Optimized detection and segmentation of nuclei in gastric cancer images using stain normalization and blurred artifact removal

Oleg Martos a, Md Ziaul Hoque a,*, Anja Keskinarkaus a, Niko Kemi b, Juha Näpänkangas b, Maarit Eskuri b, Vesa-Matti Pohjanen b, Joonas H. Kauppiä c, Tapio Seppänen a

a Center for Machine Vision and Signal Analysis, Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland
b Department of Pathology, Oulu University Hospital, Finland, and University of Oulu, Finland
c Department of Surgery, Oulu University Hospital, Finland, and University of Oulu, Finland

A R T I C L E   I N F O

Keywords:
Computer-aided diagnosis
Histopathology
Adaptive thresholding
Color deconvolution
Gastric cancer
Nuclei detection and segmentation

A B S T R A C T

Histological analysis with microscopy is the gold standard to diagnose and stage cancer, where slides or whole slide images are analyzed for cell morphological and spatial features by pathologists. The nuclei of cancerous cells are characterized by nonuniform chromatin distribution, irregular shapes, and varying size. As nucleus area and shape alone carry prognostic value, detection and segmentation of nuclei are among the most important steps in disease grading. However, evaluation of nuclei is a laborious, time-consuming, and subjective process with large variation among pathologists. Recent advances in digital pathology have allowed significant applications in nuclei detection, segmentation, and classification, but automated image analysis is greatly affected by staining factors, scanner variability, and imaging artifacts, requiring robust image preprocessing, normalization, and segmentation methods for clinically satisfactory results. In this paper, we aimed to evaluate and compare the digital image analysis techniques used in clinical pathology and research in the setting of gastric cancer. A literature review was conducted to evaluate potential methods of improving nuclei detection. Digitized images of 35 patients from a retrospective cohort of gastric adenocarcinoma at Oulu University Hospital in 1987–2016 were annotated for nuclei (n = 9085) by expert pathologists and 14 images of different cancer types from public TCGA dataset with annotated nuclei (n = 7000) were used as a comparison to evaluate applicability in other cancer types. The detection and segmentation accuracy with the selected color normalization and stain segregation techniques were compared between the methods. The extracted information can be supplemented by patient’s medical data and fed to the existing statistical clinical tools or subjected to subsequent AI-assisted classification and prediction models. The performance of each method is evaluated by several metrics against the annotations done by expert pathologists. The F1-measure of 0.854 ± 0.068 is achieved with color normalization for the gastric cancer dataset, and 0.907 ± 0.044 with color deconvolution for the public dataset, showing comparable results to the earlier state-of-the-art works. The developed techniques serve as a basis for further research on application and interpretability of AI-assisted tools for gastric cancer diagnosis.

1. Introduction

Cancer is the second leading cause of death and disease-adjusted life-year burden only after cardiovascular disease in the world, with gastric cancer being the third leading cause of cancer death [1]. The five-year survival rate of gastric cancer is only 25% [2], and up to 44% [3] in patients undergoing curatively intended surgery. Therefore, timely cancer detection, effective neoadjuvant treatment, and radical surgery are crucial factors in improving survival of patients with gastric cancer. Currently, the decision of administration of neoadjuvant therapy and surgical resection is based on cancer stage, as well as the fitness and comorbidity of the patient [4]. However, there are no robust methods to evaluate if individual’s cancer will respond to neoadjuvant therapy or will develop a disease recurrence. As both neoadjuvant therapy and surgery carry significant morbidity, result in transiently or permanently reduced quality of life, and even mortality [4,5], it is of clinical interest to produce methods for predicting these outcomes.

Since shortly after the invention of the microscope, the prediction of
mortality risk has been one of the key aims of histological analysis done by a pathologist by trial-and-error. Risk prediction is currently mainly based on tumor stage, histological grade of differentiation, and some selected cancer-specific features, e.g., tumor budding [6], allowing categorization of each cancer into few clinically relevant risk groups for treatment decisions. However, the prognosis for patients in each group is widely variable, and more accurate and individual prediction of treatment responses is needed for improving survival and reducing the negative burden of suboptimal treatments. With the introduction of digital microscopy and associated digital processing methods, assessment of predictive factors in cancers, either new or consuming previously too much time in routine workflow, has become possible. The image analysis algorithms applied in cancer pathology are often based on cell or nuclei detection and segmentation, which allow to derive various morphological, textural, and spatial features that are used either on their own or as a complementary information for advanced classification and prediction methods [7]. Such examples include neoadjuvant treatment response prediction in breast cancer [8], rectal cancer [9] and non-small lung carcinoma [10], as well as survival prediction by analyzing nuclei features in renal cell carcinoma [11]. However, these studies and clinical applications are still very sparse, based on small cohorts, and have not proven their reliability to be adapted in regular clinical practice [12].

To produce clinically usable, reliable, and robust image analysis methodologies for treatment response or prognosis prediction, it is essential to conduct large studies with high generalizability. In practice, these studies need to cover multiple countries and centers with preferably thousands of patients, and be applicable to a variety of settings, laboratory practices and scanner models. However, even a single individual’s whole-slide image (WSI) can contain tens of thousands of nuclei, which need to be correctly identified and annotated prior to evaluation of nuclear features using image analysis-based models. Manual annotation of nuclei is a repetitive, cumbersome, and time-consuming task, done by a pathologist. Therefore, the adaptation and evaluation of automated nuclei detection and segmentation supported by image harmonization are needed to produce efficient and robust results.

The purpose of this paper is to review some of the existing nuclei detection and segmentation frameworks for H&E-stained images, and to compare their performance in the setting of gastric and other cancers. The adaptation of digital image and data analysis techniques is expected to match or even surpass the level of manual assessment. The automation of such tasks does not aim to replace the existing practices completely, but to enhance and simplify the overall workflow and function as an assistive tool for clinicians. The novelty is in the combination of basic preprocessing, artifact removal, adaptive thresholding, and color normalization techniques to the gastric cancer H&E-stained images, as no clear base for such digital analysis is established yet in this field. The learning results can be utilized in the development of an automatic evaluation pipeline for cancers, and to support or complement manual clinical pathological analysis.

We organize our article as follows: Section 2 covers the literature review, and Section 3 provides an exhaustive description of the proposed method with detailed steps. The experimental results and comparative evaluation of different techniques are reported in Section 4. Finally, Section 5 and Section 6 summarize our main findings with concluding remarks and future perspectives.

2. Literature review

The segmentation and classification of nuclei and cells are core analysis steps in many histopathology image evaluation tasks. Accurate segmentation is a cornerstone of this process. The process includes detection of the candidate pixels in an image and delineation of these pixels from the background or other elements by a distinctive criterion [7]. In real-world applications, complications arise from the large variations in tissue and cell appearance due to their natural biological diversity, irregular staining and imaging conditions, clustered structures, out-of-focus areas, folding artifacts, disease stages and other reasons. Although extensive research has been put into the accurate nuclei detection and segmentation, it remains a major challenge [13]. For example, the most often used methods, which are based on the evaluation of image intensity or color levels and global or adaptive thresholding, fail with multi-tissue data that exhibit a wide variety of nuclei shapes and intensity levels [14]. Experimentation and extensive iterative adjustments are required to achieve the best performance. Thus, most algorithms lack robustness when applied to an independent dataset, because they are tailored to the images of certain types, magnification level and most often applicable only to a specific cancer [15].

