IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 July 2023, accepted 11 July 2023, date of publication 19 July 2023, date of current version 3 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3296848

APPLIED RESEARCH

Multi-Configuration Analysis of DenseNet Architecture for Whole Slide Image Scoring of ER-IHC

WAN SITI HALIMATUL MUNIRAH WAN AHMAD[1], MOHAMMAD FAIZAL AHMAD FAUZI[1], (Senior Member, IEEE), MD JAHID HASAN[1], ZAKA UR REHMAN[1], JENNY TUNG HIONG LEE[2], SEE YEE KHOR[3], LAI-MENG LOOI[4], FAZLY SALLEH ABAS[5], AFZAN ADAM[6], ELAINE WAN LING CHAN[7], AND SEI-ICHIRO KAMATA[8], (Senior Member, IEEE)

[1]Faculty of Engineering, Multimedia University, Cyberjaya 63100, Malaysia
[2]Department of Pathology, Sarawak General Hospital, Kuching, Sarawak 93586, Malaysia
[3]Hospital Seberang Jaya, Seberang Jaya, Penang 13700, Malaysia
[4]Department of Pathology, University Malaya Medical Center, Kuala Lumpur 59100, Malaysia
[5]Faculty of Engineering and Technology, Multimedia University, Ayer Keroh, Malacca 75450, Malaysia
[6]Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
[7]Fusionex AI Laboratory, International Medical University, Kuala Lumpur 57000, Malaysia
[8]Kamata Laboratory, Waseda University, Kitakyushu-shi 808-0135, Japan

Corresponding author: Mohammad Faizal Ahmad Fauzi (faizal1@mmu.edu.my)

This work was supported by the Ministry of Higher Education (MOHE) Malaysia through the Research Excellence Consortium, Artificial Intelligence for Digital Pathology (AI4DP), under Grant KKP-2020.

ABSTRACT Nuclei classification is a mandatory process to obtain scoring information for whole slide images (WSIs). In immunohistochemistry (IHC) staining specifically for estrogen receptor (ER) biomarker, an Allred score based on the proportion and intensity of cancer nuclear staining is widely used in histopathology practice to predict response to hormonal treatment. This manually exhaustive process can be accelerated with the help of computational intelligence. In this article, we present a thorough analysis of 37 WSIs of breast cancer cases with over 2.8 million segmented nuclei. ER-stained nuclei were classified into negative, weak, moderate and strong intensities using DenseNet deep learning architecture, contributing to Allred scoring. Seven different models and configurations were exhaustively analysed in six tests to obtain the scoring reaching the best concordance of 56.8% and 81.1% with the pathologist's manual score and suggested hormonal treatment. We also discussed in detail the causes that lead to the non-concordances. This study follows the pathologists' workflow in obtaining the Allred score but is fully automated. It provides a basis for the development of more complex deep learning models, particularly for nuclei classification and achieving accurate scoring of ER-IHC stained WSIs.

INDEX TERMS DenseNet, ER-IHC, nuclei classification, PyTorch, TensorFlow, whole slide image.

I. INTRODUCTION

Digital pathology (DP) is a technology that allows pathological information created from digitalized images to be accessed, handled, and interpreted. Using optical scanners, traditional histopathology sections mounted on glass slides can be transformed into digitized histopathology images that can be viewed on a computer monitor [1]. Today whole slide scanners create images that replicate glass slides in high resolution. This whole slide scanned image is remotely consulted. It saves time, costs, and the physical transportation of slides. Hence, DP combines pathology and computers to replace the traditional diagnosis based on a microscope. This DP makes the sharing and annotating slides much simpler

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

and offers new opportunities for e-learning in health science applications [1].

DP can be one of the gears for translational medicine where interventions of new strategies using computer science to the clinical results can be implemented to benefit the patient [2]. It also helps in closing the gap between two domains (medical and computer science) to help clinicians and patients to make more informed choices and point-of-care decisions [3]. Artificial intelligence (AI) models can be trained to understand cancer nuclear using whole slide images (WSIs) with initial input knowledge from practising pathologists and clinicians, to fill the gap between basic sciences and clinical sciences. DP allows promising techniques to manipulate digital images in novel ways for diagnoses, second opinions, telepathology, quality assurance, archiving and sharing and many other uses [4]. General standards related to implementations, pathologist experiences, reliability, as well as approval of taking over routine human evaluations in diagnostics have been discussed [5], [6], [7]. Specific standards in different clinical settings have also been considered such as in oral pathology [8] and nephropathology [9]. Technical aspects of DP have been reviewed comprehensively by García-Rojo in 2016 [10], taking account of the international clinical guidelines provided by several well-known organizations. Review on particular technical features related to WSI such as the interpretation of color by the human visual system and its relations with WSI representation of pathology [11], display characteristics and its impact on diagnostic performance [12] and a white paper on tissue image analysis, software solutions and analysis strategies have been explained in detail [13].

Analysing the WSI is incredibly challenging due to its high resolution, high dimensionality, and variability in slide stain representation. These were addressed in several studies, discussing computer-aided pathologic diagnosis [14], predictive modelling object detection and tissue classification [15], and state-of-the-art nucleus and cell segmentation for different types of microscopic images [16]. Various aspects of AI tools in contributing to DP as a whole were studied [1], [5], [17], [18], as well as interpreting WSI at the cellular or nuclei level [13], [19]. A roadmap for DP in developing new AI tools and components for clinical use has been presented by NCRI (National Cancer Research Institute, UK) Cellular Molecular Pathology Initiative (CM-Path) with joint forces with the British In Vitro Diagnostics Association (BIVDA) [20]. This roadmap highlighted a few practical applications including immunohistochemistry (IHC) biomarker detection and scoring, disease quantification, morphometrics, tumor detection and cancer grading, and rare event screening (such as emphasizing tumor sample location).

The traditional way of diagnosing cancer is that pathologists, using the microscope, will study the histopathology images of the cancer biopsy taken from patients, and from their personal experience of cancer images, they will reach a decision on the presence or type of cancer they have visualised. A crucial step is the evaluation of cell patterns in the biopsy tissue. Computer-aided automatic detection methods provide assistance in this step. However, there are specific challenges. When histological sections of the biopsy are stained to allow visualization of cells, uneven distribution of stains in the tissue will induce noise - this is the first challenge. So, pre-processing steps are required to eliminate such noise. Nucleus or cell segmentation is the second challenge that has to be solved, which completely depends on the properties of the cellular details (e.g. nuclei) in the image [21]. Image segmentation is the procedure for extracting the region of interest through an automatic or semi-automatic process.

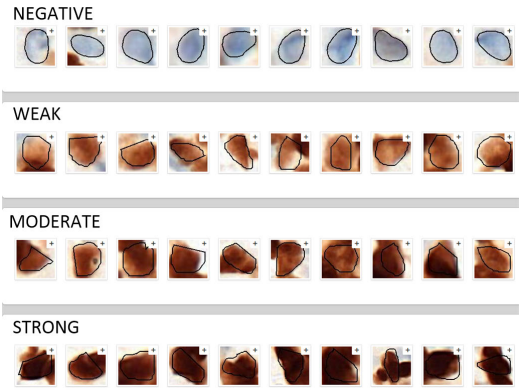

FIGURE 1. Example of nuclei classes from ground truth 22k dataset: negative, positive-weak, positive-moderate and positive-strong. The nuclei classes are categorized according to intensity.

DP can also help to increase the productivity of the surgical pathologist by automating time-consuming tasks. Some such tasks which are current aspects of breast cancer evaluation include mitosis counting, lymph node scanning for metastatic deposits and immunohistochemical scoring for specific protein expressions [22]. IHC scoring for estrogen receptor (ER) protein expression is an important means whereby the pathologist categorises cancer to predict response to endocrine therapy. The Allred score is an example of such a predictive evaluation and is based on deriving the ER status of breast cancer by combining the percentage of ER-positive cells (proportion score) and the ER staining intensity (intensity score) expressed by the tumor cell nuclei, creating a range from 0 to 8 [23]. Patients with Allred scores ≥ 3 respond to adjuvant endocrine therapy and statistically has better disease prognosis than those with Allred scores < 3 [24]. In the scoring methodology, the negative-staining nuclei are assigned an intensity score of 0, and the intensity of positive-staining nuclei is divided into three grades: weak, moderate and strong, depicted as scores 1, 2 and 3 respectively. Examples of nuclei for these grades are shown in Fig. 1. The nuclei classification into different intensity classes can be automated in DP. The aim of our work is to find the best deep learning (DL) pre-trained model for the classification of nuclei in ER-IHC-stained histopathology images of breast cancer.

II. RELATED WORKS

DL can deal with the very complex patterns of histological images. It can highlight the underlying features which cannot be identified by human eyes, and uncover unrecognized features to assist diagnosis in digital pathological images. DL stands out on accuracy, computational efficiency and generalizability in analyzing DP images, specifically on segmentation (e.g. tumor region identification), detection (e.g. metastasis detection) and classification (e.g. patient prognosis). Two important components in the DL framework are the pre-processing and post-processing stages; where the former is to optimally prepare the input, and the latter is to improve the results of the network output.

In pre-processing of DL models, three main tasks for DP images have been identified, namely tissue and artefact detection, stain normalization and patch selection. Recent work by [25] on tissue and artefact detection is based on changing the color space and adaptive thresholding, which can process WSI quicker and increase the DL performances. Work on stain normalizations used various methods ranging from global color normalization to color transfer using generative adversarial networks (GANs). In GANs, the model can learn from color distribution and histopathological patterns, thus stable performances can be achieved especially when the stains come from different sources. For patch selection, available methods used thresholding, color deconvolution and active contour model. This task can increase the overall accuracy of DL performances because patches contain significant information for particular problems.

Post-processing in DL is important to achieve the ultimate computer vision tasks: classification, detection, and segmentation. The classification task is used to predict class labels, using patch aggregation methods such as adopting simple max voting procedures or complex models like random forest and nearest-neighbour classifiers. Patch aggregation can increase DL performance, teaching the model to be more robust to low-confidence predictions and single-patch misclassifications. For detection, it was done by identifying the centroid location or bounding box of the object. Common tasks in pathological images are the detection of lymphocytes and mitosis. The main post-processing method for detection is using a non-maxima suppression algorithm, which removes the overlapping bounding boxes but maintains a high level of sensitivity. In DP images, pixel-level segmentation was commonly carried out for segmenting nuclei and tubules/glands. These were carried out using either one of the strategies: two-class or three-class pipelines. For binary segmentation and traditional morphological operations or watershed transform, a two-class pipeline is used. Recent methods using DL were using a three-class pipeline, where it can simultaneously estimate the background, the border and the inside of the object of interest.

From the literature study, CNN has achieved state-of-the-art object recognition performance in various DP applications [26], [27]. Authors [28] have adopted a deep CNN to separate abnormal from normal cervical cells in Pap-stained and hematoxylin and eosin (H&E) stained images. They proposed a pre-trained feature extraction ConvNet model on the ImageNet dataset, and data pre-processing on the cervical cell dataset. Transfer learning is also applied, whereby the pre-trained network parameters are used to initialize a new ConvNet. This ConvNet is then fine-tuned on the pre-processed training samples. Others [29] have trained a CNN to classify cells in fluorescence microscopy images, and [30] have combined a CNN with a deep autoencoder for individual cell classification. For Ki67 stains, [31] used pixel-to-pixel learning for single-stage nucleus recognition. The fully convolutional network (FCN) model takes advantage of weak labels, i.e. ROI region annotation, to assist individual nucleus identification. This auxiliary task boosts nucleus identification by encouraging the network to learn more general representations. More importantly, it can reduce human effort for fine-grained nucleus annotation, which is much more expensive. Most of the above methods only use H-channel images in the network. So they still introduce noise, and discarding E channel images means losing some structural information. The current effort in DP is on examining large pathology WSI datasets and applying AI or DL approaches to identify novel prognostic factors including tumor nesting, nuclear features, tumor cell density and stromal cell features.

