Human papilloma virus detection in oropharyngeal carcinomas with in situ hybridisation using hand crafted morphological features and deep central attention residual networks

Shereen Fouad*

School of Computing and Digital Technology, Birmingham City University, Birmingham, United Kingdom

Gabriel Landini

Oral Pathology Unit, School of Dentistry,University of Birmingham, Birmingham, United Kingdom

Max Robinson

Centre for Oral Health Research, Newcastle University, Newcastle, United Kingdom

Tzu-Hsi Song[1,1]

Oral Pathology Unit, School of Dentistry,University of Birmingham, Birmingham, United Kingdom

Hisham Mehanna

Institute for Head and neck studies and education (Inhanse), University of Birmingham, Birmingham, United Kingdom

Abstract

Human Papilloma Virus (HPV) is a major risk factor for the development of oropharyngeal cancer. Automatic detection of HPV in digitized pathology tissues using *in situ* hybridisation (ISH) is a difficult task due to the variability and complexity of staining patterns as well as the presence of imaging and stain-

*Shereen Fouad

Email address: Shereen.Fouad@bcu.ac.uk (Shereen Fouad)

[1] Current affiliation: Vascular Biology Program, Boston Children's Hospital, Boston, United States

[2] Professor Mehanna is a National Institute for Health Research (NIHR) Senior Investigator. The views expressed in this article are those of the author(s) and not necessarily those of the NIHR, or the Department of Health and Social Care.

ing artefacts. This paper proposes an intelligent image analysis framework to determine HPV status in digitized samples of oropharyngeal cancer tissue microarrays (TMA). The proposed pipeline mixes handcrafted feature extraction with a deep learning for epithelial region segmentation as a preliminary step. We apply a deep central attention learning technique to segment epithelial regions and within those assess the presence of regions representing ISH products. We then extract relevant morphological measurements from those regions which are then input into a supervised learning model for the identification of HPV status. The performance of the proposed method has been evaluated on 2,009 TMA images of oropharyngeal carcinoma tissues captured with a ×20 objective. The experimental results show that our technique provides around 91% classification accuracy in detecting HPV status when compared with the histopatholgist gold standard. We also tested the performance of end-to-end deep learning classification methods to assess HPV status by learning directly from the original ISH processed images, rather than from the handcrafted features extracted from the segmented images. We examined the performance of sequential convolutional neural networks (CNN) architectures including three popular image recognition networks (VGG-16, ResNet and Inception V3) in their pre-trained and trained from scratch versions, however their highest classification accuracy was inferior (78%) to the hybrid pipeline presented here.

Keywords: histology, human papilloma virus, in situ hybridisation, deep learning, machine learning

1. Introduction and Background

Oropharyngeal Carcinoma (OPC), is a type of head and neck cancer affecting the epithelium of the oropharyngeal mucosa which has seen a significant increase in incidence over the last decade [1]. Infection with certain types of Human Papilloma Virus (HPV) (called ‘high risk’) is considered a major risk factor for the development of OPC. According to [1], HPV is responsible for over 70% of OPC in Europe and the USA. It has been found in [2] that patients with tumours positive for HPV infection (HPV+) tend to have better prognosis than those with HPV negative (HPV-) tumors. Therefore, there is interest in histological assessment of HPV status in OPC samples as both diagnostic and prognostic features. This is often performed using microscopy on biopsy or resection samples processed with, for example, *in situ* hybridization (ISH), a laboratory method that enables the direct histological detection of viral genome sequences in tissues [2, 3]. Owing to the high morphological diversity and complexity of pathology samples, HPV assessment/diagnosis remains a challenging task to histpathologists [1, 2, 3]. Figure 1 shows HPV+ and HPV- samples of OPC images processed by ISH.

Over the past years there have been significant improvements in Digital Pathology (DP) that allow whole histopathology slides to be digitised as high-resolution images. This, in turn, has opened the possibility of using image processing routines and intelligent techniques to resolve a variety of histopathology problems (typically tumor *vs.* non-tumour classification). Various approaches have been proposed to integrate computer vision, machine learning and deep learning tools to analyse histopathological data, e.g. [4, 5, 6]. Those efforts in general aim to reduce the workload of histopathologists and therefore reduce cost and time of diagnosis while performing at known levels of accuracy. Furthermore, it has been suggested that they could be used as a *second readers* helping to reduce inter- and intra-observer variability among pathologists [7].

