not rotation invariant, it is better to apply a rotation-invariant recognition technique. In our experiment, we compared two approaches for rotation-invariant artifacts detection; training data augmentation by including rotated images and the use of rotation-invariant features $^{46, 47}$. We tested the method in $^{47}$, in which principal component analysis (PCA) was applied to the circulant matrix of GLCM features to derive the eigenvalues as the new rotation invariant features.

Here the details of the features used in the experiment are explained; we extracted 46 features from the images, which include 44 GLCM based features and 2 features related to physical property. The GLCM features were the same for both artifacts, while the 2 physical properties were specific to each artifact.

For the image feature calculation, we use the luminance $^{48}$ derived from the pixel values of red, green and blue channels as

$$
Luminance = 0.299 \times Red + 0.587 \times Green + 0.114 \times Blue
$$

(1)

The air bubble poses a different refractive index from the surrounding medium, resulting in a dark boundary when captured by the microscopic imaging system. The characteristics of such a dark boundary can be described by the number of black pixels (luminance $< T_{black}$, where $T_{black}$ is the threshold for the black pixel) and edge pixels in a block. The edge pixels were detected using the Canny’s edge detection algorithm.

In the case of tissue fold, we consider the fact that the tissue fold poses a higher thickness compared to the regular tissue area and have a higher saturation-luminance difference $^{38}$. Then, we use the number of high saturation and low luminance pixels and the
number of edge pixels. High saturation and low luminance pixel mean the pixel whose saturation is higher than luminance, where saturation and luminance are the S and V values of HSV space\(^\text{38}\).

Then GLCMs are calculated for each artifact from the luminance image at a 1-pixel distance for four different directions, which include 0, 45, 95, and 135 degree angles. Among the GLCM features presented by Haralick\(^\text{46}\), we use 11 features which are "contrast," "correlation," "homogeneity," "energy," "entropy," "variance," "sum average," "sum variance," "sum entropy," "difference variance" and "difference entropy." Thus, the 46 candidate features include the 44 GLCM feature derived from the 11 GLCM features calculated for four angles, plus the 2-physical property-based features.

After deriving the candidate features, SFS was used to rank the features for each approach in rotation-invariant artifact detection experiment. Then, we compared the performance of classifiers of different approaches for rotation-invariance to select the suitable classifier for the proposed system.

3.3 Quality Estimation Method

The quality of the WSI is estimated from the high-resolution image based on the sharpness and noise measurements using RQM. The sharpness is the spread of edges where the width of an edge is calculated as the distance between local maxima and local minima. The maximum local gradient is detected for each edge's direction. If the maximum local gradient is higher than the threshold, the corresponding pixel is considered as an edge pixel and the sharpness is estimated for that edge. The width between the local maxima and local minima is measured for all detected edge pixels. Then, the total width of all edges is divided by the number of edges which gives the sharpness degradation index, derived as

\[
\text{Sharpness degradation index}, s = \frac{1}{N} \sum_{i=1}^{N} w(i),
\]

where \(w(i)\) is the width of the \(i\)th edge, and \(N\) is the number of edges.

Out of focus region has a small gradient on edge pixels compared to the sharp regions. Thus, out of focus image blocks have high sharpness degradation value. The noise is measured based on the assumption that it is a random variable, independent of the color channels. Thus, the noise is estimated for each of the RGB channels independently. An unsharp masking technique is utilized to estimate noise where the Gaussian-blurred version of the image is subtracted from the original image. Then, the noise is estimated from the resultant image as

\[
\text{Noise index}, n = \frac{1}{M} \sum_{i=1}^{M} [d_{\text{min}}(j)]^2, \quad (3)
\]

where \(M\) is the number of pixels in the image and \(d_{\text{min}}(j)\) is the minimum difference of pixel values in a 3x3 window centered at the \(j\)th pixel. RQM estimates the subjective or objective quality based on the sharpness and noise index using a linear regression model as:

\[
q = \alpha s + \beta n + \gamma, \quad (4)
\]

where \(q\) is the predicted quality of image blocks, \(s\) is sharpness degradation index, \(n\) is noise index and \(\alpha\), \(\beta\) and \(\gamma\) are coefficients of prediction.

The robustness of the linear model was confirmed in the previous work by comparing it with objective and subjective evaluation scores. In our experiment, we checked the prediction of objective quality score, i.e., the mean square error (MSE). The values of \(\alpha\), \(\beta\) and \(\gamma\) were derived by the linear regression analysis using MSE. We used image blocks of 2000 \times 2000 pixels (20x). Then digitally degraded them to generate blur, noisy and excessive sharp images by adding Gaussian noise (standard deviation was varied from 2 to 20 at an interval of 2), Gaussian blur (standard deviation was varied from 0.2 2 at 0.2 intervals) and applying unsharp masking, accordingly. After that, we compared the quality index values of images with its MSE and found a high correlation\(^\text{11}\). In our experiment the values of \(\alpha\), \(\beta\) and \(\gamma\) were \(-948\), \(156\) and \(31.6\), respectively, similar to the previous report\(^\text{7}\).

Finally, each of the evaluated blocks is ranked as good, medium, and poor block based on the estimated quality \(q\). Then, the percentage of poor areas is estimated for the slide which can be used for selecting slides necessary to rescan.

4. Experiment

4.1 Dataset

4.1.1 WSI Data

In total, 52 slides were used in the experiment, received from the Biomedical laboratory of Shinshu University Hospital. The dataset for air bubble detection, tissue fold detection, and quality evaluation were prepared from those slides. During the experiment, we have evaluated H&E, IHC, and PAS stained slides.
The slides used for the experiments contained major tissue organs (i.e., liver, lung, stomach, brain, intestine, ovary, spleen, heart, etc.) produced from human and rat biopsy. These slides were scanned by two different scanners HAMAMATSU NanoZoomer and 3DHISTECH Pannoramic desk scanner, which ensure the robustness of the proposed method for different scanners.

### 4.1.2 Image Set of Air Bubble Detection

A set of 200 image ROIs (100 x 100 pixels) was exported from 26 whole slide images at 1x magnification as tiff file for the air bubble detection experiment, which included 80 positives and 120 negatives. The image set was divided into two groups: \(AB_{SFS}\ Set\) and \(AB_{Val}\ Set\) where each group contained 100 images of which 40 were positives and 60 were negatives. \(AB_{SFS}\ Set\) was used for feature selection and the \(AB_{Val}\ Set\) was used for validating the classifier. The \(AB_{Val}\ Set\) was further subdivided into \(AB_{Train}\) and \(AB_{Test}\ set\); each contained 50 images of which 20 positives and 30 negatives. While \(AB_{Test}\) additionally contained the 23 rotated versions of its 50 images rotated at 15-degree intervals starting from 15 to 345 degrees. The assignments of images done randomly to have an equal number of positives and negatives in a group. In our experiment, 50% of the dataset was used for feature selection and the other 50% for validating the classifier. The 4-fold cross validation was performed on the entire dataset.

### 4.1.3 Image Set of Tissue Fold Detection

A set of 200 image ROIs (100 x 100 pixels) was exported from 20 whole slide images at 1x magnification as tiff file for the tissue fold detection experiment, which included 80 positives and 120 negatives. The image set was divided into two groups: \(TF_{SFS}\ Set\) for selecting features and \(TF_{Val}\ Set\) for validating classifier. Each group contained 100 images which contained 40 positives and 60 negatives. The \(TF_{Val}\ Set\) was further subdivided into \(TF_{Train}\) and \(TF_{Test}\ set\) to contain 20 positives and 30 negatives in each. \(TF_{Test}\) set additionally contained the 23 rotated versions of its 50 original images. The distribution of data for feature selection, classifier validation, and 4-fold cross validation of tissue fold was the same as the air bubble experiment. The image assignment process was random, like the air bubble experiment as well.