ER-IHC stain is specific to breast cancer, and the study is still very limited even though it is the leading cause of cancer with high mortality rates among women; hence it is crucial to conduct more deep-learning-related research to pave the way for computer-aided pathological image analysis in breast cancer. There is no similar work on ER-IHC from other authors in the literature, and no standard or public dataset is available for ER-IHC baseline or benchmarking. We recently published the initial work done on the Allred scoring for ER-IHC in [32], introducing the proof of concept for hormone receptor status in WSIs. In this work, a different approach was applied where a well-known StarDIST object detection method [33] is utilized for nuclei segmentation, obtained using Cytomine platform [34]. StarDIST was trained using H&E stained images but seemed to produce decent results with other stains. The nuclei were then classified into four classes using several DL approaches based on the DenseNet model with the aim of getting the highest scoring concordance with the pathologists at the whole slide level. DenseNet is chosen for its outstanding performance in classifying ER-IHC nuclei in our earlier work [35]. The main contribution of this paper is listed below:

1) Exhaustive ground truth (GT) generation for ER-IHC on the whole slide level and the cellular level. Our collaborating pathologists helped on identifying the whole slide regions for scoring; validating the segmented and classified nuclei to further train our DL models; and providing concordance validation on the 3k nuclei dataset.
2) Modified trimmed and lightweight DenseNet configuration with only 21 dense layers with a reduction of 97.5% model size, faster computation and

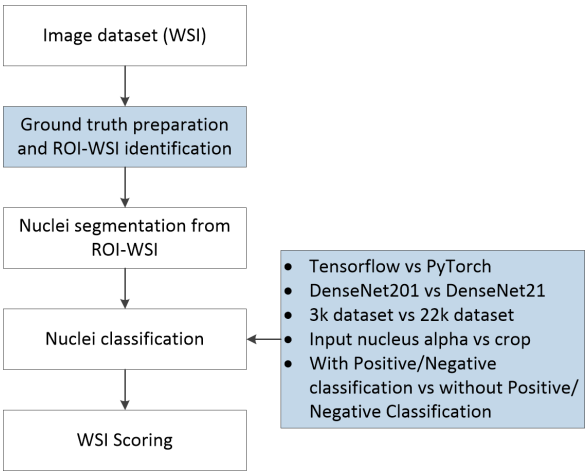

FIGURE 2. General process flow of the methodology. The focus of this paper is highlighted in the blue boxes.

better concordance with the pathologists' manual interpretation.

3) Modified positive-negative nuclei classification to first filter out negative nuclei of ER-IHC for better separation of the two classes before further classifying them into weak, moderate and strong classes using the DL model.
4) Comprehensive comparative evaluation on the DL models where seven different configurations are experimented in six tests, to find the highest concordance with the pathologists manual evaluation. We dive into details on why the models misclassified some of the nuclei, which are elaborated in separate sections for each test in Section IV.
5) All underlying issues related to the proposed nuclei classification for ER-IHC are highlighted in a section before concluding this article, together with some potential improvements.

III. METHODOLOGY

There are five general steps in the methodology of this article, as illustrated briefly in Fig. 2. Main contributions are highlighted in the blue boxes: the GT preparation and different setup of DL environment for the nuclei classification process.

A. IMAGE DATASET

Our image database consists of a total of 44 ER-IHC stained WSIs of invasive breast carcinoma provided by our collaborating hospital, Universiti Malaya Medical Center. The WSIs were scanned using 3DHistech Pannoramic DESK at 20× magnification with an approximate dimension of 80,000 pixels width and 200,000 pixels height per WSI. These 44 WSIs comprise an initial cohort of 40 which had Allred scoring by the collaborating pathologists with a breakdown of 17 ER-negative and 23 ER-positive. Of these 40 WSIs, 37 had region annotation by the pathologists (as in Fig. 3). The remaining 4 of the 44 WSIs (subsequently added) were not annotated and did not have manual Allred scoring. The WSI dataset used in this research is available at the IEEE Dataport https://dx.doi.org/10.21227/9gbq-gz50. The usage is allowed only for scientific research and must be cited with ethical attribution to this article.

B. GT PREPARATION AND ROI-WSI IDENTIFICATION

For GT preparation, there are three separate processes: one for Allred scoring of WSIs, and another two for nuclei classification of negative, weak, moderate and strong (NWMS) classes. The details for these steps are explained below.

1) GT FOR ALLRED SCORING

1) Only 37 WSIs are evaluated
2) Our pathologists identified useful large ROIs for each WSI (example in Fig. 3) and manually delineated the regions. These regions will be referred to as ROI-WSI.
3) Single scoring will be given per WSI based on the ROIs. The GT has 37 Allred score for 37 WSIs.

2) GT FOR NUCLEI CLASSIFICATION

1) 37 regions
   a) The pathologists identified one small patch per WSI, of size around 500 by 500 pixels. 37 patches are identified, one each from the 37 WSIs.
   b) The patches underwent nuclei segmentation using Stardist object detection method [33] in Cytomine platform [34]. Stardist was trained using H&E stained images, but seemed to produce decent results with other stains, and can be applied for IHC stains.
   c) A total of 3333 nuclei were segmented and manually classified into the four classes by two junior pathologists for concordance comparison. After the classification, there are 233 nuclei with class disagreement, and their class was then determined by a senior pathologist (with >35 years of experience in breast cancer pathology diagnostics). As a final result, 2428 nuclei are classified as negative, 135 are positive-weak, 367 are positive-moderate, 249 are positive-strong and 154 are not classified due to incomplete or inaccurate segmentation, yielding 3179 validated nuclei dataset. This dataset will be called as 3k dataset for the rest of this article.
2) 220 regions
   a) The pathologists identified five small patches per WSI, of size around 500 by 500 pixels. 220 patches are identified from the 44 WSIs.
   b) These patches will undergo similar processes as the 37 patches, segmentation of nuclei using Stardist in Cytomine for all 220 patches.
   c) The segmented nuclei are pre-classified into four different classes (negative, weak, moderate and strong) using DenseNet-201, which was

preliminarily trained with the GT nuclei from the 3k dataset.

d) Two pathologists were involved in validating the detection and segmentation of the nuclei, by accepting the correctly segmented nuclei, correcting the under- or over-segmented nuclei, and manually adding the delineation of missing nuclei. They also validated the pre-classified nuclei, by accepting the correctly classified nuclei, correcting the wrongly classified nuclei, and adding the class of the missing nuclei. The validation process of both segmentation and classification was done simultaneously, patch by patch. It was done by two pathologists, with 110 patches for each of them. The resulting number of validated nuclei was 22431 (16209 negative, 1391 weak, 3078 moderate and 1753 strong). This dataset will be called as 22k dataset for the remainder of this article.

The overall process flow of the GT generation is shown in Fig. 4 with three main components: WSI database, scoring GT generation and class GT generation. Since the initial scoring and classification were done based on the 37 WSIs, the related process is highlighted using blue texts, boxes and arrows for distinction. It is noticeable that both 3k and 22k datasets are imbalanced, and can be addressed by using either data augmentation, generative adversarial network, or hyperparameter tuning. In this experiment, we used data augmentation from Albumentations [36] image transformations, based on vertical and horizontal flip, shifting hue, saturation and value limits, rotation and random size crop.

C. NUCLEI SEGMENTATION FROM ROI-WSI

The identified large region from the WSIs (ROI-WSI) for the GT Allred scoring was extracted from the WSI for cellular level analysis, i.e.: nuclei detection and segmentation. There were a total of 359 ROI-WSI extracted from the WSIs, from one up to 26 large regions per WSI. The size of ROI-WSI was very big, ranging from 6,227 to 28 million micron^2 (480 by 642 to 10,000 by 8,543 pixels) containing hundreds of thousands of nuclei. Each WSI had more than one ROI-WSI, with the total area listed in Table 1 together with the total of segmented nuclei in each image. The process had taken place in September 2021, and was also done using the Stardist object detection method [33] in Cytomine platform [34] (S_CellDetect_Stardist_HE_ROI v1.0.4) based on ''ROI-WSI'' term. This particular version of the Stardist detection algorithm used during the experiment had the limitation of processing only up to 5,000 by 5,000 pixels, hence the large ROI-WSI is divided into smaller regions with a maximum of 2,048 pixels sides, as illustrated in Fig. 5. The code for this operation can be found in https://github.com/mizjaggy18/S_ROI_splitpoly. Processing smaller regions at a time is also computationally efficient because only a smaller space will be allocated rather than having to allocate one large space or memory for one large region. The total number of segmented nuclei was 2,878,290, with an average time taken of 0.06s per nucleus without GPU setting. The total time taken was 172,697.4s (47.9715 hours). The computational time can be reduced with GPU runtime but the setup had not taken place during this experiment.

TABLE 1. List of images with their respected total ROI-WSI area in micron^2, number of ROI-2048 blocks and total number of segmented nuclei. This table is sorted in ascending order according to the total ROI-WSI area.

Image ID	Total ROI-WSI area (micron^2)	ROI-2048 blocks	Total nuclei in ROI-WSIs
4301099	3538285.39	36	8979
5155522	3885449.28	73	20199
6278990	5482445.22	57	23629
4305373	5720592.12	41	38079
5060537	5818141.78	55	55114
4305361	6209926.72	53	44608
4305247	7225306.23	46	17793
4305435	10819140.66	103	35146
4305385	11534372.10	102	34502
4305343	14135724.12	96	70742
4305273	14428585.03	91	21109
4305317	15288197.06	140	75331
4305453	15414392.55	139	96661
4305267	15997965.93	123	38950
4305349	16328058.92	128	40658
4305305	16544798.53	133	45306
4305337	17454623.69	174	109716
4305403	17598983.11	129	78509
4305299	18000125.74	109	65235
4305293	19836296.89	132	57307
4305255	21903409.21	144	51632
4305447	22108125.86	171	117978
4305397	22924183.54	168	59930
4305421	24257008.53	156	117108
4305285	24921165.75	165	101619
4305441	25086813.28	166	44896
4305379	25327788.22	172	102846
4305459	25730449.58	172	98102
4305427	26099794.94	174	62281
4305415	26596513.08	175	85803
4305279	27059812.12	182	70404
4305391	29565956.95	208	148694
4305323	30108868.89	192	69178
4305329	41025950.92	262	163241
4305367	43047588.03	240	171153
4305465	47998679.25	288	146215
4305409	51847533.64	349	289637

D. NUCLEI CLASSIFICATION

The nuclei classification process is crucial to determine the scoring of the ER-IHC biomarker. The score expressed by the cancer informs the decision on whether hormonal treatment is suitable for the patient. The pathologists have furnished us with the GT Allred scoring per WSI, based on the identified regions. From the regions, every single nucleus is extracted and classified according to the four NWMS classes. The count of the resulting classes will determine the final intensity and proportion scores, which in combination will yield the Allred score. In order to match the pathologists' scoring and to obtain reasonable computational time, we tested several DL setups for the classification and scoring performance of breast carcinoma on ER-IHC WSIs. The approaches are

FIGURE 3. WSI with the annotated large regions by the pathologist. The regions are referred as ROI-WSI.

listed in Fig. 2 and all the models involved in the experiments are summarized in Table 2. All DL models are trained from scratch without pre-trained weight with data augmentations from Albumentations [36] using 100 epochs with batch size 16.