The segmentation methods used in academic and clinical research and applications are represented by watershed segmentation [16] and its variations, template or pattern matching such as local binary pattern (LBP) [17], fast marching [18], graph-cuts-based binarization and multiscale Laplacian-of-Gaussian (LoG) filtering [19], model-based approaches such as ellipse descriptor [20], minimum-model [21] and contour-based minimum-model segmentation [22], active contours from basic level sets [23] to more advanced distance regularized level set evolution [24], marked point process approach [25], multi-pass adaptive voting [26], and others.

The recent advances in machine learning (ML) and deep learning (DL) based image processing have been utilized widely for nuclei, cell, and tissue segmentation. The ML frameworks may apply Support Vector Machines (SVM) [27], Naïve Bayesian, decision trees and random forests or k-means and fuzzy C-means clustering [28,29], whereas DL frameworks employ Fully Convolutional Networks (FCN) [30], generic or task-tailored convolutional neural network (CNN) models, deep CNNs (DCNNs) [31–33], Region Based CNN (R-CNN) and Faster R-CNNs [34], Generative Adversarial Networks (GANs) [35], autoencoders (AEs) such as U-Net [36,37], Context Encoder Network (CE-Net) [38], and many other generic and tailored models.

2.1. Traditional nuclei segmentation methods

The active level sets have been applied successfully in microscopy, including H&E-stained images, for nuclei, lymphocytes, and gland segmentation in breast cancer histopathology, as well as in liver, gastric mucosa, and bone marrow images among many other examples [39]. However, level sets work in general poorly with objects that exhibit blurred or inconsistent boundaries, and often fail to separate tightly clustered elements, which is especially applicable for overlapping nuclei [29]. A fuzzy C-means method addresses such issues to separate nuclei with blurred edges, achieving better results than active contours, while being nearly ten times faster computationally [29].

The LoG is a robust technique that works both with shape and intensity information in the image, but it is sensitive to noise and may not work well with objects of a wide scale range [40]. For nuclei segmentation, Al-Kofahi et al. [19] applied multi-scale LoG to overcome scaling issues. The addition of a clustering algorithm helped to avoid errors due to excessive sensitivity to minor peaks in distance map. To improve the results even further, the authors employed the graph-cuts-based refining segmentation as iterative binary labeling problem with a sequential graph coloring method, to outline borders of individual nuclei in tightly clustered blobs. In general, graph cuts are positioned as an alternative technique to active contours to detect object boundaries, utilizing a weighted undirected graph, where pixels represent nodes and weighted edges outline similarity between these pixels [7].

Classical k-means clustering, and probabilistic models as its extension, have been applied to nuclei segmentation, where cluster separation can be done either by pixel-level color, intensity, texture, or location characteristics [7]. But these methods depend on initialization points and require repetitive executions to achieve acceptable results [41]. The Hough transform has been used for cell and nuclei detection, e.g., as part
of the color normalization technique [42], but it is computationally expensive and is targeted to work with nearly circular objects, which does not fit well to cancerous nuclei segmentation [19].

Watershed segmentation [16] is a popular segmentation technique, and it is one of the most widely used algorithms in nuclei and cell segmentation as it is computationally simple, readily available, and effective [14,39]. The algorithm can be improved by defining local minima only at the objects of interest, which serve as markers. This marker-controlled watershed has much higher efficiency in segmenting and separating only these objects of interest, “flooding” all other noisy elements. The marker selection can be done in multiple ways and often depends on the application-specific conditions [39]. The adaptive H-minima transform has been utilized for foreground marker generation producing enhanced nuclei shape detection [43]. For H&E-stained images of breast cancer, multiple marker types were utilized for segmentation with merging of the resulting nuclei segmentations to obtain the final improved combined view [44]. Two marker types were employed: a) fast radial symmetry transform (FRST) that exploits radial symmetry of nuclei producing foreground and background markers, and b) regional minima with own set of foreground and background markers, adjusted with the help of solidity, boundary saliency and mass displacement evaluations, and the area-based correction for each segmented element. The results can be improved further iteratively, subjecting segmented regions to its own local thresholding [13].

Besides the initial segmentation, the separation of tightly clustered nuclei is in general an overly complex problem and remains a great challenge despite all recent advances [13]. The main issue with widely used watershed transform is the proper separation of overlapping nuclei into individual ones. The FRST performs well in most cases, but it may fail to generate accurate markers on asymmetrical or overly elongated nuclei. This can be compensated by segmentation on regional minima [39], H-maxima transform has been utilized for foreground marker detection, where the final refinement on each segmented region is performed with the gradient-weighted distance transform, k-means, and minimum spanning tree, and submitted to the objective function fitting the result to ellipsoid shapes [44].

The combination of both intensity and distance map of the image can be used as input to the watershed segmentation, reducing over-segmentation levels [13]. The authors [13] improve the result further by model-based merging and model-based correction, utilizing volume and convexity characteristics of detected nuclei. However, such an approach requires definition of models based on trained or statistical data. Alternatively, the local gradient information with minimal a priori information can be used for contour-based “minimum-model” nuclei detection and segmentation [22]. Such a model does not imply that nuclei or cells satisfy a regular “roundish” shape nor specific size, and, thus, it is expected to work well with elements of extreme shapes with simple set of parameters and without prior training.

The marked point processes (MPP) originally introduced for the separation of objects with simple shapes have been successfully applied for accurate extraction of complex-shape objects in breast cancer images [25]. The voting-based methods have been introduced that exploit symmetry properties of cells and nuclei, aggregating votes for each pixel on the contour and, as a result, allowing to identify the center region of nuclei, as they exhibit higher votes than its surrounding areas [26]. The ellipsoid descriptor analysis with cone-shape voting has been proposed that improves the watershed segmentation once the initial distance map is corrected by contours adjusted by elliptical model [20]. The multi-pass adaptive voting (MPAV) utilizes convex property of nuclear boundary and adaptively selects and adjusts gradient information of nuclei contours, producing more accurate results than other voting-based methods [26].

Some images pose higher complexity, especially in detecting epithelial cancerous nuclei. Such irregular shaped nuclei with non-uniform chromatin distribution cause great challenge in proper detection and segmentation. One study for breast cancer proposed the usage of saliency maps with tensor voting, utilizing both magnitude and orientation estimation to detect nuclei, and Loop Back Propagation algorithm driven by Markov Random Field to obtain their boundaries [45]. That was shown to outperform classical algorithms such as the “minimum-model” approach [22] and improved marker-controlled watershed segmentation [39].

Large variations in morphology and texture structures in tumor cells impair their accurate detection and segmentation. The manually derived model-based features often fail on more complex tissue structures and are usually tailored only to specific set of images or cancer types. The texture-based image features can provide more accurate results for instance segmentation. These features can be extracted with gray-level co-occurrence matrices (GLCM), LBP, histograms of oriented gradient and Gabor filters [46]. The pattern-matching ML algorithms can then be applied for image segmentation.