### 4.1.4 Images for Quality Evaluation

In the regression of eq. (4) for quality evaluation, we used 55 images, which were exported from a set of 10 WSI at 20x magnification. The WSI set was different from the one used for the artifact detection experiment. Moreover, the dataset used for the quality evaluation contained an additional 5 WSIs, which were different from the ones used for artifact detection experiment and for training the linear regression model.

### 4.2 Feature Selection

We compared classifiers of different approaches where each classifier contained a different combination of features. The features of the classifier that achieves the best performance in the rotation invariance test were selected.

We performed the baseline detection (BS) method without consideration of rotation invariance, the rotated training image-based (RTI) approach, and the rotation-invariant feature-based (RIF) approach. In the case of RTI approach, training images were generated by rotating images in 23 different angles starting from 15 to 345 degrees at 15-degree intervals. The RIF classifiers were trained using only the original images, but the rotation-invariant features were used.

In the feature selection experiment, we ranked the features of each approach using the SFS method which utilized the 10-fold cross validation with SVM binary classifier. Then, for each approach, we evaluated the performance of the classifier using different number of features starting from 5 to 21 with a step of 2. Fig. 5 shows the performance of classifiers using different number of features for each approach for detecting the air bubble. We relied on the area under curve (AUC) to evaluate the classifier performance and selected the classifier that achieved the highest AUC. Fig. 6 shows the receiver operating characteristic (ROC) curves for the best classifiers of each approach in case of air bubble detection. We selected the RTI classifier with 9 features for air bubble detection as it achieved the highest AUC, shown in fig. 6.

A similar experiment was performed for tissue fold detection. Fig. 7 shows the AUC of classifiers using different number of features for each approach. The ROC curves for the best cases of each approach are plotted in fig. 8. RTI based classifier with 9 features also achieved the highest AUC for tissue fold detection and was selected accordingly.

We utilized different datasets for feature ranking using SFS and for validating classifiers based on AUC. Further, in the validation experiment different images were used for training and testing the classifier. Fig. 9 shows the data distribution for evaluating air bubble classifiers. Tissue fold classifiers were validated.
The SFS method selects a suboptimal set of features based on the average classification error from the cross validation experiment. SFS selects the training and test data randomly for each feature in the cross validation. In our experiment to rank features, we executed the SFS method 3 times and selected 3 subsets of 21 features for each approach. Then the classifier was tested by taking the same number of features for 3 subsets. Each subset of features yielded similar classification performance though the order of features varies in each subset. This experiment ensured that the SFS method selected efficient features. Table A1, A2, and A3 in appendix A show the subsets of 21 features for baseline, RTI, and RIF approaches accordingly to detect air bubbles. Table B1, B2, and B3 in appendix B show the subsets for tissue fold detection.

According to the results obtained above, we have selected the RTI approach with 9 features for both air bubble and tissue fold detection. The lists of selected features are shown in table 1 where the error is the average classification error for a feature in 10-fold cross validation.

In the feature selection experiment, features were selected based on the accuracy in 10-fold cross validation. While the performance of the classifiers was tested using AUC, shown in fig. 5 and 7. If more features are used, it is expected to improve the accuracy, but overfitting may result in high accuracy and a low AUC. Therefore, we expect that there can be a peak AUC as similarly.

The SFS method selects a suboptimal set of features based on the average classification error from the cross validation experiment. SFS selects the training and test data randomly for each feature in the cross validation. In our experiment to rank features, we executed the SFS method 3 times and selected 3 subsets of 21 features for each approach. Then the classifier was tested by taking the same number of features for 3 subsets. Each subset of features yielded similar classification performance though the order of features varies in each subset. This experiment ensured that the SFS method selected efficient features. Table A1, A2, and A3 in appendix A show the subsets of 21 features for baseline, RTI, and RIF approaches accordingly to detect air bubbles. Table B1, B2, and B3 in appendix B show the subsets for tissue fold detection.

According to the results obtained above, we have selected the RTI approach with 9 features for both air bubble and tissue fold detection. The lists of selected features are shown in table 1 where the error is the average classification error for a feature in 10-fold cross validation.

In the feature selection experiment, features were selected based on the accuracy in 10-fold cross validation. While the performance of the classifiers was tested using AUC, shown in fig. 5 and 7. If more features are used, it is expected to improve the accuracy, but overfitting may result in high accuracy and a low AUC. Therefore, we expect that there can be a peak AUC as similarly.

The SFS method selects a suboptimal set of features based on the average classification error from the cross validation experiment. SFS selects the training and test data randomly for each feature in the cross validation. In our experiment to rank features, we executed the SFS method 3 times and selected 3 subsets of 21 features for each approach. Then the classifier was tested by taking the same number of features for 3 subsets. Each subset of features yielded similar classification performance though the order of features varies in each subset. This experiment ensured that the SFS method selected efficient features. Table A1, A2, and A3 in appendix A show the subsets of 21 features for baseline, RTI, and RIF approaches accordingly to detect air bubbles. Table B1, B2, and B3 in appendix B show the subsets for tissue fold detection.

According to the results obtained above, we have selected the RTI approach with 9 features for both air bubble and tissue fold detection. The lists of selected features are shown in table 1 where the error is the average classification error for a feature in 10-fold cross validation.

In the feature selection experiment, features were selected based on the accuracy in 10-fold cross validation. While the performance of the classifiers was tested using AUC, shown in fig. 5 and 7. If more features are used, it is expected to improve the accuracy, but overfitting may result in high accuracy and a low AUC. Therefore, we expect that there can be a peak AUC as similarly.

The SFS method selects a suboptimal set of features based on the average classification error from the cross validation experiment. SFS selects the training and test data randomly for each feature in the cross validation. In our experiment to rank features, we executed the SFS method 3 times and selected 3 subsets of 21 features for each approach. Then the classifier was tested by taking the same number of features for 3 subsets. Each subset of features yielded similar classification performance though the order of features varies in each subset. This experiment ensured that the SFS method selected efficient features. Table A1, A2, and A3 in appendix A show the subsets of 21 features for baseline, RTI, and RIF approaches accordingly to detect air bubbles. Table B1, B2, and B3 in appendix B show the subsets for tissue fold detection.

According to the results obtained above, we have selected the RTI approach with 9 features for both air bubble and tissue fold detection. The lists of selected features are shown in table 1 where the error is the average classification error for a feature in 10-fold cross validation.

In the feature selection experiment, features were selected based on the accuracy in 10-fold cross validation. While the performance of the classifiers was tested using AUC, shown in fig. 5 and 7. If more features are used, it is expected to improve the accuracy, but overfitting may result in high accuracy and a low AUC. Therefore, we expect that there can be a peak AUC as similarly.

The SFS method selects a suboptimal set of features based on the average classification error from the cross validation experiment. SFS selects the training and test data randomly for each feature in the cross validation. In our experiment to rank features, we executed the SFS method 3 times and selected 3 subsets of 21 features for each approach. Then the classifier was tested by taking the same number of features for 3 subsets. Each subset of features yielded similar classification performance though the order of features varies in each subset. This experiment ensured that the SFS method selected efficient features. Table A1, A2, and A3 in appendix A show the subsets of 21 features for baseline, RTI, and RIF approaches accordingly to detect air bubbles. Table B1, B2, and B3 in appendix B show the subsets for tissue fold detection.