1) TensorFlow VS PyTorch

These two frameworks were evaluated in terms of scoring accuracy and computational time. The DL model used for the nuclei classification was DenseNet-201, based on its outstanding performance in our previous study on pancreatic cancer of pathological images [37]. We also compared 32 DL models for ER-IHC nuclei classification done in [35], where DenseNet-169 gave the best performance but with a very minimal difference from DenseNet-201. The confusion matrix of DenseNet-201 showed more balanced results than DenseNet-169, especially on the Strong and Weak classes. Hence DenseNet-201 is chosen for this exhaustive experiment. The scoring is highly dependent on the nuclei segmentation and classification process.

2) DenseNet-201 VS DenseNet-21

DenseNet architecture is known for its high training time and large model size. Hence, we trimmed the dense layers by limiting the convolutional layers with only block configurations of (2,2,2,2) yielding 21 layers altogether, instead of (6,12,48,32) for DenseNet-201. This configuration is faster because the convolutions are repeated only twice for each dense block, as opposed to 6, 12, 48 and 32 respectively for DenseNet-201. This is followed by evaluating the effect of the lighter dense layers using DenseNet-21. Reducing the model's layer will also reduce the model size by 97.5% (5MB vs 200MB), hence a reduction in parameters and time complexity as well. The impact of the lightweight model will be evaluated against the scoring.

3) 3k DATASET VS 22k DATASET

The best performing DenseNet model (201 vs 21) will be retrained using a higher number of nuclei images, the 22k dataset. The performance of the two models (the one trained using the default 3k dataset against the 22k dataset) will be evaluated. This test is important to see the effect of getting pathologists' concordance on the validation, compared to single pathologist validation.

4) INPUT NUCLEUS IMAGE (CROP VS ALPHA)

After the nuclei segmentation process in section III-C, there is an option to take the output image as crop or alpha. For the default setups, the crop-nucleus input image is used for model training and classification. The difference between these two is clearly shown in Fig. 6. The crop-nucleus will be the nucleus cropped around its bounding box, and the alpha-nucleus will only be the nucleus itself with the exact boundary and the background will be set to alpha

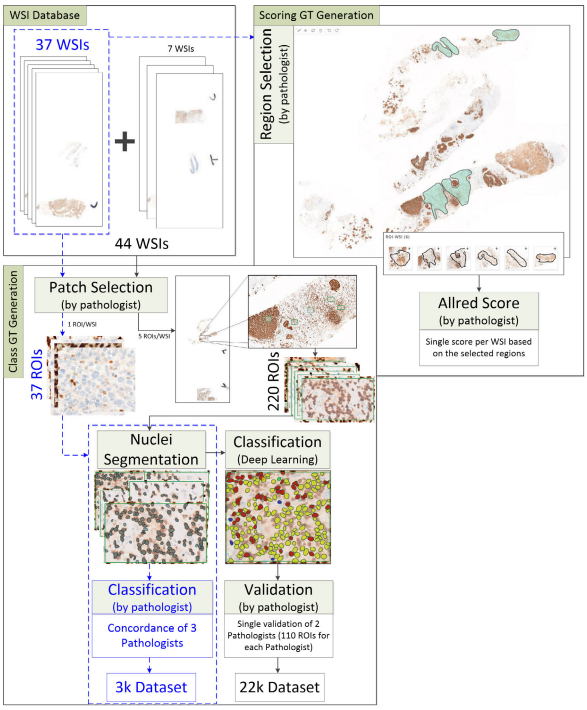

FIGURE 4. Process flow for GT generation: from WSI database to whole slide scoring and nuclei classes, provided by the collaborating pathologists.

(transparent). The alpha image will have 4 channels (RGBA), instead of only RGB. This setup was carried out on the 22k dataset, where the DL model was trained using alpha-nuclei images, and compared with a model trained using crop-nuclei images in the 3k vs 22k dataset setup.

5) PN CLASSIFICATION (WITH VS WITHOUT)

Positive-Negative (PN) classification was proposed in our previous work [38] for p53 expression. The same idea is used for ER expression, as shown in Fig. 7, which will be explained in the following subsection. In this setup, we filtered out negative nuclei using the PN classification algorithm, and only the remaining nuclei were classified using the best DenseNet model. The classes will still be NWMS, where the PN algorithm will act as a first filter, and DenseNet classification as a second filter. This approach is called a hybrid approach, and is expected to produce lesser false positive classification involving negative- and weak-class nuclei.

E. POSITIVE-NEGATIVE (PN) CLASSIFICATION

For the last setup of the nuclei classification experiment, we filtered out negative nuclei using the PN classification algorithm, before performing DenseNet classification at the positive output. Even though the PN algorithm has filtered out the negative nuclei, there were still negative nuclei remaining in the positive class, having dark and intense blue color, which cannot be filtered out with simple threshold filtering. Hence DenseNet classification will still categorize the nuclei into the four classes. The process flow is shown in Fig. 7, where

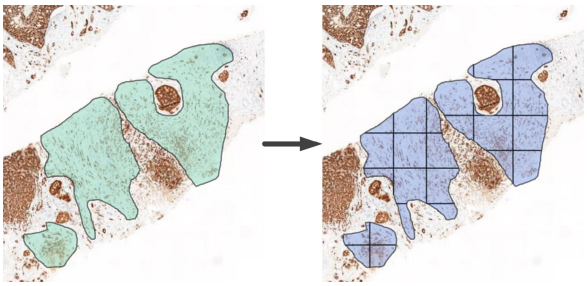

FIGURE 5. Division of large ROI-WSI into smaller regions with a maximum of 2048 pixels sides. The division will minimize the resources used by the algorithm compared to running on large ROI-WSI.

TABLE 2. Summary of the models for nuclei classification.

Model name	Model details
TF-DN201	TensorFlow DenseNet201
PT-DN201	PyTorch DenseNet201
PT-DN21	PyTorch DenseNet21
PT-22k-DN21-crop	PyTorch DenseNet21 using 22k dataset with crop input image
PT-22k-DN21-alpha	PyTorch DenseNet21 using 22k dataset with alpha input image
PN+PT DN21	PyTorch DenseNet21 hybrid with PN classification
PN+PT-22k DN21	PyTorch DenseNet21 using 22k dataset hybrid with PN classification

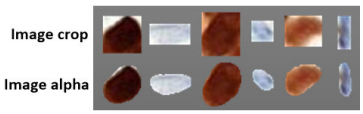

FIGURE 6. Input nucleus image: crop vs alpha. For crop image, an exact bounding box of the detected nucleus boundary is taken; while for alpha image, the exact detected nucleus boundary is used, and the outside boundary area will be transparent, using an additional alpha color channel.

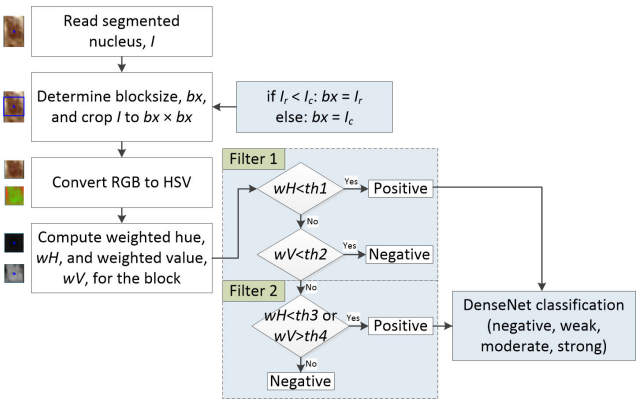

FIGURE 7. Process flow of obtaining PN classification, followed by DenseNet classification (negative, weak, moderate, strong) at the positive output. Examples of nucleus images are shown for each step.

the blue highlighted box showed the differences between this approach on ER expression compared to the p53 expression, as proposed in [38]. Since the nucleus (I) size varies, we take the smallest side of the nucleus bounding box, either

I_r or I_c, to crop it as a square block. Then, the nucleus block will be converted to HSV (Hue-Saturation-Value) color space. Weighted values of Hue and Value will be calculated using the reciprocal of Euclidean's distance, against the block's centroid; the closer the pixels are to the centroid, the higher the weighted values. IHC stain produced a brownish color nucleus in response to ER-positive expression, and blueish for ER-negative. Here, two filters with different threshold values are applied to separate the positive and negative nuclei, according to their wH and wV values.

F. WSI SCORING

The Allred scoring for ER-IHC is determined by combining the intensity score and the proportion score of the regions identified by the pathologists. From the classified nuclei, the positive portion will contribute to the scoring. The percentage of positive nuclei over the total of whole detected nuclei will determine the proportion score, and the highest class of positive nuclei (either weak, moderate or strong) will determine the intensity score. For example, a WSI with regions containing a total of 8979 nuclei (from Fig. 3 regions), and the classifications are 2113 negative, 194 weak, 6088 moderate and 584 strong, with total positive nuclei is 6866 (76.47%). By referring to the standard ER status evaluation, as listed in Table 3, the proportion score is 5, and the highest positive class is moderate, where the equivalent intensity score is 2. From here, the Allred score is the summation of both scores, which equals 7. The scores for the hormonal treatment for ER-IHC are categorized into two groups. Allred scores of 0 and 2 are considered negative and not actionable for hormonal treatment, while scores of 3 to 8 are considered positive and recommended for hormonal therapy. A score of 1 is not a possible outcome because a proportion score of 0 means there are no positive nuclei (i.e. the intensity score will be 0); and a proportion score of 1 means there are some positive nuclei and the intensity score must be at least 1.

In manual practice, the evaluation is done by the pathologists based on manual counting and estimation, and the speed is depending on their level of confidence. Expert pathologists can perform the evaluation in less than a minute per WSI. For a computer algorithm, it has to go through nuclei detection and segmentation, classification and lastly scoring, involving every single nuclei. The computational time can be very extensive depending on the region size, but computer-aided evaluation allows the pathologists to attend to more urgent diagnostic aspects of work than tedious manual counting.

G. EVALUATION METHOD

All the different approaches proposed in Section III-D will be evaluated based on the final Allred score and computational time for each WSI. The score will be compared with the manual GT score provided by the pathologists. The best approach is one with the most agreement with GT scores and is less computationally expensive. There will be six comparison experiments of the exhaustive cellular level analysis for the 37 WSIs, summarized in Table 4. For each analysis, only one setting is changed (with emphasized font), to observe the effect of each tested parameter, namely the number of dense layers, framework, dataset for model training, and input nucleus image. It is known that DL models perform differently with different applications and types of images. While the evaluation of this "grid search" method is exhaustive, the analysis takes every possible parameter to find the best model to get the highest scoring concordance with the pathologists at the whole slide level for ER-IHC stain.

TABLE 3. Allred score for Estrogen and Progesterone receptor evaluation. Allred score = Proportion score + Intensity score.

ER status (% of positive nuclei)	Proportion score	Nuclei intensity	Intensity score
0	0	None	0
<1	1	Weak	1
1 to 10	2	Moderate	2
11 to 33	3	Strong	3
34 to 66	4		
$\geq$ 67	5		

TABLE 4. Summary of proposed approaches for evaluation. The six tests are from seven different model configurations in terms of framework, dense layer, dataset and input nucleus, as listed in Table 2.