2.2. Deep learning methods for nuclei segmentation

Although classical or ML-based segmentation algorithms require no prior or small amount of model training, they depend greatly on manual parametrization and feature definition which is a laborious process that involves a lot of tweaks and adjustments. The DL-based algorithms show highly accurate results in natural image classification and segmentation surpassing classical methods and human expert level.

CNN models have been used for producing significantly enhanced high-resolution optical microscopy images by improving their resolution, field of view, and depth of field [46]. An augmented reality microscope with near real-time image segmentation with the pretrained FCN has been demonstrated [30]. More advanced models exist already that may offer better performance: (Faster) R-CNN, RetinaNet, Feature Pyramid Net (FPN), U-Net, etc. The AEs, including classical U-Net and its more advanced architectures, such as denoising AE and variational AE, have been applied for nuclei, cell, and tissue segmentation. In medical imaging, U-Net has been employed for cell counting, detection, and morphometry, and for automatic annotation with results comparable in quality to manual labeling [36]. Context encoder network (CE-Net) captures higher-level information and preserves spatial information for medical image segmentation, outperforming original U-Net and other state-of-the-art methods [38]. The HoVer-Net utilizes tailored 50-layer Preact-ResNet50 and subsequent parallel network branches to perform simultaneous nuclei segmentation and classification [47]. The hybrid methods integrate, e.g., DCNNs with marker-controlled watersheds for overlapping nuclei segmentation in histopathology images [37].

Transformers, introduced by Google Brain in 2017 [48], are utilized in large language models and provide remarkable results in natural language processing. Employing sequential processing and self-attention mechanisms, they utilize encoder-decoder approach to learn long-range dependencies in input data. Pre-trained initially on massive web-scale datasets in supervised or self-supervised manner, they can be applied to a wide variety of downstream tasks, offering great generalization [49]. Vision Transformer (ViT) showed that such models can be extended to computer vision tasks matching or even surpassing image classification methods based on CNNs [50]. In application to H&E-stained images on GasHisSDB gastric cancer dataset [51], ViT was demonstrated to outperform classical ML methods, based on color histogram, GLCM and LBP feature extraction techniques. Hybrid methods, e.g., GasHis-Transformer [52], show even greater accuracy. This model combines CNN with a multi-scale visual transformer and applies image normalization to speed up model learning. Initially designed for gastric cancer classification, the authors showed it to be general enough to be transferable to breast cancer and lymphoma image classification, outperforming earlier state-of-the-art DL models.

Using Masked AE (MAE) with pre-trained ViT, Segment Anything Model (SAM) exhibits excellent capabilities of transformers in image segmentation [49]. Even though SAM demonstrates excellent results in
natural image processing, it does not reach sufficient performance on medical images [53]. MedSAM is targeted explicitly for that purpose and provides better results, especially in segmenting small objects with weak boundaries [53].

However, to achieve good performance the transformer-based models need to be pre-trained on very large-scale datasets, which is often problematic with medical images [54]. The Medical Transformer (MedT) works only on self-attention mechanism and is based on gated position-sensitive axial attention and Local-Global (LoGo) training strategy [54]. It overcomes the need for extensive pretrained and still performs sufficiently well, surpassing the performance of classical CNNs and regular transformer-based architectures in segmentation of ultrasound and H&E-stained images. MedT performance has been evaluated in comparison to classical U-Net and Seg-Net [55] on EBHI-Seg H&E-stained image dataset to colon cancer image analysis [56].

3. Methods

3.1. Dataset description

Dataset 1: A cohort of 601 gastric adenocarcinoma patients undergoing gastrectomy at Oulu University Hospital during 1983–2016 were identified. Of these, 18 were excluded based on unavailability of histological slides, leaving 583 patients for analysis. The prospectively collected H&E-stained slides that were originally used in the clinical decision-making were viewed with light microscope, and the slides with deepest cancer invasion were digitized using Aperio AT2 scanner (Leica Biosystems, Wetzlar, Germany). The follow-up was retrieved from 100% complete Causes of Death Registry held by Statistics Finland until December 31st, 2019. The Oulu University Hospital Ethics Committee approved the study (15.2.2016 [51]) and the Finnish National Authority for Medicolegal Affairs (VALVIRA) waived the need for informed consent.

The dataset consists of malign cancer specimens only, as the images are used for nuclei segmentation using the proposed framework with traditional threshold algorithms. From this dataset, expert pathologists manually selected 35 ROIs of assorted sizes to represent a diverse array of gastric cancer image samples. The chosen ROIs illustrate wide variety in tissue heterogeneity and intensity complexity, and extensive variation in H&E staining strength, intentionally including images with artifacts (out-of-focus parts, tissue folds or tear, etc.), muscle tissue, or blood clots.

Dataset 2: For the comparative assessment, the nuclei segmentation is evaluated on the public MoNuSeg dataset collected from The Cancer Genome Atlas (TCGA) in the reference nuclei detection and segmentation study [66], called herein as TCGA dataset. The dataset contains 30 images (~21500 annotated nuclei) cropped to 1000 × 1000-pixel size, which is split into training and test sets of respectively 16 and 14 images. The test set has ~13500 nuclei and the test set ~7000 nuclei. The test set provides two images per cancer type: bladder, breast, colon, kidney, liver, prostate, and stomach. The performance is compared against the original study [66], where evaluation and comparison were done between CellProfiler, Fiji, two CNNs and MedT. For MedT training and inference, the images are resized to 512 × 512 pixels.

3.2. Image annotation processing and ROI selection

The initial hospital dataset preparation and ROI selection was performed by a pathologist in Aperio ImageScope. The manual ROI selection (shown in Fig. 2a) serves as a guide for further nuclei-level annotations. The QuPath is used for pixel-level nuclei boundary annotations. The bounding box has been marked for each ROI to simplify search and mapping of all annotated nuclei to the respective ROIs, as exported annotations may not reflect proper region hierarchy in GeoJSON files. The laborious manual labelling process of 35 selected ROIs produced 9085 nuclei annotations.

The annotated slides were processed by developed scripts into binary mask images (as shown in Fig. 2b and Fig. 2c) that serve as a ground truth (GT). For optimization purposes, the WSI and annotation file handling for ROI identification and annotation grouping are performed as a preliminary step independently from the preprocessing, segmentation, and feature extraction steps.

3.3. Artifact removal with spatially varying blur detection

The Spatially Varying Blur Detection (SVBD) based on the high frequency multiscale fusion and sort transform (HIFST) of gradient magnitudes is selected to detect and remove out-of-focus areas often present in WSI slides [67]. The method was introduced originally as an alternative to various techniques based on the estimation of blur kernels and deconvolution, and the authors describe advantages and superior...
performance in comparison to other state-of-the-art solutions. It does not require any prior knowledge of blur type and level and camera or scanner properties, and it is robust against image distortions and noise (both the original noise and, e.g., the product of image compression). The HiFST utilizes the Discrete Cosine Transform (DCT), which translates a signal from spatial to frequency domain. As the blur smoothens the intensity and color variations, it is associated with the reduction of high frequencies in the respective parts of the image. For each pixel in an image of size $N \times M$, a patch of selected size is taken around the pixel, which the high-frequency DCT coefficients are computed on. The results are combined in a vector and sorted by their absolute values. The multiscale transform employs a set of patches of varied sizes (denoted as a resolution) instead of a single patch size and sorts all coefficients from the union of vectors at different resolutions producing a single vector of multi-scale sorted DCT coefficients $L_{ij}$ for each pixel at coordinates $i$ and $j$ at $m$ different scales:

$$L_{ij} = \text{sort}\left(\bigcup_{r=1}^{m} H_{M_r}^{ij}\right)$$  \hspace{1cm} (1)$$

where $H_{M_r}^{ij}$ is the vector of the absolute values of the high-frequency coefficients for the respective pixel at each scale, and $M_r = 2^{2^{r-1}} - 1$ if odd \cite{67}.