According to the results obtained above, we have selected the RTI approach with 9 features for both air bubble and tissue fold detection. The lists of selected features are shown in table 1 where the error is the average classification error for a feature in 10-fold cross validation.

In the feature selection experiment, features were selected based on the accuracy in 10-fold cross validation. While the performance of the classifiers was tested using AUC, shown in fig. 5 and 7. If more features are used, it is expected to improve the accuracy, but overfitting may result in high accuracy and a low AUC. Therefore, we expect that there can be a peak AUC as similarly.

The SFS method selects a suboptimal set of features based on the average classification error from the cross validation experiment. SFS selects the training and test data randomly for each feature in the cross validation. In our experiment to rank features, we executed the SFS method 3 times and selected 3 subsets of 21 features for each approach. Then the classifier was tested by taking the same number of features for 3 subsets. Each subset of features yielded similar classification performance though the order of features varies in each subset. This experiment ensured that the SFS method selected efficient features. Table A1, A2, and A3 in appendix A show the subsets of 21 features for baseline, RTI, and RIF approaches accordingly to detect air bubbles. Table B1, B2, and B3 in appendix B show the subsets for tissue fold detection.

According to the results obtained above, we have selected the RTI approach with 9 features for both air bubble and tissue fold detection. The lists of selected features are shown in table 1 where the error is the average classification error for a feature in 10-fold cross validation.

In the feature selection experiment, features were selected based on the accuracy in 10-fold cross validation. While the performance of the classifiers was tested using AUC, shown in fig. 5 and 7. If more features are used, it is expected to improve the accuracy, but overfitting may result in high accuracy and a low AUC. Therefore, we expect that there can be a peak AUC as similarly.

The SFS method selects a suboptimal set of features based on the average classification error from the cross validation experiment. SFS selects the training and test data randomly for each feature in the cross validation. In our experiment to rank features, we executed the SFS method 3 times and selected 3 subsets of 21 features for each approach. Then the classifier was tested by taking the same number of features for 3 subsets. Each subset of features yielded similar classification performance though the order of features varies in each subset. This experiment ensured that the SFS method selected efficient features. Table A1, A2, and A3 in appendix A show the subsets of 21 features for baseline, RTI, and RIF approaches accordingly to detect air bubbles. Table B1, B2, and B3 in appendix B show the subsets for tissue fold detection.

According to the results obtained above, we have selected the RTI approach with 9 features for both air bubble and tissue fold detection. The lists of selected features are shown in table 1 where the error is the average classification error for a feature in 10-fold cross validation.

In the feature selection experiment, features were selected based on the accuracy in 10-fold cross validation. While the performance of the classifiers was tested using AUC, shown in fig. 5 and 7. If more features are used, it is expected to improve the accuracy, but overfitting may result in high accuracy and a low AUC. Therefore, we expect that there can be a peak AUC as similarly.

The SFS method selects a suboptimal set of features based on the average classification error from the cross validation experiment. SFS selects the training and test data randomly for each feature in the cross validation. In our experiment to rank features, we executed the SFS method 3 times and selected 3 subsets of 21 features for each approach. Then the classifier was tested by taking the same number of features for 3 subsets. Each subset of features yielded similar classification performance though the order of features varies in each subset. This experiment ensured that the SFS method selected efficient features. Table A1, A2, and A3 in appendix A show the subsets of 21 features for baseline, RTI, and RIF approaches accordingly to detect air bubbles. Table B1, B2, and B3 in appendix B show the subsets for tissue fold detection.

According to the results obtained above, we have selected the RTI approach with 9 features for both air bubble and tissue fold detection. The lists of selected features are shown in table 1 where the error is the average classification error for a feature in 10-fold cross validation.

In the feature selection experiment, features were selected based on the accuracy in 10-fold cross validation. While the performance of the classifiers was tested using AUC, shown in fig. 5 and 7. If more features are used, it is expected to improve the accuracy, but overfitting may result in high accuracy and a low AUC. Therefore, we expect that there can be a peak AUC as similarly.
the trade-off between the discriminability and the overfitting. The reason for lower AUC in the case when # of features =11 to 15 is not obvious, but it can be related to the selected features by SFS technique. If we look at table B3, the selected features in RIF technique vary considerably depending on subsets. In other cases (table A1~3, B1-2), many features are common in three subsets. In RIF, PCA is applied to the circulant matrix of GLCM features to derive the eigenvalues as the new features, whose significance may depend on the data subset.

4.3 Artifact Detection

The efficacy of the SVM classifiers for detecting artifacts was tested using 4-fold cross validation. For the 4-fold cross validation of air bubble detection, the entire dataset of 200 images was used as mentioned in section 4.1.2. 50 images were randomly assigned to four groups to have 20 positives and 30 negatives in each group. Then, the images of each group were rotated in 23 different angles from 15 to 345 degrees at 15-degree intervals. After that, the SVM binary classifier for detecting the air bubble was trained and tested using the 4-fold cross validation. The results of the experiment are shown in Tables 2 and 3. The average accuracy was over 98% in 4-fold validation. Further, we evaluated the performance of the SVM classifier using ROC curve analysis and the AUC was 0.9911 ± 0.0094.

A similar 4-fold cross validation experiment was conducted for the tissue fold detection. The dataset of 200 images for tissue fold experiment was apportioned to four groups randomly, so that each group had 20 positives and 30 negatives. Then, the images of each group were rotated in a similar way to have 23 different orientations of each image. After that, the SVM binary classifier selected for detecting the tissue fold was trained and tested in the 4-fold cross validation manner. The performance of the tissue fold detection classifier in 4-fold cross validation is shown in Tables 4 and 5. The average accuracy was over 97%, and the AUC in the ROC analysis was 0.9911 ± 0.0088.

In our experiment, we compared the SVM-based method described above with CNN-based classifiers to investigate whether the current CNN approach is applicable or not to our system in which the amount of training data is limited. We applied data augmentation
by rotating images as explained in section 4.2. Then, we utilized transfer learning by using the trained VGG16 network from the ImageNet data, and then used our dataset for a secondary training\(^{34}\). We freeze the convolution base of the network and derived the bottleneck features for our dataset. The bottleneck features are the last activation maps before the fully-connected (FC) layers of the VGG16 model. Then trained a small FC model using the bottleneck features from our dataset on top of the convolution base. The top FC model contained 3 dense layers including a dropout. 50% of the dataset was used for training and validating the model which includes 40 artifact and 60 undegraded images of 100x100 pixels. A set of 2400 images were generated from the 100 images by augmentation for training and validation. The other 50% was used for testing and augmented in the same way. This experiment was carried out for training two separate models for detecting air bubble and tissue fold.

In the experiment of CNN based classifiers, we investigated the performance of the models for different epochs, batch sizes, learning rates, dropouts, and optimizers, shown in table 6. In air-bubble detection, the best validation accuracy was 0.8733, and in the tissue-fold detection, the best validation accuracy was 0.9217, but a significant difference between training loss and validation loss was noticed while training the VGG16 models as shown in fig. 10. The results indicate overfitting. The AUC in air bubble case was 0.88 and that in tissue fold case was 0.90, which is significantly lower than the SVM based classifiers. Based on these results, we have adopted the SVM classifiers with optimally selected features in the following experiment, as it provided stable and higher classification accuracy than CNN based model for our task.