Experiment	Framework	Dense layer	Dataset	Input nucleus
Test 1	**TensorFlow vs PyTorch**	201	3k	crop
Test 2	PyTorch	**201 vs 21**	3k	crop
Test 3	PyTorch	21	**3k vs 22k**	crop
Test 4	PyTorch	21	22k	**crop vs alpha**
Test 5	**PyTorch vs PN+PyTorch**	21	3k	crop
Test 6	PN+PyTorch	21	**3k vs 22k**	crop

IV. RESULTS AND DISCUSSIONS

The ROI-WSI for each WSI underwent nuclei detection, segmentation, and classification to obtain the Allred score. Detection and segmentation tasks were done only once, and the classification experiment was repeated comprehensively with different settings for Test 1 to Test 6, with parameters as described in Table 4. The resulting Allred score for every 37 WSIs of all settings is listed in Table 5, with the GT scores manually estimated by the pathologists. The table is sorted according to the manual GT scores, from 0 to 8. Out of 37 WSIs, 17 of them are considered negative (scores 0 and 2), and the other 20 are considered positive (scores 3 to 8). The comparison of Allred score was evaluated on the exact agreement with the pathologist's scores, and also on the clinical utility of suggestion for hormonal treatment. The WSI scores for the latter were marked with an asterisk (*), and the total agreement for both comparisons is summed up at the bottom of the table.

A. TEST 1: TensorFlow VS PyTorch USING DenseNet201 (TF-DN201 VS PT-DN201)

The first test was to investigate two main DL frameworks: TensorFlow and PyTorch, using the same model,

TABLE 5. Allred score for 37 WSIs comparative to the manual GT score done by the pathologists, sorted according to the GT score.

Image No (ID)	Manual (GT)	TF-DN201	PT-DN201	PT-DN21	PT-22k-DN21-crop	PT-22k-DN21-alpha	PN+PT DN21	PN+PT-22k DN21
		Test 1	Test 1,2	Test 2,3,5	Test 3,4,6	Test 4	Test 5,6	Test 6
1 (4305349)	0	3	2*	2*	3	3	2*	3
2 (5155522)	0	2*	2*	2*	2*	2*	2*	2*
3 (4305367)	0	4	4	5	4	4	5	4
4 (4305305)	0	4	3	3	4	4	3	3
5 (5060537)	2	2*	2*	2*	2*	2*	2*	4
6 (6278990)	2	2*	2*	2*	2*	2*	2*	2*
7 (4305459)	2	3	3	3	3	3	3	3
8 (4305447)	2	3	3	3	3	3	2*	3
9 (4305441)	2	5	5	5	5	5	5	5
10 (4305409)	2	2*	2*	2*	2*	2*	2*	4
11 (4305403)	2	3	3	2*	3	3	2*	3
12 (4305385)	2	2*	2*	2*	2*	2*	2*	2*
13 (4305343)	2	3	3	3	3	3	2*	3
14 (4305337)	2	2*	2*	2*	2*	2*	2*	2*
15 (4305329)	2	2*	2*	2*	2*	2*	2*	2*
16 (4305317)	2	3	3	3	3	3	2*	3
17 (4305267)	2	4	4	4	4	4	4	4
18 (4305361)	3	3*	3*	3*	3*	3*	3*	3*
19 (4305373)	3	3*	3*	3*	3*	3*	3*	3*
20 (4305453)	3	3*	3*	3*	4*	4*	3*	3*
21 (4305465)	3	4*	4*	4*	4*	4*	4*	4*
22 (4305427)	3	4*	4*	4*	4*	4*	4*	4*
23 (4305421)	3	3*	3*	3*	3*	3*	3*	3*
24 (4305415)	3	4*	3*	3*	4*	4*	3*	4*
25 (4305397)	3	3*	3*	3*	3*	3*	3*	3*
26 (4305391)	3	3*	3*	3*	3*	3*	3*	3*
27 (4305379)	3	2	2	2	2	2	2	2
28 (4305285)	3	2	2	2	3*	3*	2	2
29 (4305279)	3	3*	3*	3*	4*	4*	3*	3*
30 (4301099)	7	7*	7*	7*	7*	7*	7*	7*
31 (4305247)	7	6*	6*	6*	6*	6*	6*	6*
32 (4305435)	7	7*	7*	7*	8*	8*	7*	8*
33 (4305299)	7	6*	6*	6*	6*	6*	6*	6*
34 (4305293)	7	6*	6*	6*	6*	6*	6*	6*
35 (4305255)	8	8*	7*	7*	8*	8*	7*	8*
36 (4305323)	8	8*	8*	8*	8*	8*	8*	8*
37 (4305273)	8	7*	6*	6*	7*	7*	6*	7*
Agreement on Allred score/ hormonal treatment*		17/25*	17/26*	18/27*	15/26*	15/26*	**21/30***	14/23*

*Allred scores of 0 and 2 are negative (i.e.: not actionable), and scores of 3 to 8 are positive (i.e.: recommended for hormonal therapy).

DenseNet201. Both scoring results show that TensorFlow and PyTorch are able to get 17 out of 37 WSIs exact same score as the GT. Even though the total agreement is the same, the individual WSI scores for both models vary from one another. We can see that there is no concordance for the first four WSIs with a GT score of 0 for all models. This is due to inaccurate segmentation (segmenting other than nuclei) and also sensitive classification by the model. The GT scores were prepared by the pathologists by manual estimation based on their experience and expertise, whereas classification by computer was according to the learnt model, nucleus per nucleus. None of the nuclei were neglected unless it was missed out during the segmentation. A single nucleus classified as positive-weak will result in a proportion score of 1 and an intensity score of 1, which translates to an Allred score of 2. Some of the examples of these segmentation and classification issues will be discussed further at the end of this section.

Other than WSIs with GT 0, the difference in score results by these two models can be seen in images 24, 35 and 37; one score concordance for each TensorFlow (image 35) and PyTorch (image 24) model. Detailed comparisons of these images are shown in Table 6 together with the subsequent tests. Concordance scores or closest to the GT are marked with an asterisk in the column ‘‘Allred Score’’. For image 24, the PyTorch model is better in differentiating negative (Class 0) and positive-weak (Class 1) nuclei by more than 6000 nuclei, with a total of +7.3% more positive nuclei in the TensorFlow model. For images 35 and 37, TensorFlow is better in differentiating positive-moderate (Class 2) and positive-strong (Class 3), where the difference is on the intensity score, with similar ER status proportion. By looking at

their agreement on hormonal treatment with GT, 25 WSIs from the TensorFlow model have the same suggestions as the GT, and 26 WSIs for PyTorch. We have to stress here that accuracy on the agreement for treatment is far more important than the exact score itself because the decision will determine whether the patient should undergo hormonal treatment or not. Since the PyTorch model had more treatment concordance, we proceeded with PyTorch for the following experiments, first on lighter-weight DenseNet.

B. TEST 2: DenseNet201 VS DenseNet21 USING PyTorch (PT-DN201 VS PT-DN21)

This experiment will compare the performance of the previous model (PT-DN201) with a lighter DenseNet, with only 21 layers instead of 201, using the same PyTorch framework. Surprisingly the lightweight model performed better with more concordance with GT on both the Allred score and hormonal treatment (18 and 27 correspondingly). For most of the WSIs, the Allred scores were exactly the same as the heavier model, except it produced better results on image 11 which concordance with the GT score. The difference can be seen in Table 6 where the positive proportion is only +0.189% for the heavier model but gives a different proportion score. Meanwhile for image 3, even though scores from the DN201 model are not a concordance, the difference with GT is lesser than scores from DN21.

Upon checking the detailed result, out of 171153 nuclei in the ROI-WSIs for image 3, DN201 classified them as 170404, 11, 2 and 736 for each NWMS class, which contribute to 0.438% positive nuclei, with proportion and intensity scores of 1 and 3 respectively. For DN21, the NWMS classifications are 165254, 751, 0 and 5148 accordingly, with 3.447% positive and 2+3 for proportion+intensity scores. While we do not have a GT breakdown for each class, since the GT Allred score for this class is 0, DN21 clearly misclassified many nuclei as positive-strong. The reason is some parts of the nuclei are having intense dark blue stains, which the model confused with the positive-strong nuclei. DN201 has extra 180 dense layers to learn more about the features, hence it is able to learn the different nuclei characteristics better.

C. TEST 3: PyTorch DenseNet21 USING 3k VS 22k DATASET (PT-DN21 VS PT-22k-DN21-CROP)

We have seen slight improvement by reducing the DenseNet layer from 201 to 21 in Test 2. In this Test 3, the DN21 model is trained using a 22k dataset, as explained in section III-B b. (ii) iv. We expect more concordance on both exact scores and hormonal treatment, but the results show otherwise. The model trained with the 22k dataset obtained 15 and 26 respectively for concordance on scoring and hormonal treatment, whereas the 3k model achieved 18 and 27 correspondingly. On closer comparison, the 3k model is better for six images, with five concordances on both score and treatment (images 11, 20, 24, 29 and 32) and one on treatment only (image 1). For the 22k model, there are three images which performed better than the 3k model, namely images 28, 35 and 37, with the first two being concordance on both score and treatment and the last one only on treatment. Detailed comparisons for these images are tabulated in Table 6. For images with low ER status, the 3k model is better because it is able to differentiate the negative and positive-weak nuclei much better than the 22k model, producing lower positive nuclei proportion. While for images with higher ER status, the main difference between the two models is their bias on the positive-moderate class (3k model) and positive-strong class (22k model).

Images 3, 4, 7, 8, 9, 13, 16, 17 and 27 remain as non-concordance issues for all tests so far. Image 28 previously was non-concordance for all other models, but the 22k model now can successfully get the concordance score. Out of 101619 nuclei in the ROI-WSIs, the positive proportion by the 22k model is 1111, 115 and 50 for weak, moderate and strong classes, while the 3k model gets 605, 32 and 26, which contribute to proportion scores of 2 (1.256%) and 1 (0.652%) for each 22k and 3k models. The difference is very small but gives different proportion scores.

D. TEST 4: PyTorch DenseNet21 USING 22k DATASET USING INPUT IMAGE CROP VS ALPHA (PT-22k-DN21-CROP VS PT-22k-DN21-ALPHA)

This test will compare if there is any difference using the crop or alpha input nucleus, as explained in Section III-D-d. The difference between the two can be clearly seen in Fig. 6, and we are expecting some improvements using the alpha input because it takes the exact nucleus area without the surrounding pixels. The classification results somehow turned out to be exactly the same for both, class per class, hence it shows that limiting the nucleus area without its surrounding pixels did not affect the model's learning process. We did not dive into details, for example comparing nucleus per nucleus, because the aim is to get the best model for WSIs scoring. Both models obtained a total of 15 concordances for the score and 26 for hormonal treatment. Since these two models gave less concordance as compared to PT-DN21 in Test 3, the next test will take the PT-DN21 model for comparison.

E. TEST 5: PyTorch DenseNet21 NON-HYBRID VS HYBRID WITH PN CLASSIFICATION (PT-DN21 VS PN+PT DN21)

This test will take a hybrid setup considering positive and negative nuclei classification using procedures explained in Section III-E with a combination of the PT-DN21 model, to be compared with the non-hybrid PT-DN21 model. This hybrid model gave the most favorable results as compared to the rest of the five models so far. The concordance for Allred score is 21 and for hormonal treatment is 30. The improvements for both concordances are +8% when measured with the non-hybrid model, which can be observed in images 8, 13 and 16. For these three images, all other models gave false positive results with a score of 3 which is positive and recommended for hormonal therapy, while the patients' GTs indicate their ER status is negative and not actionable.