The multiscale approach allows us to deal both with the small-scale and large-scale structures that may be present in the image. For each pixel, the $k$th element of the $L_{ij}$ vector is selected, and the combination of all pixels produces an $L_t$ layer, the matrix of the same size $N \times M$. For better comparison of blurred and unblurred regions, the $L_t$ is subsequently normalized between $[0,1]$ to get $\tilde{L}_t$ layer. The final HiFST consists of a set of normalized $L_t$ layers.

As denoted in \cite{67}, the preprocessing step performs Gaussian filtering to remove high-frequency noise, and the gradient magnitude image $G$ is calculated with the convolving kernels to remove duplicate elements but preserving image structure and shape components:

$$G = \sqrt{(B_x^*h_x)^2 + (B_y^*h_y)^2}$$  \hspace{1cm} (2)$$

where $B_{x}$ is the Gaussian filtered image, $h_x = \begin{bmatrix} 1 & 0 \\ \end{bmatrix}$, and $h_y = \begin{bmatrix} 0 & -1 \\ \end{bmatrix}$.
The image \( G \) is fed to the multiscale HiFST to compute DCT coefficients vector for each pixel. The matrix \( T \) is obtained by max pooling the output \( L \) layers. Additionally, the entropy map \( \omega \) of \( T \) is calculated, which acts as a weighting factor to highlight salient regions in the image. Finally, the blur detection map \( BD \) is computed as the Hadamard product (element-wise multiplication) between matrices \( T \) and \( \omega \):

\[
BD = T \odot \omega
\]

(3)

where \( \circ \) represents the Hadamard product operation.

To reduce the impact of outliers and to emphasize the boundaries, the blur map is smoothed by edge-preserving filters [68]. The WSI blur artifact removal algorithm applied in the developed image processing pipeline is based on the Python version [69] of SVBD from the original paper [67]. Fig. 3 presents the algorithm flow diagram, extending the referred high-level view [69] with the additional steps. The generated blur map is subjected to simple thresholding to produce a binary mask, which is applied to the input image to remove detected blurred areas.

The total area of blurred parts is calculated and subtracted from the ROI area to account for the removal in the GT mask, segmentation performance metrics and any area-based estimations and comparisons for segmented structures.

3.4. Color normalization and color deconvolution

The high-level workflow for color normalization and deconvolution is shown in Fig. 1. The methods that are selected in our experiment are outlined below.

3.4.1. Method 1: contrast-limited adaptive histogram equalization

The method based on adaptive contrast enhancement is used as a base comparison for more advanced algorithms. The RGB image is transformed to HSV space, and the hue channel is selected for further processing. The contrast-limited adaptive histogram equalization (CLAHE) is performed, which enhances the contrast individually for each small tile of the image against the exponential histogram distribution. The contrast factor of 0.02 is selected, that prevents over-saturation of the homogenous parts in the image, where pixels are falling in the same gray level range. This basic method does not produce a separate hematoxylin channel, so the enhanced hue channel is used as input for further segmentation steps.

3.4.2. Method 2: color deconvolution

With the CD method based on Hoque et al. [63], CD matrix is estimated to perform image separation into eosin, hematoxylin, and background channels. The source image is converted to optical density (OD) space [70], according to Eq. 4, excluding the incident light intensity \( I_0 \):

\[
OD = -\log_{10}\left(\frac{I}{I_0}\right) = Ac
\]

where the transmitted light intensity \( I = I_0e^{-Ac} \), \( c \) is the absorption factor of stain in each RGB channel and \( A \) is the amount of the stain with the known level of the incident light intensity \( I_0 \). The stain concentration matrix \( C \) is defined as a composition for all stain vectors. For the non-square stain color appearance (SCA) matrix \( M, C \) can be calculated as the product of Moore–Penrose pseudoinverse matrix, \( M^{-1} \), and the normalized \( OD_{norm} \) image. In practice, it can be calculated via a system of linear equations. The inverse matrix is denoted as a color deconvolution matrix \( D \). The result produces the orthogonal representation of stains’ concentration in the image [70], as shown in Eq. 5. Once \( C \) is obtained, the image can be reconstructed for each stain separately, providing the individual view for each stain.

\[
OD_{norm} = MC \rightarrow C = M^{-1}OD_{norm} \rightarrow C = OD_{norm}
\]

(5)

Here-in, the optimal stain vector estimation is done on the normalized image in OD space by the Principal Component Analysis (PCA) [71]. The PCA decreases dimensionality of OD pixels, producing the principal components coefficients as a \( 3 \times 3 \) matrix by singular value decomposition (SVD) algorithm. The coefficients are then projected as vectors on the plane in 3D space and stain vectors are calculated, using as a seed the original standard hematoxylin and eosin vectors, \( H = [0.64, 0.72, 0.27] \) and \( E = [0.09, 0.95, 0.28] \) proposed originally by Ruifrok et al. [70]. As a result, both deconvolution and reconvolution matrices are obtained. The CD matrix is used for stain separation in the input image, and reconvolution matrix can be used to reconstruct the normalized RGB image from the deconvoluted one.

3.4.3. Method 3: color normalization with SVD

This classical color normalization algorithm, proposed by Macenko et al. [62], converts image to the OD space using Eq. 4 with transmitted light intensity \( I_0 \) set to 255. The transparent pixels, below \( \beta = 0.15 \) threshold, are removed, as these areas have no or little stain concentration, and their exclusion helps to improve stain vector estimation. The SVD-geodesic method is applied to produce SVD directions, which are then projected on the plane and normalized to calculate the angle of each projected point to the first SVD direction. For each stain, the intensity histogram is produced from all pixels in the image that this stain is prevalent in. To account for the intensity variation across the slide, the 99th percentile (\( \alpha = 1 \)) of these intensity values is taken to provide robust approximation of the maximum stain concentration. It allows to normalize the set of images to the same intensity level. Applying the 99th percentile to angles, robust extremes are obtained, and with the conversion of the extreme angles back to OD space, the optimal stain vectors are derived. The stain concentration matrix \( C \) is estimated using optimal stain vectors, according to Eq. 5.

The normalized \( C_{norm} \) matrix is used to produce color normalized RGB image \( I_{norm} \) as given in Eq. 6:

\[
I_{norm} = I_0 e^{-W_I C_{norm}}
\]

(6)

where \( W_I \) is the target stain appearance matrix and \( C_{norm} \) is a transformation matrix of the size respective to the number of stains and color channels. In our trials, \( W_I \) is set to the reference H&E OD matrix, \( HE_{ref} = [0.5626, 0.7201, 0.4062; 0.2159, 0.8012, 0.5581] \). The reference \( HE_{ref} \) can be also obtained from the target images. Similarly, the hematoxylin and eosin images are calculated, which are utilized in further segmentation processing.