Further, we have estimated the performance of overall artifact detection for whole slide images. For this experiment, we used 6 whole slide images, which included 3 different types of slide: slide with contained both air bubble and tissue fold, slide contained no tissue artifacts and slides contained either air bubble or tissue fold. Each slide was evaluated by the air bubble detection classifier and by the tissue fold detection classifier. Finally, the result of the artifact detection was measured as a whole. Table 7 shows the overall artifact detection result for 6 slides. The false-negative rate (1-TPR) was considerably low for the overall artifact detection. Slide 4 was the slide that contained no tissue artifacts; thus, no artifacts were detected by any of the classifiers.

### Table 6. Parameter values explored for the hyperparameters.

| Hyperparameter       | Optimization space          |
|----------------------|-----------------------------|
| Epoch                | [20, 30, 50, 70, 100, 150]  |
| Batch size           | [5, 10, 15, 20, 30]         |
| Learning rate        | [0.01, 0.003, 0.05, 0.07]   |
| Dropout threshold    | [0.5, 0.6, 0.7, 0.8]        |
| Optimizers           | Adam, SGD, RMSprop          |

### Table 7. Overall artifact detection accuracy for whole slide images. TPR: true positive rate, SPC: true negative rate & ACC: accuracy, -: no true positives.

|       | TP  | SPC | ACC | # of blocks \(\text{with detected artifacts}\) | # of blocks \(\text{with overlooked artifacts}\) |
|-------|-----|-----|-----|-----------------------------------------------|-----------------------------------------------|
| Slide 1 | 0.75  | 1   | 0.98 | 3 | 1 |
| Slide 2 | 1     | 1   | 1   | 4 | 0 |
| Slide 3 | 0.96  | 0.86 | 0.88 | 24 | 2 |
| Slide 4 | -     | 1   | 1   | 0 | 0 |
| Slide 5 | 1     | 1   | 1   | 3 | 0 |
| Slide 6 | 1     | 0.99 | 0.99 | 1 | 0 |

### 4.4 Quality Evaluation

The linear regression model of eq. (4) was trained using 220 images which included 55 original images and 165 digitally degraded images to derive the quality of image blocks from sharpness degradation index and noise index. The degraded images included simulated blur images, simulated noisy images, and excessive sharp images. The Pearson correlation coefficient between MSE and quality was 0.865.

We have evaluated the quality of WSIs to confirm the validity of the proposed method. In the experiment, 20x...
WSI was divided into non-overlapped image blocks of 2000 × 2000 pixels for quality estimation. Then, the quality of WSI was decided based on the percentage of poor-quality blocks in the slide. In the artifact detection, the SVM classifiers trained and tested using the 4-fold cross-validation were used. The SVM binary classifiers resulted in 1 for artifact detection and 0 otherwise. Fig. 11 shows the quality evaluation result for an H&E-stained slide scanned by Hamamatsu NanoZoomer at 20x. Proposed method estimated the quality of the whole slide image by evaluating the entire slide area with excluding the blocks of detected artifacts.

We have evaluated an image set of 5 WSIs using the proposed method and compared its results with the RQM. The evaluation time and the number of false positives due to artifacts were estimated, shown in table 8. The proposed method resulted in a low percentage of poor-quality regions for a slide compared to RQM in the presence of artifacts in the slide. The proposed method can evaluate the quality of slides without resulting in false positives due to tissue artifacts. Slides necessary to rescan can be selected based on the index of poor-quality regions in slide; shown in table 8. The p-value was 0.028 in one tailed paired t-test between the methods in terms of percentage of poor region for slides which was less than the level of significance (\(\alpha = 0.05\)). Therefore, the null hypothesis is rejected, and the alternative hypothesis is accepted which states that the RQM results in greater poor regions than the proposed method for a slide. Further quantitative evaluation will be performed in a more extensive clinical test in the next step.

The evaluation time was lower for the proposed method compared to RQM in the presence of artifacts in the slide. However, in the absence of artifacts, the time was slightly higher than RQM. Then, the time requirements of the proposed method were analyzed for the same dataset as table 8. For this experiment, a personal notebook PC with a 2.6 GHz Intel Core i5 processor was used without any external graphics card. Fig. 12 shows the process-wise time distribution for the proposed method, which includes the artifact detection and the quality estimation for blocks. The artifact detection process takes a fairly short time while the quality estimation occupies the majority of the total evaluation time. The total time required to evaluate a slide also depends on the amount of tissue area in slide.

| Method | WSI dimensions in pixels | % of poor regions in slide | # of false positives due to artifacts | Time (sec) |
|--------|--------------------------|----------------------------|-------------------------------------|-------------|
| Slide 1 | RQM 18Kx14K              | 34.92                      | 19                                  | 206.7       |
|        | Proposed                 | 4.76                       | 0                                   | 164.9       |
| Slide 2 | RQM 40Kx20K              | 13.5                       | 26                                  | 411.6       |
|        | Proposed                 | 0.5                        | 0                                   | 334.4       |
| Slide 3 | RQM 40Kx26K              | 33.46                      | 11                                  | 443.1       |
|        | Proposed                 | 29.23                      | 0                                   | 408.4       |
| Slide 4 | RQM 20Kx10K              | 42                         | 7                                   | 200.1       |
|        | Proposed                 | 28                         | 0                                   | 147.9       |
| Slide 5 | RQM 40Kx14K              | 76.42                      | 76                                  | 267.2       |
|        | Proposed                 | 22.85                      | 1                                   | 122.0       |

Fig. 11 Quality evaluation of a typical WSI (cropped view). Green overlay indicates artefact affected block. Dark, light and no shade indicate poor, medium and good quality block, accordingly. The blue, cyan, magenta and yellow box indicates false positive in artefact detection, glass, slightly artefact affected, and out of focus block.

Fig. 12 Time requirements of proposed slide quality evaluation method.
the amount of artifact, and glass block detected. A larger tissue area yields in higher evaluation time while a higher number of artifacts detected blocks significantly reduce the evaluation time by eliminating artifact regions from the evaluation. Moreover, the quality estimation process is more time costly than artifact detection. The time requirements are shown in fig. 12 included I/O time (i.e. writing results to the file) which ranges from one to several seconds per slide.

Furthermore, we analyzed the time requirements of the quality estimation process, shown in table 9. The quality estimation process mainly involves the sharpness calculation, noise calculation, detection of glass blocks and I/O operations. From table 9, it can be observed that sharpness calculation is a time-costly step and it occupies the majority of the quality estimation time. In our experiment, it covers from 48% to 60% of the total quality estimation time depending on the slide condition. The sharpness estimation subprocess loop through each edge of the image blocks whose gradient is higher than the threshold. Thus, it is a time-consuming task. However, for a slide highly affected by the focus blur, the sharpness calculation takes a shorter time like slide 5.

Noise estimation costs lower time than the sharpness estimation. In our experiment, noise calculation occupied 30% to 45% of the total time with an exception in the case of slide 5. Detecting glass blocks take a relatively short time, usually less than 5 seconds. The rest of the time is consumed by the I/O and other operations which varies from slide to slide. The time required by each process varies slightly from time to time, usually less than 5 seconds. In our experiment, we measured the time for each process 3 times and took the average, reported in tables 8, 9, and fig. 12.

5. Discussion

The proposed method evaluates the quality of the entire slide by eliminating artifacts and enables efficient judgment of the slide to rescan slide or to store the digital image for further analysis and diagnosis. For this purpose, it is necessary to decide on the threshold for the low-quality index to select slides to rescan. The threshold for rescanning the slide could be decided by the pathologists and experts depending on the purpose of evaluation such as diagnosis or image analysis. Further, the evaluation results can be utilized for analysis using automated tools. Poor quality blocks and artifacts will be ignored for analysis.