Upon closer inspection in Table 6, the significant difference given by the hybrid model is in its ability in

differentiating negative and positive-weak nuclei for these three cases. Previously we observed the non-hybrid model is better in this type of classification in Test 3 when compared to the non-hybrid 22k model, and a combination with PN classification further improves it. For the non-hybrid model, total nuclei for the positive-weak class are 1313, 652 and 817 for each image 8, 13 and 16, which make the total positive percentage of 1.122%, 1.009% and 1.112%. To get an Allred score of 2, the proportion score has to be 1, where the ER status has to be less than 1% of positive nuclei. The hybrid model is able to achieve this, with the total positive-weak nuclei of 924, 592 and 601, and ER status of positive nuclei of 0.791%, 0.923% and 0.822% for each of the three cases. The small differences make significant changes to the proportion score due to the strict negative ER status range (less than 1%) which minimized the false positive prediction.

In Test 3, we listed the nine non-concordance images where all tested models failed to get the correct score or hormonal treatment. In this test, the scores of the aforementioned images (images 8, 13 and 16) have successfully achieved GT concordance but an issue remains for image 28 in addition to the other six non-concordance images (images 3, 4, 7, 9, 17 and 27). These images will be further analyzed later in this section.

F. TEST 6: PyTorch DenseNet21 HYBRID WITH PN CLASSIFICATION USING 3k VS 22k DATASET (PN+PT DN21 VS PN+PT-22k DN21)

In Test 3 and Test 4, we observed there is nothing much to promise by the 22k model, apart from the concordance for image 28. This test is to see whether or not a hybrid of PN classification is able to improve the 22k model to supersede the PN+PT DN21 model. By looking at the score concordance, it is clearly not with the 14-Allred score and 23-hormonal treatment concordances. The difference is too much that it seems unfair to compare these two hybrid models, so we deemed it fit to make a comparison with its non-hybrid model instead. The hybrid model is better for images 20 and 29 in getting the exact concordance for the Allred score, even though both have the same agreement on the hormonal treatment.

However, the opposite happened for images 5 and 10, where the hybrid of 22k model produced the worst scores, with Allred score of 4 for both images, while all other models are not having any issues getting the concordance for the exact score and hormonal treatment (Allred score of 2). When checking the detailed classification scores for each class, we found the cause of this issue, as tabulated in Table 6. For image 5 particularly, the positive classes for the non-hybrid model are 5, 1 and 5 for each WMS, and 4, 1 and 5 for the hybrid model. The algorithm decides that the highest intensity score for the non-hybrid model is weak (class 1), and strong (class 3) for the hybrid model, even though their difference is only one nuclei in class 1. We looked into the classified nuclei, as shown in Figure 8, and the positive-strong class were mixed of artefacts and misclassified nuclei. Better classification can be obtained after addressing artefact segmentation.

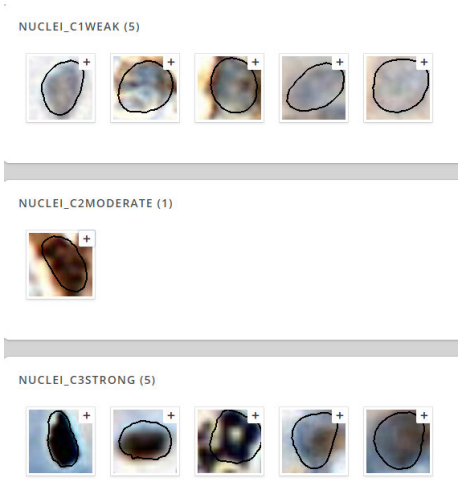

FIGURE 8. Classification results for positive classes of weak, moderate and strong nuclei for image 5 using non-hybrid PT-22k DN21. The first three cells in the positive-strong class are actually artefacts, wrongly segmented as nuclei. The remaining 2 nuclei.

For image 10, the hybrid model is clearly better in separating the negative and positive-weak nuclei with only a total of 330 positive proportions as compared to the non-hybrid model (506 positive nuclei). However, the intensity of the ER status for the hybrid model is biased towards class 3 (positive-strong) even though the number is lesser than the non-hybrid model, due to a much lesser nuclei in class 1 (positive-weak).

G. TIME COMPLEXITY

In the previous subsections, we compared the performance of each model in terms of their concordance of scores and hormonal treatment suggestions with the GT. The comparison is purely on the scoring accuracy without taking the time complexity into consideration, which will be elaborated in this section. Our system uses the 6th generation Intel® Core™ i7-6700K at 4.00GHz with 4 cores and 8 threads CPU, GeForce GTX 1060 6GB GPU and 64GB memory. It is running on Ubuntu 18.04 with CUDA 11.4.

From the distribution plot in Figure 9, we can see that the longest time taken is when executing the TensorFlow DN201 model. The average time taken per WSI is approximately 3.6 hours (0.16s per nucleus), as compared to the PyTorch DN201 model, 2.4 hours (0.1s per nucleus). Apart from the good score performance of PyTorch, its time complexity is also much lower, which is in our favor. For the three subsequent PyTorch DN21 models (3k model, 22k-crop model and 22k-alpha model), the time taken is even much lower: 1.72, 1.85 and 1.76 hours respectively per WSI, and an average of 0.07s per nucleus for all models. Since the size and weight of these lighter models are lesser than DN201, is it expected to compute in lesser time, but the fact that it produced better score performance than the heavier model is a new finding for

TABLE 6. Detailed comparison for selected images for each test set.

Test Set	Image no. (Image ID)	Model name	Class 0	Class 1	Class 2	Class 3	Total positive (% positive)	Allred Score
Test 1	24 (4305415)	TF-DN201	72575	13012	150	66	13228 (15.417%)	4
		PT-DN201	78834	6856	95	18	6969 (8.122%)	3*
	35 (4305255)	TF-DN201	3742	506	17377	30007	47890 (92.753%)	8*
		PT-DN201	3736	750	28890	18256	47896 (92.764%)	7
	37 (4305273)	TF-DN201	7173	661	5996	7279	13936 (66.019%)	7*
		PT-DN201	7169	820	7369	5751	13940 (66.038%)	6
Test 2	3 (4305367)	PT-DN201	170404	11	2	736	749 (0.438%)	4*
		PT-DN21	165254	751	0	5148	5899 (3.447%)	5
	11 (4305403)	PT-DN201	77688	791	10	20	821 (1.046%)	3
		PT-DN21	77836	648	2	23	673 (0.857%)	2*
Test 3	1 (4305349)	PT-DN21	40369	278	1	10	289 (0.711%)	2*
		PT-22k-DN21-crop	40094	542	4	18	564 (1.387%)	3
	11 (4305403)	PT-DN21	77836	648	2	23	673 (0.857%)	2*
		PT-22k-DN21-crop	76977	1470	23	39	1532 (1.951%)	3
	20 (4305453)	PT-DN21	87171	9437	52	1	9490 (9.818%)	3*
		PT-22k-DN21-crop	81603	14872	168	18	15058 (15.578%)	4
	24 (4305415)	PT-DN21	79608	6097	72	26	6195 (7.220%)	3*
		PT-22k-DN21-crop	72263	13192	246	102	13540 (15.78%)	4
	28 (4305285)	PT-DN21	100956	605	32	26	663 (0.652%)	2
		PT-22k-DN21-crop	100343	1111	115	50	1276 (1.256%)	3*
	29 (4305279)	PT-DN21	66546	3802	19	37	3858 (5.48%)	3*
		PT-22k-DN21-crop	62184	7991	120	109	8220 (11.675%)	4
	32 (4305435)	PT-DN21	4847	1542	15816	12941	30299 (86.209%)	7*
		PT-22k-DN21-crop	4883	943	12973	16347	30263 (86.107%)	8
	35 (4305255)	PT-DN21	3748	1117	30703	16064	47884 (92.741%)	7
		PT-22k-DN21-crop	3641	664	17140	30187	47991 (92.948%)	8*
	37 (4305273)	PT-DN21	7280	767	7011	6051	13829 (65.512%)	6
		PT-22k-DN21-crop	7062	834	5525	7688	14047 (66.545%)	7*
Test 5	8 (4305447)	PT-DN21	116654	1313	1	10	1324 (1.122%)	3
		PN+PT DN21	117045	924	1	8	933 (0.791%)	2*
	13 (4305343)	PT-DN21	70028	652	22	40	714 (1.009%)	3
		PN+PT DN21	70089	592	22	39	653 (0.923%)	2*
	16 (4305317)	PT-DN21	74493	817	5	16	838 (1.112%)	3
		PN+PT DN21	74712	601	5	13	619 (0.822%)	2*
Test 6	5 (5060537)	PT-22k-DN21-crop	55103	5	1	5	11 (0.02%)	2*
		PN+PT 22k DN21	55104	4	1	5	10 (0.018%)	4
	10 (4305409)	PT-22k-DN21-crop	289131	287	40	179	506 (0.175%)	2*
		PN+PT 22k DN21	289307	137	37	156	330 (0.114%)	4
	20 (4305453)	PT-22k-DN21-crop	81603	14872	168	18	15058 (15.578%)	4
		PN+PT 22k DN21	88080	8414	150	17	8581 (8.877%)	3*
	29 (4305279)	PT-22k-DN21-crop	62184	7991	120	109	8220 (11.675%)	4
		PN+PT 22k DN21	64677	5507	116	104	5727 (8.134%)	3*

* concordance scores or closest to the GT

this ER-IHC-stained WSI. For the two hybrid models combined with PN classification, both have similar computational times, 2.76 hours for the hybrid 3k model (PN+PT DN21) and 2.77 hours for the hybrid 22k model (PN+PT-22k DN21) per WSI, with 0.11s and 0.12s per nucleus respectively.

From Table 5, the second top-performing model is the non-hybrid of PT-DN21 with 1.72 hours per WSI and 0.07s per nucleus, which is one hour less than the hybrid model per WSI. However, this system can be left overnight or for a few days for the experts to attend to other important issues and come back to the system when it has completed the analysis to validate it. Both score and time are important but at this stage, we will take the highest-performing model to be considered the best configuration.

H. ISSUES AND RECOMMENDATIONS

There are two main issues which have been identified throughout this experiment and these will be discussed in detail here. The first issue is the non-concordance of image results', and the second issue is inaccurate segmentation, which can be divided into non-nuclei segmentation and over-sensitive segmentation.

1) NON-CONCORDANCE IMAGES

In Test 5 (Section IV-E), we listed out seven images that remain non-concordance for the rest of the tests, including image 28 for not being concordance for the best performing model (PN+PT DN21). In this section, we will look into the images at the cellular level to analyse them in detail as to why the models could not achieve concordance. The images are images 3, 4, 7, 9, 17, 27 and 28. Only part of the ROI-WSI will be captured as examples, together with the classified nuclei which are overlayed with colors according to the classes (blue: negative, yellow: positive-weak, orange: positive-moderate, red: positive-strong). These images have similar scores for all models, and the classified nuclei in

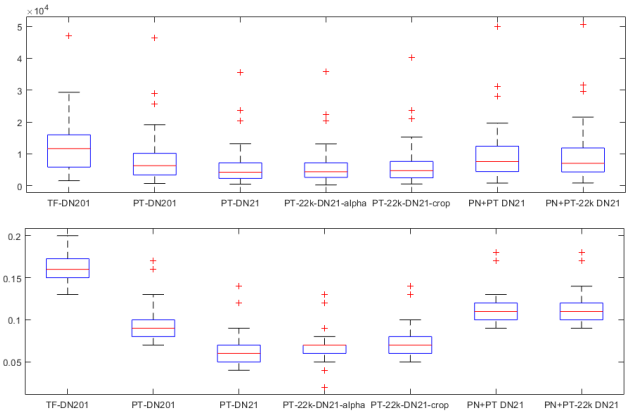

FIGURE 9. Distribution of time complexity in second (s) for all models in classifying the nuclei (top) per WSI; (bottom) per nuclei. Non-hybrid DN21 models show the lowest computational time, due to the lower number of dense layers and parameters.