3.4.4. Method 4: color normalization with stain color descriptor and color-based classification

Khani et al. [61] introduced a global stain color descriptor (SCD) to determine stain concentration matrix \( C \) both in the source and target
images. The image-specific C matrix is not produced by classical color deconvolution, but instead calculated by a color-based classifier (the readily available basic Random Forest Classifier is used). The algorithm learns the color palette via color quantization, principal color histograms (used for low-dimensional embedding) and classification models utilizing RGB and SCD information. The learning generates stain-specific probability maps, and maps for the image background. For color normalization of the source image a set of statistics (mean, 5th and 95th percentile) is calculated for each channel of the deconvolved target and source image, which are then mapped using a spline-based nonlinear mapping, producing normalized stain channels. The normalized image is reconstructed by Eq. 6.

3.4.5. Method 5: stain deconvolution with ICA in the wavelet domain

Alsubaie et al. [65] performed stain deconvolution with blind source separation via Independent Component Analysis (ICA), on the basis that pixels of each stain would be distributed along the principal axes of different independent components. As highlighted by the authors, the ICA works on the assumption that the source signals are independent and have non-Gaussian distribution, but this is too strong assumption for stained images where multiple stains contribute to the colorization and intensity level of each pixel. In their study, this was corrected by the addition of a distributed wavelet transform (DST) that decomposed each color channel into a series of narrow sub-band images with multiple levels of decomposition.

The input image is converted to OD space by Eq. 4. Then a single-level discrete 2-D DST is applied for each OD channel for 5 levels of decomposition. For all sub-bands at various levels, a kurtosis measure is calculated as a representation of their conformity to Gaussian distribution. All measured values are combined into a single concatenated matrix and sorted in descending order. The 20 sub-bands with the highest kurtosis measure are then selected and fed to the FastICA algorithm [72], that outputs 3 × 3 stain mixing matrix, allowing us to generate stain concentration matrix C, according to Eq. 5. The stain matrix is normalized and sorted by rows having single first minimum and maximum values, producing vectors for the hematoxylin and eosin stains and the background. The normalized image and hematoxylin, eosin and background images can be then constructed by Eq. 6. The algorithm performance can be adjusted by the number of levels and sub-bands.

3.4.6. Method 6: adaptive color deconvolution

The adaptive color deconvolution (ACD) proposed by Zheng et al. [64] does not rely on stain classification as in the color normalization with SCD and color-based classifier by Khan et al. [61]. Instead, the parameters of hematoxylin and eosin stains are estimated simultaneously via an integral optimization. Herein, only main equations for the modified stain concentration matrix, model’s objective function and variables are listed, as adapted from [64]. The learning model samples random set of pixels from the input image and converts them to OD space, where stain concentration matrix C is determined. To improve the model for normalizing stain intensities, the stain-weight matrix W is added as a variable to Eq. 5, extending the regular CD technique:

\[ C = WDOD_{stains} \]

where D is the inverse of SCA matrix M, and \( W = \text{diag}(w_h, w_e, 1) \) with \( w_h \) and \( w_e \) representing weights for hematoxylin and eosin stains, respectively.

The W and M matrices are learned with a gradient descent algorithm on reference images, where M is represented by six independent degree variables denoted as a collection variable \( \varphi \), which defines contribution of each color channel to hematoxylin and eosin stains and background. The objective function \( L(\varphi) \) for the learning model is defined on the prior knowledge as a sum of three separate objective functions: \( L_p(\varphi) \) as control for distribution of pixels for each stain and the minimization of the residual separation, \( L_b(\varphi) \) as balance of the two stains and \( L_s(\varphi) \) as overall density of staining, adapted from [64]:

\[ L(\varphi) = L_p(\varphi) + \lambda_1L_b(\varphi) + \lambda_2L_s(\varphi) \]

where three weight lambda hyper-parameters, \( \lambda_p \), \( \lambda_b \), and \( \lambda_s \) are derived from the prior knowledge of stain composition, \( L_p(\varphi) \) utilizes parameter \( \lambda_p \), \( L_b(\varphi) \) is controlled by the balance parameter \( \eta \), and \( L_s(\varphi) \) by the parameter \( \gamma \) that denotes the level of desired density of staining. The optimization to obtain variables \( \varphi \) and \( W \) can be performed as follows:

\[ \langle \hat{\varphi}, \hat{W} \rangle = \arg \min_{\langle \varphi, W \rangle} L(\varphi, W) \]

In our work, the default \( \lambda_p = 0.002, \lambda_b = 10, \) and \( \lambda_s = 1 \) settings are used, and on the base of experiments \( \eta \) is chosen to be set to 0.6 for both datasets, and \( \gamma \) to 0.5 for the main image dataset, and 0.25 for the public TCGA image dataset [66]. For the improved results, the background pixels with OD levels below 0.3 are filtered out.

3.5. Nuclei detection and segmentation

Each color normalization or deconvolution algorithm produces a normalized image and separate hematoxylin and eosin image. If the hematoxylin image is presented in RGB space, then the red channel or the mean value of red and green channels is selected for further processing, denoted as H channel. As the estimation of candidate threshold proposals is performed on the normalized image, this procedure can be oversensitive to the images with large background (bright) areas, an extra measure is taken to discard such erroneous candidates. For this, the median intensity of the normalized image converted to grayscale is calculated, which will function as the minimum acceptable threshold level. Once the candidate proposal estimation on H channel is completed, the H channel is subjected to Gaussian filter (\( \sigma = 1 \)) to remove pixel noise and to smoothen nuclei borders, which improves the segmentation results.

The entire nuclei segmentation procedure is then repeated till the best threshold candidate is found, on which the final segmentation will be performed and refined by area-based correction procedure. The workflow is shown in Fig. 4.

3.5.1. Local adaptive thresholding with candidate threshold proposals

The automatic threshold selection is adapted from the Multiscale Adaptive Nuclei Analysis (MANA) algorithm [14]. MANA utilizes object-based detection with area-based correction to perform the initial nuclei segmentation, which subsequently separates tightly clustered nuclei into individual ones evaluating solidity of all detected objects. For the initial threshold proposals, the progressive weighted mean, \( PWM_{\text{CURVE}} \), is computed from the grayscale histogram of the source image, as shown in Eq. 10, adapted from [14]:

\[ PWM_{\text{CURVE}} = \frac{\sum_{i=0}^{L}s_iw_i \times x_i}{\sum_{i=0}^{L}s_iw_i} \]

where \( w_i \) represents the counts in each histogram bin and \( x_i \) the respective bin index for all bins (0 ≤ P ≤ 255). The \( PWM_{\text{CURVE}} \) represents the color distribution in the image.