The artifact detection is implemented using SVM with SFS in this work. In the experiment, it was compared with the CNN-based method. The use of handcrafted features is still beneficial if the target objects of classification are specific, as it achieves stable classification results while trained with a comparatively smaller dataset. In the future, by collecting more data for tissue artifacts, it will be possible to achieve higher performance on CNN.

The artifact detection in the proposed method achieves good accuracy, but small numbers of false positives and false negatives are unavoidable. False positives of artifact detection are those blocks that are not an artifact but detected as an artifact. Such blocks will be ignored from quality evaluation and later from analysis if the slide is selected for analysis. On the other hand, false negatives are those that contained artifacts but were not detected. Such blocks would be detected as poor regions during quality estimation. The proposed method is designed to integrate with automated analysis, and poor regions will be eliminated from the analysis. Thus, such FN blocks will be ignored during analysis.

The method presented in this paper utilizes several parameters that need to be determined in practice. As the definition of luminance, we used the equation defined in ITU-R BT.601 \( \text{eq. (48)} \), although the definition would not affect the results much. The threshold values for white and black pixels, \( T_w \) and \( T_b \), are selected empirically by observing the scanned image histogram.

|    | # of blocks in slide | # of quality estimated blocks | Total quality estimation time (Sec) | Sharpness estimation time (Sec) | Noise estimation time (Sec) | White pixels test time (Sec) | I/O and others time (Sec) |
|----|----------------------|-------------------------------|------------------------------------|-------------------------------|-----------------------------|----------------------------|--------------------------|
| Slide 1 | 63                    | 41                            | 160.59                             | 97.88                         | 57.08                       | 1.1                       | 4.53                     |
| Slide 2 | 200                   | 120                           | 330.09                             | 180.29                        | 127.57                      | 2.53                      | 14.7                     |
| Slide 3 | 260                   | 144                           | 401.92                             | 211.27                        | 180.04                      | 2.7                       | 7.91                     |
| Slide 4 | 50                    | 43                            | 138.09                             | 84.6                          | 46.56                       | 1                         | 5.93                     |
| Slide 5 | 140                   | 45                            | 115.87                             | 43.26                         | 63.61                       | 1.22                      | 7.78                     |
They are not sensitive, and we used the same $T_w$ presented in ref. 7. In the training data augmentation, the image was rotated at 15 degree intervals, as there was no significant difference in performance between 7.5 and 15 degree intervals for rotations. The artifact detection was implemented with 1x WSI. It was because the accuracy is satisfactory, and time is very short in 1x, whereas in the case of 5x WSI, the time is longer but no significant improvement in the accuracy. WSI scanner selects focus points from the low-magnification image, and this type of artifact detection can be utilized to ignore selecting points from artifact areas.

In our experiment, we used slides of different staining types and different organs to ensure the robustness of the method. The slides were scanned by two different scanners to confirm the efficacy of the method regardless of the scanner models. The resolution of the scanned image varies a little depending on the sampling rate of the scanners at the same magnification. Colors also differ for these scanners. In the future, we plan to experiment with slides scanned by more scanners and staining types.

Time is an important issue for the practical implementation of the WSI scanner system. With the advent of modern scanners, the entire slide can be scanned in less than a minute. The proposed evaluation method is currently computationally expensive that demands 2 to 5 minutes depending on the slide condition. For the practical implementation, the time requirements can be minimized utilizing the parallel computing environment.

The primary goal of this research is to deploy the method within the computer attached to the scanner and evaluate slide quality to select slides for analysis and decide for rescanning. The evaluation results will be incorporated with analysis tools to ignore artifact and poor-quality blocks for analysis.

Further, this method can be integrated within the scanner to select focus points while rescanning the slides. Selecting the focus points in the artifacts affected regions is meaningless and could affect the image quality. Artifact affected regions could be avoided from focus point selection by integrating the proposed method with the scanner's focus point selection algorithm.

5. Conclusion

In this paper, we proposed a method to evaluate the quality of the slide for further operation in the pathology workflow. The proposed method detected tissue artifacts from a low-resolution image and then evaluated the quality of slide from the high-resolution image by eliminating artifacts. It can evaluate the quality of scanned images without being affected by false positives due to artifacts. This primary analysis confirmed the effectiveness of the proposed method for WSI scanner-based pathology. The proposed method enabled the efficient and practical selection of slides to rescan. Thus, it will reduce the rescanning rate of the scanner which is an important criterion to judge scanner performance.

Acknowledgement

This work was supported (in part) by JST-Mirai Program Grant.

References

1) Gilbertson, J. R., Ho, J., Anthony, L., Jukic, D. M., Yagi, Y., & Parwani, A. V.: “Primary histologic diagnosis using automated whole slide imaging: A validation study” *BMC Clinical Pathology*, 6(1) (2006)

2) Ho, J., et al.: “Use of whole slide imaging in surgical pathology quality assurance: design and pilot validation studies.” *Human Pathology*, 37(3), 322-331 (2006)

3) Ghaznavi, F., Evans, A., Madabhushi, A., & Feldman, M: “Digital Imaging in Pathology, Whole-slide Imaging and Beyond” *Annual Review of Pathology: Mechanisms of Disease*, 8(1), 331-359 (2013)

4) Pantanowitz, L.: “Digital images and the future of digital pathology” *Journal of pathology informatics*, 1(1), 15 (2010)

5) Pantanowitz, L., Farahani, N., & Parwani, A.: “While slide imaging in pathology: advantages, limitations and emerging perspectives” *Pathology and Laboratory Medicine International*, 23 (2015)

6) Molin, J., Thorsen, S., & Lundström, C. “Implementation of large-scale routine diagnostics using whole slide imaging in Sweden: Digital pathology experiences 2006-2013” *Journal of Pathology Informatics*, 5(1), 14 (2014)

7) Hashimoto, N., et al.: “Referenceless image quality evaluation for whole slide imaging” *Journal of Pathology Informatics*, 3(1), 9 (2012)

8) Simson, S., Malathi, L., Kumar, G., & Balachander, N.: “Artifact in Histological Section” *Biomedical and Pharmacological Journal*, 9(2), 843-845 (2016)

9) Ekundina, V., Eke, G.: “Common artifacts and remedies in histopathology (a review)” *African Journal of Cellular Pathology*, 4(1), 6-12 (2015)

10) Yagi, Y., & Gilbertson, J. R.: “A relationship between slide quality and image quality in whole slide imaging (WSI)” *Diagnostic Pathology*, 3(Suppl 1) (2008)

11) Hossain, M. S., et al.: “Practical image quality evaluation for whole slide imaging scanner” *Proc. SPIE 10711, Biomedical Imaging and Sensing Conference* (2018)

12) Okarma, K.: “Current Trends and Advances in Image Quality Assessment” *Elettronika Ir Elektrotechnika*, 25(3), 77-84 (2019)

13) Choi MG, Jung JH, Jeon JW.: “No-reference image quality assessment using blur and noise” *Int J Electr Electron Eng*;3:318-22 (2009)

14) Konstantinides K, Natarajan B, Yovanof GS. “Noise estimation and filterin using block-based singular value decomposition” *IEEE Trans Image Process*;3:479-88 (1997)

15) S. Sutharasan.: “No-reference visually significant blocking artifact metric for natural scene images.” *Journal of Signal Processing*, 89(8):1647 - 1652 (2009)