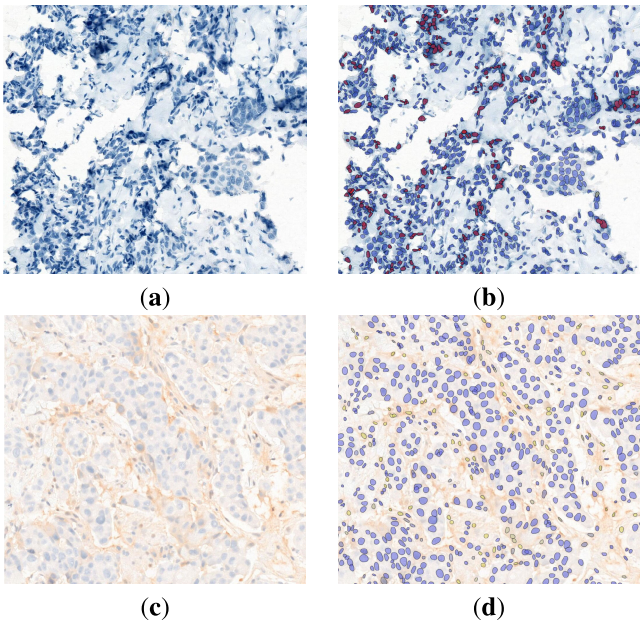

FIGURE 10. Example of misclassification in (a) and (b) Image 3 (4305367); and (c) and (d) Image 4 (4305305) which led to wrong scores. (a) and (c) are part of the ROI-WSI; (b) and (d) with the classified nuclei. For image 3, the dark blue nuclei confused the model as positive-strong nuclei (dark brown); and for image 4, the model detected the light brown nuclei as a positive-weak class.

the examples in Figure 10, 11 and 13 are captured from all models. For images 3 and 4, both have GT scores of 0, but the models gave scores of 4 or 5 for image 3, and 3 or 4 for image 4. Examples of part of the ROI-WSIs for these images are shown in Figure 10, (a) and (b) for image 3, and (c) and (d) for image 4. On looking at image 3, human eyes can clearly identify the nuclei as negative with unambiguous blue staining of various shades, from deep dark blue to the lightest blue. However, the models mistakenly classified many of the nuclei to the positive-strong class, specifically the deep dark blue color, as it might be confused with the dark brown nuclei.

For image 4, the WSI is visually having lightly brown staining, and similar-looking cells were confused with the positive-weak nuclei. The high Allred scores obtained by all models were due to the high positive nuclei proportion caused by the positive-weak nuclei, ranging from 5% to 13%, resulting in proportion scores of 2 to 3. This may be a situation where background staining of the cell cytoplasm may artefactually give a brownish hue over the nucleus. Cytoplasmic staining is not accepted in ER staining assessment.

For images 7, 9 and 17, the GT Allred scores for all were 2, but were wrongly scored by all models as 3, 5 and 4 respectively. In Figure 11, examples of ROIs from images 7 and 9 are shown. We can see that the cause leading to misclassification is similar for these images, where there are many brown-looking cells classified as positive nuclei, some are weak, moderate or strong, but many are classified as positive-weak. These are actually non-cancerous stroma cells which are in the connective tissue that provides the background on which the cancer cells grow. For image 17, we take the wrongly classified cells from the best model, a hybrid of PN classification with PT DN21 model, to be analyzed, as shown in Figure 12. There are a total of 29 positive nuclei, 10 are positive-weak and 19 are positive-strong. These gave an Allred score of 4, with a proportion score of 1 and an intensity score of 3. However, looking at the classified nuclei, they are not nuclei, but instead other artefacts in the image, just like some of the positive-strong nuclei in image 5 shown in Figure 8. The sizes are very small, with the area ranging from 2 to 46 micron^2. A typical diameter of the cancer nucleus would be 12 to 18 microns. Taking the smallest diameter and assuming an ellipse with 12 microns semi-major axis and 6 microns semi-minor axis, an area of 56.5 micron^2 is obtained. It is safe to assume a complete nucleus should be at least of this size. This type of misclassification perhaps can be mitigated or avoided by introducing another class, such as a non-nuclei class, to get cleaner results. On the other hand, another cause that contributes to this issue comes from inaccurate segmentation, where other than nuclei are segmented, such as stroma cells or artefacts.

Images 27 and 28, both have GT Allred scores of 3, but were wrongly scored as 2 for all models, except the non-hybrid 22k model obtained the correct score only for image 28. Part of the ROI-WSI for these images is shown in Figure 13, where we can see that image 27 visually looks negative ER staining. Upon review with the pathologists, image 28 clearly has some brownish-colored nuclei classifiable as positive-weak to positive-moderate, but many of these were missed by the models, except for the non-hybrid 22k model. The best model (PN+PT DN21) only detected 529 (0.521%) positive nuclei with 476, 32 and 21 for each

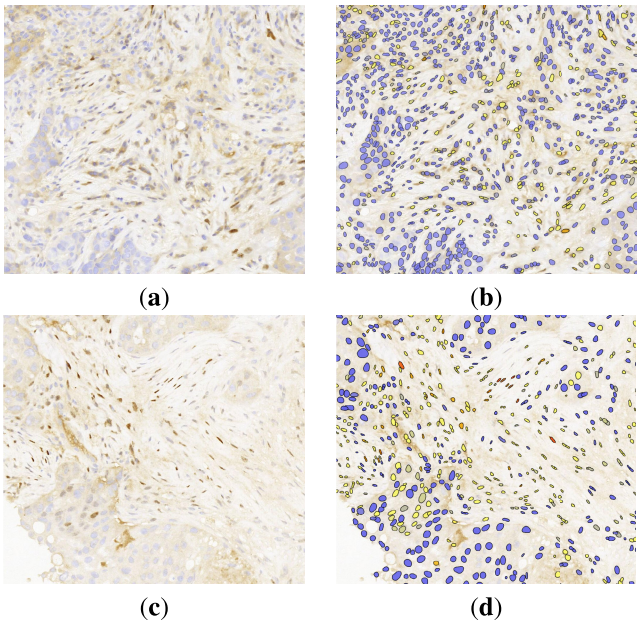

FIGURE 11. Example of misclassification in (a) and (b) Image 7 (4305459); and (c) and (d) Image 9 (4305441) which leads to wrong scores. (a) and (c) are part of the ROI-WSI; (b) and (d) with the classified nuclei. There are many of the nuclei classified as positive-weak and positive-moderate. These are actually non-cancerous stromal cells.

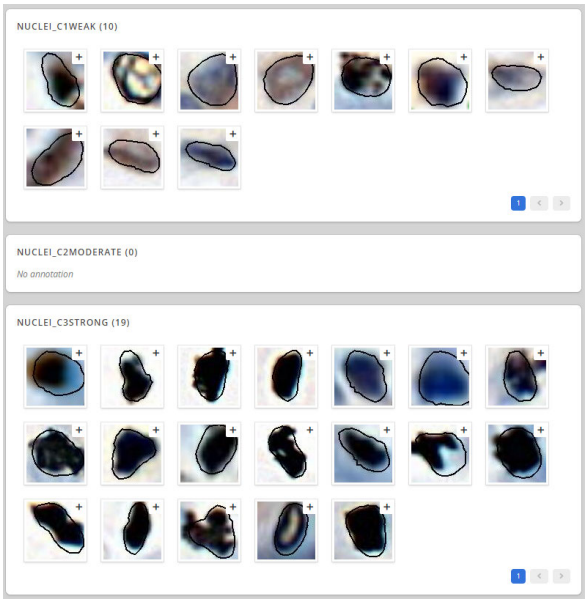

FIGURE 12. Wrongly classified cells as nuclei positive-weak and positive-strong in image 17 (4305267) which lead to an Allred score of 4. Most of these are actually artefacts, wrongly segmented as nuclei. This result is from the best model (PN+PT DN21) with the lowest classified positive proportion for this WSI.

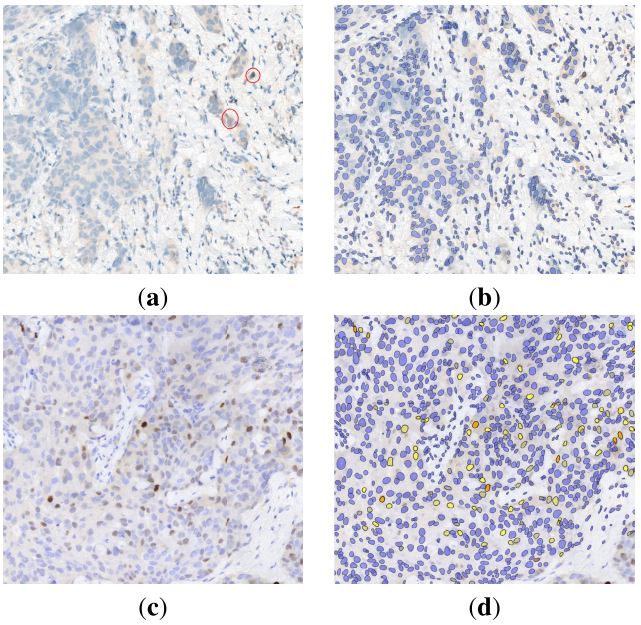

FIGURE 13. Example of misclassification in (a) and (b) Image 27 (4305379), and (c) and (d) Image 28 (4305285) with GT Allred score 3 which leads to wrong scores. (a) and (c) are part of the ROI-WSI; (b) and (d) with the classified nuclei. For image 27, although the nuclei visually look negative, on quick scanning, there are isolated positive-staining nuclei outside the main tumor cluster (see circle). For image 28, some of the nuclei are clearly positive-weak to positive-moderate in staining, but many were missed by the models.

FIGURE 14. Examples of nuclei classified as positive-weak for Image 28, with bluish and brownish hue, looking like the third last nucleus in Fig. 1 of negative class for 22k validated dataset.