It is then fitted with a 15th order polynomial, from which inflection points can be detected. The inflection point is the point on the smooth curve, where the curvature changes the sign. It is derived as a root of the second derivative of the function. The obtained values are applied as candidate thresholds. However, unlike in the original MANA, where the global Otsu’s threshold is used, in our work, the candidates are set as sensitivity levels for the local adaptive thresholding. The hematoxylin channel image (or if not available, then the source image) is transferred to grayscale level and to the [0,1] intensity range. All candidates above 0.5 are chosen for nuclei segmentation with the additional criteria that the median intensity of the input image is used as the minimum
3.5.2. Morphological operations and marker-controlled watershed segmentation

The thresholding produces a binary image with white foreground, as target segmentation structures, and black background. The set of morphological operations is applied to reduce over-segmentation, where the size of structuring elements depends on the input slide resolution. First, the filling is performed, which is a flood-fill operation, removing holes in the objects. It is followed by the opening operation with ‘disk’ structuring element of size 3 to smoothen object borders. The ‘area opening’ operation is then executed to remove all elements below the minimum nuclei area $\chi_{\text{min}}$. The nucleus area depends on its type and the image resolution, but for cancerous tissue, the size may vary greatly. This can be addressed by multi-scale threshold as defined in [39]. In our study, the experimentally derived area of 150 pixels is used as sufficient approximation. The marker-controlled watershed is utilized. The extended-minima transform (the regional minima of the H-minima transform) is applied, imposing only regional minima on binary mask, which is fed to the watershed transform. As an additional measure, the ‘area opening’ operation is applied one time more to remove small elements, that may have been introduced by watershed.

3.6. Area-based correction

Finally, the area-based correction (based on the MANA algorithm proposal) is performed to remove all segmented elements with excessively small areas, which correspond to the noise or segmentation artifacts. Although the morphological operations remove elements below $\chi_{\text{min}}$, area, due to the large variation of nuclei size in the dataset images, this additional cleanup step is necessary. For this, the mean area of all final segmented nuclei is calculated, and all structures smaller than 23% of the mean area are labeled as ‘small’ and discarded. The original algorithm can detect and mark overly large structures as well (i.e., several times larger than the mean area), but it was experimentally confirmed that excessively large but valid cancerous nuclei may be then erroneously removed, so this step was omitted.

3.7. Comparison to CNNs and MedT on TCGA dataset

Our method is compared to the accuracy of CNN models assessed by authors of MoNuSeg dataset [66], as well as to recent MedT [54]. The original CNN2 [73] represents a two-class CNN model, segmenting foreground (nuclei) and background pixels. The CNN3 model [66] is a three-class scheme, adding nucleus boundary class of pixels. Both CNNs apply nuclei detection to produce nuclear markers via a probability map generation for all pixels, showing how likely they belong to respective class. The output is subjected to thresholding and morphological operations to remove noise. Final markers are produced with distance transform with H-minima suppression. As a last step, anisotropic region growing is performed starting from markers to generate nuclei shapes and boundaries. The CNN2 and CNN3 models consist of respectively 2 and 3 convolution layers with rectified linear units (ReLU), each one followed by max pooling layer, closing in the end with two fully connected layers. The last output layer employs softmax to produce probabilities for target classes. The description of MedT model can be found from the original paper [54].

3.8. Segmentation accuracy score metrics

The pipeline performance is assessed against GT images, which provide pixel-level annotations for nuclei, by several segmentation metrics: F1-measure, Dice coefficient (DC), and Aggregated Jaccard Index (AJI) [66], as well as basic accuracy. F1-measure and DC are more robust metrics than basic pixel-level accuracy and have been widely used for assessing segmentation performance. The Dice coefficient (like closely related Jaccard Index) is utilized for pixel-level similarity comparison between GT and segmentation result:

$$
DC = 2 \times \frac{|G \cap S|}{|G| + |S|}
$$

(11)

where $G$ represents the GT image and $S$ the segmented nuclei image. The DC has its own limitations as although it penalizes segmentation errors, it does so only for true positives. In comparison to classical DC, the AJI metric (based on Jaccard Index) helps to penalize both object- and pixel-level errors and compute the ratio between aggregated intersection cardinality (number of pixels) and aggregated union cardinality for all GT and segmented objects [66]. This accounts for all falsely segmented objects, thus decreasing the score in case of under- or over-segmentation. The AJI is calculated as follows, adapted from [74]:

$$
AJI = \frac{\sum_{i=1}^{L} |G_i \cap S_i|}{\sum_{i=1}^{L} |G_i| + \sum_{i=1}^{N} |S_i| - \sum_{i=1}^{L} |G_i \cap S_i|}
$$

(12)

where $S_i$ is the segmented nucleus from set $1 < k < N$ that maximizes the Jaccard Index with GT nucleus $G_i$ from set $1 < i < L$, with the addition of falsely segmented $S_i$ nuclei, that no GT match was found for.

3.9. Nuclei density calculation

Given the segmentation results, we can calculate the number of nuclei detected in the ROI, as well as the total area of all segmented nuclei. In the present study, no extensive attempts are done to separate accurately and precisely the clustered nuclei into individual elements, as although the precise separation is beneficial for the nuclei count estimation, it does not show significant impact on the estimate of total area of all nuclei as a nuclei density feature.

For nuclei count, the ultimate erosion and dilate morphological operations are applied both on the GT image and segmented image, followed by the distance transform calculation. All resulting objects that have a minimum distance $\chi_{\text{min}}$ of 4 between adjacent objects are counted as separate nuclei. For nuclei density area $A_{\text{density}}$, a simplistic calculation is applied:

$$
A_{\text{density}} = \frac{A_{\text{nucl}}}{A_{\text{total}}}
$$

(13)

where $A_{\text{nucl}}$ is the total area of all segmented nuclei, and $A_{\text{total}}$ is the total ROI area, accounting for the parts excluded during the blur artifact removal preprocessing step.
4. Experimental results and performance study

4.1. Experimental setup

The main pipeline is implemented with MATLAB R2020b. Part of the processing is performed with Python 3.7.4 in Jupyter Notebook v6.0.1. The original WSI slide annotations are performed with Aperio ImageScope (Aperio Technologies, USA), and later enhanced with QuPath v0.2.3 [75]. QuPath software stores annotations in GeoJSON format with the export and import by Apache Groovy scripts, developed on the base of Janowczyk’s examples [76]. Main processing was performed on a general desktop with 16 GB of RAM and NVIDIA GeForce GTX 1060 3 GB. MedT model training and evaluation was performed on Nvidia Tesla V100-PCIE-16 GB with 16 GB of memory.

4.2. Blur impact on segmentation accuracy

The impact of blur presence on the segmentation accuracy can be seen in Fig. 5, as well as in Table 1, where the corresponding scores are shown for comparison. The extensive areas of the falsely segmented nuclei at the location of blur artifact (displayed in green) represent thick tissue folds that absorb transmitted light much stronger than surrounding areas. Such artifacts make them appear like the stained nuclei or dilutes nuclei boundaries, negatively impacting firstly the automatic threshold level selection, and eventually the segmentation itself. If not removed, the blur presence makes the automatic threshold selection fail in most cases, and manual sensitivity and window parameter adjustments are required to get an acceptable output.

With blurred areas removed, the segmentation accuracy is considerably higher, with AJI better by up to ~25%, Dice Coefficient by up to ~21.5% and basic accuracy by up to 11%. Additionally, the removal of large falsely segmented areas will improve the validity of extracted nuclei features.