16) Z.M. Parvez Sazzaad, Y. Kawayo and Y. Horita.; “No reference
image quality assessment for JPEG2000 based on spatial features”, Signal Process: Image Communication 23 pp.257-268(2008)
 dodge
 17) R. Venkatesh Babu, S. Suresh and Andrew Perkis.; “No-reference JPEG-image quality assessment using GAP-RBF”, Signal Processing 87 pp.1493-1503(2007)
 18) R. Ferzli and L.J. Karam.: “A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)” [IEEE Transactions on Image Processing, 18(4)1:717-722 (2009)
 19) N.D. Narvekar and L.J. Karam. “A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection” In International Workshop on Quality of Multimedia Experience, pages 87-91 (2009)
 20) S. Varadarajan and L.J. Karam.: “An improved perception-based no-reference objective image sharpness metric using iterative edge refinement” In International Conference on Image Processing, pages 401-404 (2008)
 21) Liu, L., Liu, B., Huang, H., & Bovik, A. C.: “No-reference image quality assessment based on spatial and spectral entropies” Signal Processing: Image Communication, 29(8), 856-863. doi: 10.1016/j.image.2014.06.006 (2014)
 22) W. Xuehui, Li Junli and C. Gang.: “An Image Quality Estimation Model Based on HVS’ TENCON 2006 - 2006 IEEE Region 10 Conference, Hong Kong, pp.1-4 (2006)
 23) C. Fan, Y. Zhang, L. Feng and Q. Jiang, “No reference image quality assessment based on multi-expert convolutional neural networks,” IEEE Access 6, 8934-8943 (2018)
 24) Yang, D., Peltoketo, V., & Kamarainen, J.: “CNN-Based Cross-Dataset No-Referenc Image Quality Assessment” 2019 IEEE/CVF International Conference on Computer Vision Workshop (2019)
 25) Y. Ma, X. Cui, F. Sun and S. Hao, “No-Reference Image Quality Assessment Based on Multi-Task Generative Adversarial Network,” IEEE Access 7, 146893-146902 (2019)
 26) Guan, X., Li, F., & He, L.: “Quality Assessment on Authentically Distorted Images by Expanding Proxy Labels” Electronics, 9(2), 252 (2020)
 27) Shrestha P, Kneepkens R, Vrijens J, Vossen D, Abels E, Hulsken B.: “A quantitative approach to evaluate image quality of whole slide imaging scanners” J Pathol Inform 2016;7:56
 28) Senaras C, Niaz MKK, Lozanski G, Gurcan MN: “DeepFocus: Detection of out-of-focus regions in whole slide digital images using deep learning” PLoS ONE 13(10): e0205387 (2018)
 29) M.S. Hosseini, J.A. Z. Brawley-Hayes, Y. Zhang, L. Chan, K.N. Plataniotis and S. Damaskinos: “Focus Quality Assessment of High-Throughput Whole Slide Imaging in Digital Pathology,” in IEEE Transactions on Medical Imaging, 39, 1, pp.62-74 (2020)
 30) A. Marcano-Cedeño, J. Quintanilla-Domínguez, M.G. Cortina-Januchs and D. Andina, “Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network,” IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society, pp.2845-2850 (2010)
 31) Khaire, U. M., & Dhanalakshmi, R. “Stability of feature selection algorithm: A review.” Journal of King Saud University - Computer and Information Sciences (2019)
 32) Yamashita, R., Nishio, M., Do, R. K., & Togashi, K.: “Convolutional neural networks: An overview and application in radiology.” Insights into Imaging, 9(4), 611-629 (2018)
 33) S.J. Pan and Q. Yang, “A survey on transfer learning,” IEEE Transactions on knowledge and data engineering, 22, 10, pp.1345-1359 (2009)
 34) Zhao, W.: “Research on the deep learning of the small sample data based on transfer learning” AIP Conf. Proc. 1864, 020018-1-020018-8 (2017)
 35) N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. SalakhutdiNov.: “Dropout: A simple way to prevent neural networks from overfitting,” The Journal of Machine Learning Research, 15, 1, pp.1929-1958 (2014)
 36) L. Perez and J. Wang: “The effectiveness of data augmentation in image classification using deep learning,” arXiv preprint arXiv:1712.04621 (2017)
 37) Palokangas, S., Selinummi, J., & Yli-Harja, O.: “Segmentation of Folds in Tissue Section Images” 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2007)
 38) Bautista, P., & Yagi, Y.: “Detection of tissue folds in whole slide images” 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2009)
 39) Wang, M., Kothari, S., & Phan, J.: “Eliminating tissue-fold artifacts in histopathological whole-slide images for improved image-based prediction of cancer grade” Journal of Pathology Informatics, 4(1), 22 (2013)
 40) Hossain, M. S., et al. “Automatic quantification of HER2 gene amplification in invasive breast cancer from chromogenic in situ hybridization whole slide images” Journal of Medical Imaging, 6(4), 047501 (2019)
 41) Whitney, A.: “A direct method of nonparametric measurement selection” IEEE Transactions on Computers, C-20(9), 1100-1103 (1971)
 42) Pudil, P., Novovicova, J., & Kittler, J.: “Floating search method in feature selection” Pattern Recognition Letters, 15(11), 1119-1125 (1994)
 43) Nakariyakul, S., & Casasent, D. P.: “An improvement on floating search algorithms for feature subset selection” Pattern Recognition, 42(9), 1932-1940 (2009)
 44) Somol, P., Pudil, P., Novovicova, J., & Pachl P.: “Adaptive floating search methods in feature selection” Pattern Recognition Letters, 20(11-13), 1157-1163 (1999)
 45) Nakariyakul, S., & Casasent, D. P.: “Adaptive branch and bound algorithm for selecting optimal features” Pattern Recognition Letters, 28(12), 1415-1427 (2007)
 46) Haralick, R.M., K. Shanmugan and I. Dinstein: “Textural Features for Image Classification” IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-3, 1973, pp.610-621 (1973)
 47) Putzu, L., & Roberto, C. D.: “Rotation Invariant Co-occurrence Matrix Features” Image Analysis and Processing -ICIAP 2017 Lecture Notes in Computer Science, 391-401 (2017)
 48) RECOMMENDATION ITU-R BT.601-7 (03/2011): Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios
# Appendix