WMS nuclei. For the non-hybrid 22k model, the total positively classified nuclei are 1276 (1.256%) with a breakdown of 1111, 115 and 50 for the WMS. When looking closely at the classified nuclei, the weak classified nuclei have a bluish and brownish hue, which might have confused the model, as shown in Figure 14. Referring to Figure 1, these positive-weak nuclei look more like the third last nucleus in the negative class of the 22k validated dataset. This explained why PN classification did not perform well for this image,

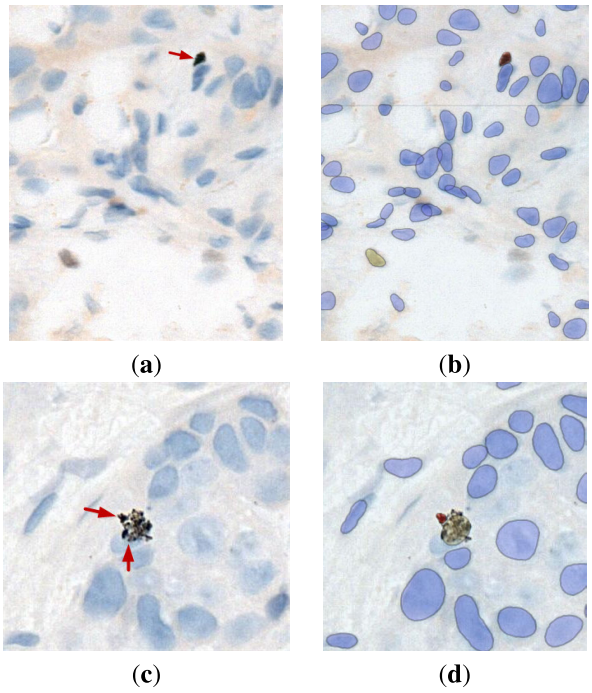

FIGURE 15. Image 1 (4305349) (a) and (c): part of the ROI-WSI; (b) and (d): with the classified nuclei. Example of segmentation and misclassification in Image 1 for other than nuclei. The stained nucleus in (a) and (b) is a dying one in pyknosis with non-specific condensation of stains with the nuclear material. For (c) and (d), the stained nucleus is in early mitosis (prophase), and also shows non-specific condensation of stains with the chromosomes. These should not be counted for ER staining assessment.

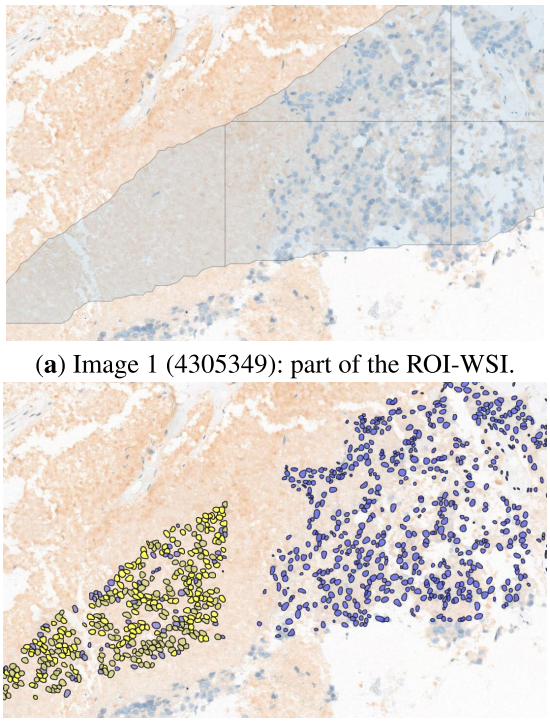

FIGURE 16. Example of inaccurate segmentation in Image 1, showing the ROI-WSI area that has no intact nuclei on the left side (necrotic area), but has many segmented nuclei and was classified as positive-weak.

because the bluish hue nuclei have been filtered out during the weighted hue threshold.

2) INACCURATE SEGMENTATION (SEGMENTING OTHER THAN NUCLEI)

Another issue that led to the misclassification of the nuclei was inaccurate segmentation, where many non-tumor-nuclei, nuclei in mitosis or artefacts were segmented and wrongly classified as one of the four classes. Some examples are shown in Figure 12, Figure 14 (second row, last image) and Figure 15. A more serious issue was identified in image 1, as shown in Figure 16. From the ROI-WSI region, the left side (light brown part) is actually a necrotic region where the tumor cells have died and nuclei have disintegrated but the Stardist segmentation algorithm detected the area as containing nuclei, and following this was the wrong classification as positive-weak nuclei. We refer to this issue as ''oversensitive segmentation''. We have identified two possible ways to overcome this issue. The first one is to train Stardist using IHC dataset, preferably ER biomarker. The current Stardist model used for segmentation was trained using H&E stain, but seems to work well even with IHC. Training the model with ER-IHC stain will further increase segmentation performance and eliminate this issue. The second way is to introduce a non-nuclei class, in addition to the NWMS classes. This class will take all non-nuclear particles or artefacts like the ones in Figure 12 into its class.

V. CONCLUSION

This paper provides a critical analysis of WSI scoring for ER-IHC stained pathological images, using several configurations of DenseNet architecture for nuclei classification. There are six tests altogether, with each test having only one change in the setting. We also introduced a modified PN classification to separate positive and negative nuclei in the first phase. The positive nuclei will be further classified into negative, positive-weak, positive-moderate and positive-strong using the DenseNet model to avoid any missed classification during the PN algorithm stage. The six tests were compared comprehensively, with the best concordance achieved by a hybrid model of PN with DenseNet of 21 layers. Out of 37 WSIs, 21 of them obtained concordance on the Allred score, and 30 of them are concordance with the suggested hormonal treatment. We also have identified several causes that lead to the non-concordances, particularly the confusion of dark blue stains, brownish small cells which are actually non-cancerous stromal cells but detected as nuclei, and also the confusion of the nuclei with both bluish hue and brownish hue. Another issue is the inaccurate segmentation, where many of the non-nuclei cells or artefacts were segmented and wrongly classified as one of the four classes. For oversensitive segmentation, there is a case where the model segmented

disintegrated nuclei on a necrotic region where the tumor cells have died, which can possibly be avoided by retraining the segmentation model with the ER-IHC dataset. The findings from this work can be a strong basis for future improvements of automated WSI scoring in ER-IHC using DL models, specifically DenseNet architecture and its hybrid algorithm.

ACKNOWLEDGMENT

The authors would like to thank their collaborating pathologists for their full support and guidance in providing the image dataset and ground truth for evaluation.

REFERENCES

[1] S. K. B. Sangeetha, R. Dhaya, D. T. Shah, R. Dharanidharan, and K. P. S. Reddy, "An empirical analysis of machine learning frameworks for digital pathology in medical science," *J. Phys., Conf. Ser.*, vol. 1767, no. 1, Feb. 2021, Art. no. 012031, doi: 10.1088/1742-6596/1767/1/012031.

[2] M. Mediouni, D. R. Schlatterer, H. Madry, M. Cucchiarini, and B. Rai, "A review of translational medicine. The future paradigm: How can we connect the orthopedic dots better?" *Current Med. Res. Opinion*, vol. 34, no. 7, pp. 1217–1229, Jul. 2018, doi: 10.1080/03007995.2017.1385450.

[3] W. Trochim, C. Kane, M. J. Graham, and H. A. Pincus, "Evaluating translational research: A process marker model," *Clin. Transl. Sci.*, vol. 4, no. 3, pp. 62–153, 2011.

[4] L. Pantanowitz, "Digital images and the future of digital pathology," *J. Pathol. Informat.*, vol. 1, no. 1, p. 15, Jan. 2010. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20922032

[5] S. W. Jahn, M. Plass, and F. Moinfar, "Digital pathology: Advantages, limitations and emerging perspectives," *J. Clin. Med.*, vol. 9, no. 11, p. 3697, Nov. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33217963

[6] C. I. Rodrigues-Fernandes, P. M. Speight, S. A. Khurram, A. L. D. Araújo, D. E. D. C. Perez, F. P. Fonseca, M. A. Lopes, O. P. de Almeida, P. A. Vargas, and A. R. Santos-Silva, "The use of digital microscopy as a teaching method for human pathology: A systematic review," *Virchows Archiv*, vol. 477, no. 4, pp. 475–486, Oct. 2020. [Online]. Available: https://link.springer.com/article/10.1007/s00428-020-02908-3

[7] A. S. Azam, I. M. Miligy, P. K.-U. Kimani, H. Maqbool, K. Hewitt, N. M. Rajpoot, and D. R. J. Snead, "Diagnostic concordance and discordance in digital pathology: A systematic review and meta-analysis," *J. Clin. Pathol.*, vol. 74, no. 7, pp. 448–455, Jul. 2021. [Online]. Available: https://jcp.bmj.com/content/jclinpath/74/7/448.full.pdf

[8] Y. Liu and L. Pantanowitz, "Digital pathology: Review of current opportunities and challenges for oral pathologists," *J. Oral Pathol. Med.*, vol. 48, no. 4, pp. 263–269, Apr. 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jop.12825

[9] L. Barisoni, K. J. Lafata, S. M. Hewitt, A. Madabhushi, and U. G. J. Balis, "Digital pathology and computational image analysis in nephropathology," *Nature Rev. Nephrology*, vol. 16, no. 11, pp. 669–685, Nov. 2020, doi: 10.1038/S41581-020-0321-6.

[10] M. García-Rojo, "International clinical guidelines for the adoption of digital pathology: A review of technical aspects," *Pathobiology*, vol. 83, nos. 2–3, pp. 99–109, 2016. [Online]. Available: https://www.karger.com/DOI/10.1159/000441192

[11] E. L. Clarke and D. Treanor, "Colour in digital pathology: A review," *Histopathology*, vol. 70, no. 2, pp. 153–163, Jan. 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/his.13079

[12] J. T. Abel, P. Ouillette, C. L. Williams, J. Blau, J. Cheng, K. Yao, W. Y. Lee, T. C. Cornish, U. G. J. Balis, and D. S. McClintock, "Display characteristics and their impact on digital pathology: A current review of pathologists' future 'microscope,'" *J. Pathol. Informat.*, vol. 11, no. 1, p. 23, Jan. 2020.

[13] F. Aeffner, M. D. Zarella, N. Buchbinder, M. M. Bui, M. R. Goodman, D. J. Hartman, G. M. Lujan, M. A. Molani, A. V. Parwani, K. Lillard, O. C. Turner, V. N. P. Vemuri, A. G. Yuil-Valdes, and D. Bowman, "Introduction to digital image analysis in whole-slide imaging: A white paper from the digital pathology association," *J. Pathol. Informat.*, vol. 10, no. 1, p. 9, Jan. 2019.

[14] S. Nam, Y. Chong, C. K. Jung, T.-Y. Kwak, J. Y. Lee, J. Park, M. J. Rho, and H. Go, "Introduction to digital pathology and computer-aided pathology," *J. Pathol. Translational Med.*, vol. 54, no. 2, pp. 125–134, Mar. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32045965

[15] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Med. Image Anal.*, vol. 33, pp. 170–175, Oct. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841516301141

[16] F. Xing and L. Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review," *IEEE Rev. Biomed. Eng.*, vol. 9, pp. 234–263, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5233461/pdf/nihms-818917.pdf

[17] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, "Artificial intelligence in digital pathology—New tools for diagnosis and precision oncology," *Nature Rev. Clin. Oncol.*, vol. 16, no. 11, pp. 703–715, Nov. 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31399699

[18] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan, "Digital pathology and artificial intelligence," *Lancet Oncology*, vol. 20, no. 5, pp. e253–e261, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470204519301548

[19] R. Pell, K. Oien, M. Robinson, H. Pitman, N. Rajpoot, J. Rittscher, D. Snead, and C. Verrill, "The use of digital pathology and image analysis in clinical trials," *J. Pathol. Clin. Res.*, vol. 5, no. 2, pp. 81–90, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cjp2.127

[20] R. Colling, "Artificial intelligence in digital pathology: A roadmap to routine use in clinical practice," *J. Pathol.*, vol. 249, no. 2, pp. 143–150, Oct. 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/path.5310

[21] S. Lal, R. Desouza, M. Maneesh, A. Kanfade, A. Kumar, G. Perayil, K. Alabhya, A. K. Chanchal, and J. Kini, "A robust method for nuclei segmentation of H&E stained histopathology images," in *Proc. 7th Int. Conf. Signal Process. Integr. Netw. (SPIN)*, Feb. 2020, pp. 453–458.