4.3. Nuclei segmentation with color normalization

The nuclei segmentation comparison with different accuracy metrics is presented for gastric cancer dataset in Table 2 with visual comparison shown in Fig. 7. The segmentation based on color normalization [62] achieves in general the best scores, 0.854 ± 0.068 F1-measure and 0.389 ± 0.095 AJI, closely matched by CD [63] with 0.847 ± 0.094 F1-measure and 0.376 ± 0.085 AJI. The violin plots in Fig. 6 represent the distribution of F1-measure and AJI scores across all images in the dataset. The complete failures are easily observed for the ACD method, but even best performing methods show low scores on some images.

The segmentation results on the TCGA dataset [66] are presented in Table 3 with visual comparison shown in Fig. 9. They show remarkably good results for all color normalization methods with up to ~7% better AJI score, which is the most penalizing one for invalid segmentation results. The CD [63] outperforms other evaluated methods with 0.907 ± 0.044 F1-measure and 0.458 ± 0.097 AJI. Among the reference methods, the specially tailored CNN3 model outperforms the proposed framework achieving AJI score of 0.508. MedT model misses small and low contrast elements, producing a lower AJI score of 0.406 ± 0.128.

As seen in the violin plots in Fig. 8, no clear outliers are observed but some images exhibit worse results than the average score across the dataset, especially by Alsubaie et al. [65] and Khan et al. [61] methods. This can also be seen in Fig. 9, where the red boxes outline extreme nuclei segmentation failures.

Adversely, segmentation failures seemed to arise in poorly cohesive or poorly differentiated carcinoma cells. In such cases, the automatic adaptive local threshold selection failed to segment deformed or overlapping nuclei which were difficult to distinguish also in visual inspection, as shown in original results in Fig. 10, or in low contrast areas such as relatively dark cytoplasm and light nucleus. In the first example, parts of the glandular structures with clustered nuclei and blurred boundaries resulted in errors. In the second example, stromal areas containing elongated nuclei within stroma consisting of collagen strands resulted in

Table 1

<table>
<thead>
<tr>
<th>Image ID</th>
<th>Basic Segmentation</th>
<th>Segmentation w/ Blur Detector</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AJI</td>
<td>DC</td>
</tr>
<tr>
<td>Sample #1</td>
<td>0.334</td>
<td>0.658</td>
</tr>
<tr>
<td>Sample #2</td>
<td>0.335</td>
<td>0.635</td>
</tr>
<tr>
<td>Sample #3</td>
<td>0.271</td>
<td>0.557</td>
</tr>
</tbody>
</table>

Fig. 5. Comparison of blur impact on segmentation results, showing overlay for the ground truth and detected nuclei (white = matching areas as true positives, purple = false negatives, green = false positives).
Segmentation accuracy results as mean and standard deviation (STD) values using selected color normalization methods applied to the Oulu University Hospital image dataset (n = 35, 9085 annotated nuclei).

<table>
<thead>
<tr>
<th>Normalization Method</th>
<th>AJI</th>
<th>F1-Measure</th>
<th>Dice</th>
<th>Accuracy, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic, HSV space</td>
<td>0.344</td>
<td>0.807 ± 0.127</td>
<td>0.643 ± 0.130</td>
<td>85.70 ± 4.84</td>
</tr>
<tr>
<td>Macenko et al. [62]</td>
<td>0.389</td>
<td>0.854 ± 0.068</td>
<td>0.695 ± 0.058</td>
<td>86.37 ± 6.29</td>
</tr>
<tr>
<td>Hoque et al. [63]</td>
<td>0.376</td>
<td>0.847 ± 0.094</td>
<td>0.674 ± 0.081</td>
<td>87.38 ± 4.84</td>
</tr>
<tr>
<td>Alsubaie et al. [65]</td>
<td>0.330</td>
<td>0.797 ± 0.168</td>
<td>0.633 ± 0.156</td>
<td>85.13 ± 6.59</td>
</tr>
<tr>
<td>Zheng et al. [64]</td>
<td>0.345</td>
<td>0.804 ± 0.106</td>
<td>0.618 ± 0.084</td>
<td>86.03 ± 5.51</td>
</tr>
<tr>
<td>Khan et al. [61]</td>
<td>0.345</td>
<td>0.831 ± 0.068</td>
<td>0.679 ± 0.066</td>
<td>84.38 ± 6.76</td>
</tr>
</tbody>
</table>

We achieved the performance of ~0.85 F1-measure for gastric cancer dataset using color normalization [62] and > 0.90 for TCGA image dataset using CD [63] with marker-controlled watershed transform for nuclei segmentation without extensive ML or DL processing. It is shown that the addition of normalization techniques plays a significant role, with the increase in F1-measure by up to 5.8% and 7.1% for the respective image datasets above the basic processing without any normalization. The performance is comparable to the state-of-the-art methods and although it is surpassed by some most recent non-DL-based works that have been published in recent years [40,77], it is seen that the higher accuracy can be achieved only with the addition of various adaptive multi-level or multi-scale thresholding and nuclei borderline correction algorithms that perform image analysis iteratively and craftily join all outputs into the common result. It is seen that nuclei segmentation pipelines can reach higher accuracy only if using advanced DL networks.

The DL-based techniques are taken in use widely as perceived in the latest trends and systematic reviews on the application of AI-assisted methods in cancer diagnosis [78–81]. As DL methods are still hindered by the lack of sufficient labeling for disease image datasets (especially for uncommon cancers), semi-supervised or weakly supervised methods can be utilized [82,83]. The “black-boxness” of DL models has been also possible to overcome in recent years with the advances in Explainable AI [84], where heatmaps, saliency maps, occlusion maps and other similar techniques can be applied to visualize the area that DL model derives its results from. Furthermore, DL model training is still utilizing comparatively small image patches. Thus, going from small images of just a few million pixels in total to WSI scale of hundreds of millions to billions of pixels is still seen as a challenge. Even if achieving better performance eventually, presently their usage may be still impractical in most cases.

It was observed that the developed nuclei segmentation framework
performs much better on the public dataset with all color normalization techniques, even if it is a multi-tissue dataset with different cancer types. The method based on Hoque et al. [63] has 6% better overall F1-measure score (0.907 vs 0.847), and 8.2% higher overall AJI (0.458 vs. 0.376). It is concluded that it is accounted to the fact that as the ROIs in the gastric cancer dataset have been selected purposefully to represent most challenging examples from the original hospital dataset, their cancerous tissue structure complexity is much more advanced and diverse. On top of that, the slide quality is worse and impaired by compression artifacts. Both these aspects are challenging for accurate nuclei detection and segmentation. Additionally, if the GT annotations are suboptimal or even missing some nuclei entirely, the resulting accuracy scores are degraded.

As the processing is impacted by input image quality, the WSI images used for digital pathological image analysis are required to be stored in a lossless or lossy, e.g., JPEG2000, format with a quality factor of at least 80 [85,86]. The SVS files in our gastric cancer dataset contain images in the basic JPEG2 with the compression quality of 50 or 60 only. Such quality level is used to reduce WSI size (and, hence, required storage space) and in most cases it is sufficient for visual analysis by an expert pathologist [86]. However, it may hinder an automated segmentation pipeline and become an obstacle in achieving satisfactory accuracy. Further study can be performed on the existing image dataset to see how compression level affects segmentation results.