## A. Features Used for Air Bubble Detection

**Table A1.** List of features used for baseline evaluation of air bubble.

| Rank | Feature               | Error | Feature               | Error | Feature               | Error |
|------|-----------------------|-------|-----------------------|-------|-----------------------|-------|
| 1    | Number of black pixels| 0.15  | Number of black pixels| 0.17  | Number of black pixels| 0.17  |
| 2    | Sum_average_0         | 0.1   | Sum_average_0         | 0.09  | Sum_average_0         | 0.08  |
| 3    | Sum_entropy_0         | 0.08  | Contrast_135          | 0.09  | Sum_average_45        | 0.07  |
| 4    | Diff_variance_135     | 0.03  | Sum_entropy_90        | 0.03  | Sum_average_90        | 0.07  |
| 5    | Sum_average_45        | 0.04  | Contrast_90           | 0.05  | Sum_average_135       | 0.07  |
| 6    | Number of edge pixels | 0.05  | Number of edge pixels | 0.04  | Contrast_90           | 0.08  |
| 7    | Contrast_0            | 0.03  | Variance_0            | 0.03  | Energy_0              | 0.08  |
| 8    | Contrast_45           | 0.03  | Entropy_0             | 0.03  | Correlation_135       | 0.05  |
| 9    | Variance_0            | 0.03  | Variance_45           | 0.03  | Homogeneity_0         | 0.04  |
| 10   | Entropy_0             | 0.02  | Diff_variance_90      | 0.03  | Entropy_45            | 0.03  |
| 11   | Variance_45           | 0.02  | Contrast_0            | 0.05  | Correlation_90        | 0.03  |
| 12   | Variance_90           | 0.01  | Sum_average_45        | 0.06  | Entropy_0             | 0.02  |
| 13   | Contrast_90           | 0.02  | Contrast_45           | 0.06  | Contrast_0            | 0.04  |
| 14   | Contrast_135          | 0.03  | Variance_90           | 0.06  | Correlation_0         | 0.04  |
| 15   | Homogeneity_45        | 0.04  | Variance_135          | 0.06  | Variance_0            | 0.05  |
| 16   | Homogeneity_90        | 0.05  | Diff_entropy_45       | 0.05  | Contrast_45           | 0.06  |
| 17   | Correlation_0         | 0.05  | Homogeneity_45        | 0.06  | Contrast_135          | 0.06  |
| 18   | Correlation_45        | 0.05  | Homogeneity_90        | 0.06  | Correlation_45        | 0.06  |
| 19   | Correlation_90        | 0.05  | Sum_average_90        | 0.06  | Number of edge pixels | 0.06  |
| 20   | Correlation_135       | 0.05  | Sum_average_135       | 0.06  | Homogeneity_45        | 0.06  |
| 21   | Energy_0              | 0.05  | Entropy_45            | 0.06  | Homogeneity_90        | 0.06  |

**Table A2.** List of features used for rotated training image-based approach of air bubble.

| Rank | Feature               | Error | Feature               | Error | Feature               | Error | Feature               | Error |
|------|-----------------------|-------|-----------------------|-------|-----------------------|-------|-----------------------|-------|
| 1    | Number of black pixels| 0.12  | Number of black pixels| 0.15  | Number of black pixels| 0.13  |
| 2    | Sum_average_45        | 0.09  | Sum_average_0         | 0.09  | Variance_0            | 0.08  |
| 3    | Correlation_0         | 0.06  | Correlation_0         | 0.08  | Energy_90             | 0.07  |
| 4    | Homogeneity_90        | 0.05  | Correlation_90        | 0.06  | Energy_0              | 0.05  |
| 5    | Contrast_0            | 0.05  | Homogeneity_0         | 0.04  | Number of edge pixels | 0.05  |
| 6    | Number of edge pixels | 0.04  | Number of edge pixels | 0.04  | Contrast_45           | 0.04  |
| 7    | Contrast_45           | 0.03  | Contrast_0            | 0.03  | Contrast_90           | 0.03  |
| 8    | Contrast_90           | 0.02  | Contrast_45           | 0.02  | Contrast_0            | 0.02  |
| 9    | Contrast_135          | 0.01  | Contrast_90           | 0.02  | Contrast_135          | 0.008 |
| 10   | Correlation_45        | 0.002 | Contrast_135          | 0.01  | Correlation_0         | 0.005 |
| 11   | Correlation_90        | 0     | Correlation_45        | 0.0004| Correlation_45        | 0.004 |
| 12   | Correlation_135       | 0     | Correlation_135       | 0     | Correlation_90        | 0     |
| 13   | Homogeneity_0         | 0     | Homogeneity_45        | 0     | Correlation_135       | 0     |
| 14   | Homogeneity_45        | 0     | Homogeneity_90        | 0     | Homogeneity_0         | 0     |
| 15   | Homogeneity_135       | 0     | Homogeneity_135       | 0     | Homogeneity_45        | 0     |
| 16   | Energy_0              | 0     | Energy_0              | 0     | Homogeneity_90        | 0     |
| 17   | Energy_45             | 0     | Energy_45             | 0     | Homogeneity_135       | 0     |
| 18   | Energy_90             | 0     | Energy_90             | 0     | Energy_45             | 0     |
| 19   | Energy_135            | 0     | Energy_135            | 0     | Energy_135            | 0     |
| 20   | Entropy_0             | 0     | Entropy_0             | 0     | Entropy_0             | 0     |
| 21   | Entropy_45            | 0     | Entropy_45            | 0     | Entropy_45            | 0     |
### Table A3. List of features used for rotation invariant feature-based approach of air bubble.

| Rank | Subset 1 | Subset 2 | Subset 3 |
|------|----------|----------|----------|
|      | Feature | Error | Feature | Error | Feature | Error |
| 1    | Number of black pixels | 0.17 | Number of black pixels | 0.17 | Number of black pixels | 0.17 |
| 2    | Contrast second | 0.1 | Sum average first | 0.1 | Contrast first | 0.09 |
| 3    | Contrast third | 0.08 | Contrast third | 0.07 | Contrast second | 0.1 |
| 4    | Contrast fourth | 0.08 | Sum variance second | 0.05 | Contrast third | 0.08 |
| 5    | Correlation first | 0.08 | Energy third | 0.05 | Contrast fourth | 0.07 |
| 6    | Correlation second | 0.07 | Variance-second | 0.06 | Correlation-second | 0.08 |
| 7    | Correlation third | 0.07 | Sum variance third | 0.06 | Correlation third | 0.07 |
| 8    | Correlation fourth | 0.08 | Sum entropy third | 0.05 | Correlation fourth | 0.06 |
| 9    | Homogeneity first | 0.08 | Correlation third | 0.05 | Homogeneity second | 0.05 |
| 10   | Homogeneity second | 0.06 | Contrast second | 0.05 | Homogeneity third | 0.05 |
| 11   | Homogeneity third | 0.06 | Diff. variance second | 0.05 | Homogeneity fourth | 0.04 |
| 12   | Homogeneity fourth | 0.05 | Diff. variance third | 0.05 | Energy second | 0.05 |
| 13   | Energy second | 0.06 | Homogeneity fourth | 0.06 | Energy third | 0.05 |
| 14   | Energy third | 0.06 | Sum variance fourth | 0.06 | Energy fourth | 0.05 |
| 15   | Energy fourth | 0.07 | Diff. entropy third | 0.06 | Entropy second | 0.05 |
| 16   | Entropy second | 0.08 | Contrast first | 0.07 | Entropy fourth | 0.05 |
| 17   | Entropy third | 0.08 | Correlation first | 0.06 | Variance second | 0.06 |
| 18   | Entropy fourth | 0.08 | Variance first | 0.05 | Variance third | 0.07 |
| 19   | Variance second | 0.09 | Diff. variance first | 0.05 | Variance fourth | 0.08 |
| 20   | Variance third | 0.09 | Contrast fourth | 0.05 | Sum average second | 0.1 |
| 21   | Variance fourth | 0.1 | Sum variance first | 0.05 | Sum average third | 0.09 |