[22] M. C. Chang and M. Mrkonjic, "Review of the current state of digital image analysis in breast pathology," *Breast J.*, vol. 26, no. 6, pp. 1208–1212, Jun. 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/tbj.13858

[23] I. Ilić, N. Stojanović, N. Radulović, V. Živković, P. Randjelović, A. Petrović, M. Božić, and R. Ilić, "The quantitative ER immunohistochemical analysis in breast cancer: Detecting the 3 + 0, 4 + 0, and 5 + 0 allred score cases," *Medicina*, vol. 55, no. 8, p. 461, Aug. 2019.

[24] M. E. H. Hammond, D. F. Hayes, A. C. Wolff, P. B. Mangu, and S. Temin, "American society of clinical oncology/college of American pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer," *J. Oncol. Pract.*, vol. 6, no. 4, pp. 195–197, Jul. 2010.

[25] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, "The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis," *Comput. Biol. Med.*, vol. 128, Jan. 2021, Art. no. 104129. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482520304601

[26] F. Xing, Y. Xie, H. Su, F. Liu, and L. Yang, "Deep learning in microscopy image analysis: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4550–4568, Oct. 2018.

[27] H. P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, "Deep learning in medical image analysis," *Adv. Exp. Med. Biol.*, vol. 1213, pp. 3–21, Jun. 2020.

[28] L. Zhang, L. Lu, I. Nogues, R. M. Summers, S. Liu, and J. Yao, "DeepPap: Deep convolutional networks for cervical cell classification," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 6, pp. 1633–1643, Nov. 2017.

[29] Z. Gao, L. Wang, L. Zhou, and J. Zhang, "HEp-2 cell image classification with deep convolutional neural networks," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 2, pp. 416–428, Mar. 2017.

[30] J. Liu, B. Xu, L. Shen, J. Garibaldi, and G. Qiu, "HEp-2 cell classification based on a deep autoencoding-classification convolutional neural network," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 1019–1023.

[31] F. Xing, T. C. Cornish, T. Bennett, D. Ghosh, and L. Yang, "Pixel-to-pixel learning with weak supervision for single-stage nucleus recognition in Ki67 images," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 11, pp. 3088–3097, Nov. 2019.

[32] M. F. Ahmad Fauzi, W. S. H. M. Wan Ahmad, M. F. Jamaluddin, J. T. H. Lee, S. Y. Khor, L. M. Looi, F. S. Abas, and N. Aldahoul, "Allred scoring of ER-IHC stained whole-slide images for hormone receptor status in breast carcinoma," *Diagnostics*, vol. 12, no. 12, p. 3093, Dec. 2022. [Online]. Available: https://www.mdpi.com/2075-4418/12/12/3093

[33] M. Weigert, U. Schmidt, R. Haase, K. Sugawara, and G. Myers, "Star-convex polyhedra for 3D object detection and segmentation in microscopy," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3655–3662.

[34] R. Marée, L. Rollus, B. Stévens, R. Hoyoux, G. Louppe, R. Vandaele, J.-M. Begon, P. Kainz, P. Geurts, and L. Wehenkel, "Collaborative analysis of multi-gigapixel imaging data using cytomine," *Bioinformatics*, vol. 32, no. 9, pp. 1395–1401, May 2016, doi: 10.1093/BIOINFORMATICS/BTW013.

[35] W. S. H. M. Wan Ahmad, M. J. Hasan, M. F. A. Fauzi, J. T. H. Lee, S. Y. Khor, L. M. Looi, and F. S. Abas, "Nuclei classification in ER-IHC stained histopathology images using deep learning models," in *Proc. IEEE Region 10 Conf. (TENCON)*, Nov. 2022, pp. 1–5.

[36] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125

[37] M. M. Sehmi, M. A. Fauzi, W. W. Ahmad, and E. W. L. Chan, "Pancreatic cancer grading in pathological images using deep learning convolutional neural networks," *F1000Research*, vol. 10, p. 1057, Nov. 2021.

[38] M. F. A. Fauzi, H. N. Gokozan, C. R. Pierson, J. J. Otero, and M. N. Gurcan, "Prognostic reporting of P53 expression by image analysis in glioblastoma patients: Detection and classification," in *Health Information Science*, X. Yin, K. Ho, D. Zeng, U. Aickelin, R. Zhou, and H. Wang, Eds. Cham, Switzerland: Springer, 2015, pp. 165–173.

WAN SITI HALIMATUL MUNIRAH WAN AHMAD received the B.Eng. degree in electronic engineering majoring in multimedia and the M.Eng.Sc. and Ph.D. degrees from Multimedia University, Cyberjaya, Malaysia. Since 2017, she has been a Postdoctoral Researcher. In 2021, she joined the Artificial Intelligence for Digital Pathology Centre, Faculty of Engineering, Multimedia University, focusing on pathological image analysis. Her research interests include medical image analysis, content-based image retrieval, and data mining.

MOHAMMAD FAIZAL AHMAD FAUZI (Senior Member, IEEE) received the B.Eng. degree in electrical and electronic engineering from Imperial College London, London, U.K., in 1999, and the Ph.D. degree in electronics and computer science from the University of Southampton, Southampton, U.K., in 2004. From May 2013 to June 2014, he was with the Clinical Image Analysis Laboratory, The Ohio State University, Columbus, OH, USA, where he involved in digital histopathology, especially on cancer and disease analysis. He is currently a Professor with the Faculty of Engineering, Multimedia University (MMU). He has published more than 100 journal and conference papers. His main research interests include signal and image processing, pattern recognition, computer vision, and medical imaging. He is currently an Executive Committee of IEEE Region 10 (Asia Pacific).

MD JAHID HASAN received the bachelor's degree in mechatronic engineering from the Faculty of Manufacturing and Mechatronic Engineering, University Malaysia Pahang, Malaysia. He is currently pursuing the master's degree with the Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia. His research interests include machine learning, deep learning, and image processing.

ZAKA UR REHMAN received the B.Eng. degree in computer system engineering majoring in multimedia from The Islamia University of Bahawalpur, Pakistan, and the M.Sc. degree in electrical engineering majoring in computer vision and machine learning for medical imaging from COMSATS University Islamabad, Pakistan, in 2017. He is currently pursuing the Ph.D. degree with the Artificial Intelligence for Digital Pathology Centre, Faculty of Engineering, Multimedia University, focusing on pathological image analysis. From 2018 to 2021, he was a Lecturer with the Department of Computer Science, The University of Lahore. His research interests include medical image analysis, content-based image retrieval, and computer vision.

JENNY TUNG HIONG LEE received the degree in medicine from Kursk State Medical University, Russia, and the master's degree in anatomic pathology from the University of Malaya. She is currently an Anatomic Histopathologist with the Sarawak General Hospital, Malaysia. She has been actively involved in clinicopathological case reports, pathological-clinical joint research, and digital pathology research. She is also involved in the training of medical officers and master's students. She is a member of the International Association of Pathology-Malaysian Division and the Malaysian Medical Association.

SEE YEE KHOR received the medical degree from the National University of Ireland, and the Master of Pathology degree from the University of Malaya. After that, she was an Anatomic Pathologist with the Hospital Queen Elizabeth Sabah. She is currently an Anatomic Pathologist with the Hospital Seberang Jaya. She has been actively involved in the clinicopathological study, joint research, and digital pathology research. She is a member of the International Association of Pathology (Malaysian Division).

LAI-MENG LOOI received the degree in medicine from the University of Singapore, trained in surgical pathology with the University of Malaya (UM), the Royal Postgraduate Medical School, U.K., and the Brigham and Woman's Hospital, USA, and the Ph.D. degree from UM. She is currently a Distinguished Professor with UM, with a concurrent appointment as a Senior Consultant Histopathologist with the University Malaya Medical Centre. Her research interests include amyloidosis, nephropathology, oncopathology, and innovative diagnostics, on which she has more than 200 publications and delivered more than 400 guest lectures. She is a Commissioner of the Lancet Commission on Diagnostics. Her contributions to medical science have been recognized with several accolades, including the ASEAN Outstanding Scientist Award, the Merdeka Award 2016 (Health, Science and Technology Category), and the Honorary Professor of the Chinese Academy of Medical Sciences-Peking Union Medical College.

FAZLY SALLEH ABAS received the Ph.D. degree in image analysis and pattern recognition from the University of Southampton, U.K., in 2004. Since then, he has been with Multimedia University, Malaysia, where he is currently an Associate Professor with the Faculty of Engineering and Technology. His research interests include digital image processing and analysis for medical/clinical applications, robot vision, surveillance, and process automation. He is a member of the Centre for e-Health, the Centre for Advanced Robotics, and the Centre for Engineering Computational Intelligence.

AFZAN ADAM received the Ph.D. degree from the University of Leeds, U.K. She is currently a Senior Lecturer with the Center for Artificial Intelligence and Technology, Universiti Kebangsaan Malaysia (UKM). Her digital pathology research projects were closely joint by Hospital UKM's pathologist; include building platforms and models for screening, detecting and classifying cancer cells for breast, blood, prostate, and cervix. She is also involved directly with various e-health research and development projects in the O&G, Ophthalmology, and Emergency Department, Hospital UKM. A few graduate students under her supervision are involved in speech recognition and prediction models on time-series and normal datasets. Her research interest includes digital pathology, besides its application into teaching machine learning and image processing skills in other domains. She is a member of the Digital Pathology Association and the Malaysian Board of Technologists. She is also one of the Head for Artificial Intelligence for Digital Pathology Consortium and the Committee for Malaysian Digital Pathology Guidelines. She has been constantly invited for her research talks nationally and internationally.

ELAINE WAN LING CHAN received the Ph.D. degree in pharmacognosy from Monash University Malaysia. She is currently a Senior Lecturer/Researcher with the Institute of Research, Development and Innovations, International Medical University (IMU). Her passion in digital science and its applications had later drove her to develop herself in programming and machine learning through online certification courses. Her desire to serve community has also driven her to lead numerous community projects to translate research in community settings. Balancing numerous responsibilities in academic administration as well as research. She has published 24 papers in acclaimed scientific journals with more than 600 citations and H-index of ten. With her passion and keen interest in artificial intelligence, she is actively involved in setting up the Malaysia diabetic cohort, developing new algorithm for computer vision in digital pathology, driving initiated for new policy and guidelines for digital pathology in Malaysia as well as creating new digital solution for healthcare digital transformation initiatives. Over the years, she had won numerous awards. Amongst these awards are Leaders in Innovation Fellowships Programme Royal Society of Engineering, U.K., in 2021; the IMU Core Value Award, in 2021; the IMU Partnership Excellence in Community Service Award, in 2020; and the IMU Leadership in Community Service, in 2020.

SEI-ICHIRO KAMATA (Senior Member, IEEE) received the M.S. degree in computer science from Kyushu University, Fukuoka, Japan, in 1985, and the Doctor of Computer Science degree from the Kyushu Institute of Technology, Kitakyushu, Japan, in 1995. From 1985 to 1988, he was with NEC Ltd., Kawasaki, Japan. In 1988, he joined the Kyushu Institute of Technology. In 1990 and 1994, he was a Visiting Researcher with The University of Maine, Orono, ME, USA. From 1996 to 2001, he has been an Associate Professor with the Department of Intelligent System, Graduate School of Information Science and Electrical Engineering, Kyushu University. Since 2003, he has been a Professor with the Graduate School of Information, Production and Systems, Waseda University, Fukuoka. His research interests include image processing, pattern recognition, image compression, and space-filling curve application. He is a member of The Institute of Image Information and Television Engineers (ITE) in Japan.