The accurate background removal helps to improve both color normalization and nuclei segmentation. As the candidate threshold proposal estimation is performed on the normalized image, this procedure can be oversensitive to the images with large background (bright) areas. Macenko et al. [62] algorithm shows great performance on both image datasets. It uses a fixed threshold level for the transparent pixel removal in OD space controlled via \( \beta \) parameter. Its value can be derived experimentally but does not generalize well, as even in a single image dataset the variation of background intensity level may vary between different slides. The performance can be improved by the addition of global Otsu’s thresholding individually for each image, which for computation cost savings can be derived from lower resolution layers of WSI [87].

The addition of the automatic sensitivity selection for local adaptive thresholding inspired by MANA method [14] with the automated candidate threshold proposals have shown reliable results. Although manual threshold parameter tweaking on each individual image may allow to achieve better segmentation accuracy, the automated

---

**Fig. 7.** Comparison of various nuclei segmentation results using selected color normalization methods applied to the selected input images from the hospital image dataset (white = matching areas as true positives, purple = false negatives, green = false positives). The red boxes outline images with obvious nuclei segmentation failures.
thresholding generalizes well without need for manual adjustments. Alternatively, a simple toolbox can offer a set of control options to allow its users to adjust or tweak automatic threshold selection on the need basis. In further studies, the background removal can be improved with more advanced detection mechanisms, which can help to remove adipose, gland lumen and other similar areas, e.g., by a series of Gabor kernels as proposed in literature [77]. The local adaptive thresholding is also affected by the selection of proper window size, and it is necessary to apply further improvements in this area with the application of adaptive morphological filter, where the size of target objects is adjusted iteratively over selected number of iterations [40].

The accurate and robust separation of clustered blobs of adjacent or overlapping nuclei has not been part of this work, but it can certainly affect accuracy results as well. Therefore, in further studies, it is beneficial to improve this area. This is especially critical for the accuracy of extracted nuclei features, where morphological and spatial features depend on the precision of segmentation for individual nuclei. Also, the number of 60 extracted features can be further extended with the addition of other color spaces as has been proposed in various works [57].

The processing time needs to be considered as well. The initial assessment of the large WSI batch (n = 583) has shown that the processing time of a single slide with just 4 ROIs can take up to 30–50 s at the finest resolution level (on the mediocre computing environment described in Section 4.1). The usage of lower levels can decrease processing time of a single slide with just 4 ROIs can take up to 30 s at 1200 pixel size) and depending on the color normalization method, the entire processing of a single ROI takes from 20 to 80 s, with the method by Khan et al. [61] with SVD and color-based classifier being the slowest one due to its higher computational complexity. Further optimizations are possible in the existing scripting pipeline to reduce running times.

### 6. Conclusion

The evaluation of clinical and academic research on digital pathological image analysis shows that despite great advanced in technology and AI-assisted methods for cancer diagnosis, the wide application of such automated pipelines in clinical care is still sparse, especially in oncology. Although AI-assisted methods are envisioned to improve cancer diagnosis, further extensive research is necessary both in the identification of proper features for ML-based methods and improved openness and interpretability of the predictions derived via DL-based methods. The study of the present state-of-the-art research results and the evaluation of nuclei detection and segmentation procedures, as well as different techniques that are introduced to improve segmentation results, show that significant improvements in this field are still required to achieve a satisfactory level of certainty and accuracy for clinical usage.

Presently, the achieved accuracy for nuclei segmentation by the developed pipeline under this work is promising. The addition of the automated artifact detection and removal, adaptive thresholding and color normalization preprocessing steps allow to improve the framework performance greatly. Accuracy could potentially be enhanced further by utilizing advanced ML or DL-based techniques. In future work, we plan to employ the promising potential of visual transformer models. Given the sparse availability of annotated datasets for gastrointestinal cancers

---

**Table 3**

Segmentation accuracy results as mean and standard deviation (STD) values using selected color normalization methods applied to the public TCGA image dataset (n = 14, 7000 annotated nuclei).

<table>
<thead>
<tr>
<th>Normalization Method</th>
<th>AJI</th>
<th>F1-Measure</th>
<th>Dice</th>
<th>Accuracy, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic, HSV space</td>
<td>0.384 ± 0.094</td>
<td>0.847 ± 0.086</td>
<td>0.650 ± 0.116</td>
<td>84.33 ± 4.64</td>
</tr>
<tr>
<td>Macenko et al. [62]</td>
<td>0.434 ± 0.085</td>
<td>0.896 ± 0.053</td>
<td>0.727 ± 0.070</td>
<td>86.70 ± 4.30</td>
</tr>
<tr>
<td>Hoque et al. [63]</td>
<td>0.458 ± 0.097</td>
<td>0.907 ± 0.044</td>
<td>0.739 ± 0.069</td>
<td>88.55 ± 3.67</td>
</tr>
<tr>
<td>Alsubaie et al. [65]</td>
<td>0.345 ± 0.098</td>
<td>0.837 ± 0.068</td>
<td>0.623 ± 0.118</td>
<td>82.47 ± 7.59</td>
</tr>
<tr>
<td>Zheng et al. [64]</td>
<td>0.445 ± 0.099</td>
<td>0.888 ± 0.056</td>
<td>0.711 ± 0.083</td>
<td>87.58 ± 4.71</td>
</tr>
<tr>
<td>Khan et al. [61]</td>
<td>0.387 ± 0.132</td>
<td>0.875 ± 0.087</td>
<td>0.704 ± 0.081</td>
<td>84.33 ± 7.82</td>
</tr>
<tr>
<td>Reference Methods (Kumar et al. [66] and Valanarasu et al. [54])</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fiji</td>
<td>0.273</td>
<td>0.665</td>
<td>0.649</td>
<td>-</td>
</tr>
<tr>
<td>CNN2</td>
<td>0.348</td>
<td>0.754</td>
<td>0.693</td>
<td>-</td>
</tr>
<tr>
<td>CNN3</td>
<td>0.508</td>
<td>0.827</td>
<td>0.762</td>
<td>-</td>
</tr>
<tr>
<td>MedT [54]</td>
<td>0.406 ± 0.128</td>
<td>0.932 ± 0.029</td>
<td>0.756 ± 0.033</td>
<td>89.90 ± 3.88</td>
</tr>
</tbody>
</table>

---

**Fig. 8.** Violin plots for F1-measure and the Aggregated Jaccard Index representing the median, the probability density, and the extreme point distribution, also depicting the outliers for the evaluated state-of-the-art methods on the public TCGA image dataset (n = 14, 7000 annotated nuclei).
that hinders training process for DL models, self-supervised or weakly supervised methods can be utilized removing the necessity for manual nuclei and tissue annotations.

Thus, the comparison of the achieved results to manual nuclei detection, segmentation, and morphological and spatial feature extraction tasks that a pathologist would need to perform for many hours to provide comparable results across the same batch of images shows the exciting potential the automation of rudimentary repetitive tasks can
Ethical approval

The Oulu University Hospital Ethics Committee approved the study (15.2.2016 §51) and the Finnish National Authority for Medicolegal Affairs (VALVIRA) waived the need for informed consent.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research has been financially supported by Academy of Finland 6 Genesis Flagship (Grant 318927) and Academy of Finland Identifying trajectories of healthy aging via integration of birth cohorts and biobank data (Grant 309112).

References