### B. Features Used for Tissue fold Detection

### Table B1. List of features used for baseline evaluation of tissue fold.

| Rank | Subset 1 | Subset 2 | Subset 3 |
|------|----------|----------|----------|
|      | Feature | Error | Feature | Error | Feature | Error |
| 1    | Correlation_90 | 0.17 | Correlation_45 | 0.17 | Correlation_90 | 0.17 |
| 2    | Diff variance_135 | 0.16 | Sum average_0 | 0.15 | Variance_0 | 0.15 |
| 3    | Energy_0 | 0.1 | Contrast_45 | 0.12 | Contrast_45 | 0.11 |
| 4    | Number of edge pixels | 0.05 | Entropy_45 | 0.06 | Sum_entropy_90 | 0.08 |
| 5    | Contrast_45 | 0.05 | Number of edge pixels | 0.04 | Number of edge pixels | 0.06 |
| 6    | Sum_entropy_90 | 0.05 | Entropy_0 | 0.04 | Energy_0 | 0.06 |
| 7    | Contrast_135 | 0.04 | Entropy_90 | 0.04 | Energy_45 | 0.06 |
| 8    | Energy_135 | 0.06 | Sum_average_45 | 0.04 | Energy_90 | 0.06 |
| 9    | Energy_45 | 0.06 | Diff_variance_45 | 0.04 | Entropy_90 | 0.06 |
| 10   | Entropy_45 | 0.06 | High saturation and low luminance pixels | 0.05 | Variance_45 | 0.06 |
| 11   | Energy_90 | 0.06 | Entropy_135 | 0.05 | High saturation and low luminance pixels | 0.06 |
| 12   | Contrast_90 | 0.06 | Energy_0 | 0.06 | Energy_135 | 0.06 |
| 13   | Correlation_135 | 0.07 | Contrast_135 | 0.07 | Sum_average_0 | 0.06 |
| 14   | Homogeneity_135 | 0.08 | Homogeneity_0 | 0.07 | Correlation_0 | 0.06 |
| 15   | High saturation and low luminance pixels | 0.08 | Energy_90 | 0.07 | Homogeneity_45 | 0.07 |
| 16   | Diff variance_0 | 0.06 | Sum_average_90 | 0.07 | Correlation_45 | 0.07 |
| 17   | Diff variance_45 | 0.08 | Variance_0 | 0.07 | Correlation_135 | 0.07 |
| 18   | Correlation_0 | 0.08 | Correlation_0 | 0.07 | Homogeneity_90 | 0.08 |
| 19   | Contrast_0 | 0.08 | Correlation_90 | 0.08 | Variance_90 | 0.08 |
| 20   | Correlation_45 | 0.09 | Correlation_135 | 0.08 | Contrast_90 | 0.09 |
| 21   | Sum_average_0 | 0.11 | Variance_45 | 0.09 | Homogeneity_0 | 0.09 |

### Table B2. List of features used for rotated training image-based approach of the tissue fold.

| Rank | Subset 1 | Subset 2 | Subset 3 |
|------|----------|----------|----------|
|      | Feature | Error | Feature | Error | Feature | Error |
| 1    | High saturation and low luminance pixels | 0.18 | High saturation and low luminance pixels | 0.18 | Correlation_45 | 0.1783 |
| 2    | Entropy_90 | 0.1113 | Entropy_0 | 0.1092 | High saturation and low luminance pixels | 0.1154 |
| 3    | Number of edge pixels | 0.0408 | Number of edge pixels | 0.0392 | Sum_average_0 | 0.0667 |
| 4    | Sum_average_135 | 0.0271 | Sum_variance_0 | 0.0246 | Number of edge pixels | 0.0283 |
| 5    | Variance_45 | 0.0167 | Correlation_0 | 0.0171 | Sum_variance_0 | 0.0154 |
| 6    | Sum_entropy_135 | 0.0129 | Sum_average_45 | 0.0121 | Sum_entropy_135 | 0.0088 |
| 7    | Correlation_90 | 0.0112 | Energy_45 | 0.0104 | Correlation_0 | 0.0071 |
| 8    | Energy_0 | 0.01 | Correlation_135 | 0.0083 | Correlation_90 | 0.0063 |
| 9    | Energy_45 | 0.0096 | Correlation_90 | 0.0063 | Variance_0 | 0.0037 |
Table B3. List of features used for rotation invariant feature-based approach of the tissue fold.

| Rank | Feature                           | Error   | Feature                           | Error   | Feature                           | Error   |
|------|-----------------------------------|---------|-----------------------------------|---------|-----------------------------------|---------|
| 1    | Correlation first                 | 0.17    | Correlation first                 | 0.18    | Correlation first                 | 0.17    |
| 2    | Contrast third                    | 0.13    | Contrast third                    | 0.13    | Contrast second                   | 0.13    |
| 3    | Entropy second                    | 0.11    | Diff. entropy second              | 0.11    | Sum average first                 | 0.11    |
| 4    | Entropy first                     | 0.1     | Entropy third                     | 0.1     | Homogeneity third                 | 0.09    |
| 5    | Energy first                      | 0.07    | Correlation fourth                | 0.1     | Homogeneity first                 | 0.09    |
| 6    | Homogeneity third                 | 0.06    | Variance second                   | 0.08    | Number of edge pixels             | 0.09    |
| 7    | Sum entropy first                 | 0.06    | Number of edge pixels             | 0.08    | Sum average second                | 0.08    |
| 8    | Correlation third                 | 0.05    | Contrast second                   | 0.07    | Entropy second                    | 0.08    |
| 9    | Entropy third                     | 0.05    | Homogeneity third                 | 0.06    | Energy third                      | 0.09    |
| 10   | Contrast fourth                   | 0.06    | Homogeneity fourth                | 0.06    | Homogeneity fourth                | 0.09    |
| 11   | High saturation and low luminance  | 0.05    | Correlation third                 | 0.06    | Entropy third                     | 0.09    |
| 12   | Diff. entropy second              | 0.05    | Sum average first                 | 0.05    | Diff. entropy third               | 0.09    |
| 13   | Homogeneity fourth                | 0.06    | Diff. entropy third               | 0.05    | Diff. entropy second              | 0.09    |
| 14   | Correlation fourth                | 0.05    | Entropy second                    | 0.06    | Sum entropy third                 | 0.09    |
| 15   | Sum variance third                | 0.07    | Diff. variance third              | 0.06    | Contrast third                    | 0.1     |
| 16   | Contrast first                    | 0.09    | Energy fourth                     | 0.07    | High saturation and low luminance | 0.1     |
| 17   | Number of edge pixels             | 0.11    | Energy third                      | 0.09    | Variance fourth                   | 0.1     |
| 18   | Contrast_second                   | 0.12    | Entropy second                    | 0.1     | Diff. variance second             | 0.11    |
| 19   | Energy third                      | 0.12    | Variance fourth                   | 0.11    | Sum average third                 | 0.11    |
| 20   | Diff. variance first              | 0.12    | Sum average second                | 0.11    | Contrast first                    | 0.12    |
| 21   | Diff. entropy third               | 0.12    | Diff. variance second             | 0.11    | Diff. variance first              | 0.14    |

Hossain Md Shakhawat is a Doctoral Student in the School of Engineering, Tokyo Institute of Technology. Simultaneously, he is working as a Graduate Research Assistant in Memorial Sloan Kettering Cancer Hospital. His research roots in digital and computational pathology.

Tomoya Nakamura is an Assistant Professor at Tokyo Institute of Technology. He received his PhD in information science from Osaka University. His current research interest is design of new computational imaging methods.

Fumizaku Kimura received his Ph.D. degree in Medical Science from Kitasato University in 2010. He was a research fellow in Tokyo Institute of Technology from 2010 to 2015. Currently, he is working as a Junior Associate Professor in the School of Health Sciences, Shinshu University.

Yukako Yagi is the Director of Pathology Digital Imaging at Memorial Sloan Kettering Cancer Center. She received her Doctor of Medicine degree from Tokyo Medical University. She has a broad interest in various aspects of medical science, with a primary focus on the development of new technologies in digital imaging.

Masahiro Yamaguchi is a Professor in the School of Engineering, Tokyo Institute of Technology. Since 1989 he has been a faculty member of the same institute. His research includes color and multispectral imaging, holography, and pathology image analysis. From 2011-2017, he was the chair of CIE TC8-07 "Multispectral Imaging."