Deep learning techniques have made an important contribution in histopathology image classification tasks, specially those involving convolutional neural networks (CNNs). For example, in [6] CNN was used to classify breast cancer histopathology images independent of magnification, by exploiting two different CNN architectures: single-task and multi-task to predict malignancy and image magnification level, respectively. A CNN-based approach was also proposed in [4] for the binary classification (carcinoma *vs.* non-carcinoma)of H&E stained histological images of breast cancer by extracting information at different scales, from nuclei to overall tissue organization.

Although deep learning-based techniques have proved in general a superior performance against traditional methods based on handcrafted features in digital pathology tasks (e.g. [8]), they suffer from several drawbacks. First, training a deep CNN from scratch requires large amounts of annotated images, which is very expensive and difficult to obtain in practice. Secondly, deep learning is also deemed as complex process, computationally expensive and with poor explainability (i.e., a black box approach). In addition, histological techniques often suffer from variability in staining uptake, distortion and sectioning and staining artefacts. These challenges make it necessary to sometimes design bespoke approaches that suite the complexity of digital pathology tasks. As a consequence, several methods have been proposed in the literature to integrate *biologically interpretable* handcrafted features with deep learning models to improve performance on complex histopathology images. Such hybrid models have shown to perform better than their deep learning counterparts alone. E.g. in [9], textural handcrafted features were augmented in the input of the CNN to improve the detection of tumor cells in H&E histology images. In [10] mitosis detection was implemented combining a light CNN mode with handcrafted features (morphology, color, and texture features) which provided a more accurate and faster solution with lower computational requirements when compared to other existing approaches.

Despite AI-based techniques have made an important contribution to the analysis of tumour images, their applications for the assessment of HPV status are still lacking. Some progress in assessing tumour HPV status has been made using multivariate statistics on radiomics data, e.g. [11, 12, 13]. In [14] machine learning models were used to classify HPV status in computed tomography (CT) using texture analysis applied to regions of interest in contrast-enhanced neck CT of OPC cases. However, the study utilized a small experimental sample (n=107 subjects) and the accuracy achieved when compared to ground truth clinical cohorts was not particularly high (75.7%) despite it being higher those achieved by two blinded neuroradiologists (accuracy of 44.9 and 55.1% respectively). However, the analysis of histological data should provide a more direct, richer and more accurate results, based on morphological features at the cellular and tissue level. HPV status in tissues has been investigated in [15] using texture features of the nuclear chromatin condensation from images to indirectly infer HPV positivity. The overall accuracy, however, was low (68%) to be useful for robust diagnostic purposes. To the best of our knowledge, analysis and detection of HPV status from ISH histopathology images using intelligent methods (deep learning combined with machine learning) have not yet been investigated.

In [16] we reported a preliminary machine learning/imaging workflow to identify high risk HPV genomes in digitized tissue micro-array samples of OPC processed by ISH. The segmentation algorithm used mathematical morphology to extract ISH-stained regions, however, the identification of HPV in [16] was performed on features extracted from whole core tissue images which are most often heterogeneous in composition, i.e. they include neoplastic epithelium as well as non epithelial (stromal) regions and can potentially contain staining artefacts. In reality, robust HPV status assessment requires the conditional detection of ISH chromogen precipitation (blue stained features with the methodology used here) in epithelial tumour regions of the specimen while ignoring other tissue regions and artefacts (typically non-specific precipitation of the chromogen, drying artefacts, non-specific leukocyte cytoplasm staining) all of which can lead to false positive readings. Not surprisingly, accurate identification of positive ISH products in epithelial regions can be a difficult task for histopathologists and it is further complicated for two reasons: 1) the blue hybridisation staining is fine patterned and 2) the counterstain dye used to reveal the general morphology of tissues (Red Counterstain II) is not tissue type-specific and has similar staining uptake in stroma and epithelium, therefore discriminating between these two tissue types poses additional hurdles to the observer.

This paper presents an extension of our preliminary work introduced in [16] for automated identification of HPV status in ISH processed sections, this time guiding the analysis to the epithelial tissue compartment, while excluding features in non-epithelial tissues (e.g. connective stroma, background) and image artefacts. To this end, we exploited a deep central attention residual network, proposed in [17], to segment epithelial regions from the information provided by the counterstain dye and then restricted the identified ISH precipitiation region (indicating the presence of HPV genomes) to the segmented epithelial compartments. The proposed pipeline mixes handcrafted feature extraction with a deep learning for epithelial region segmentation as a previous step. The extracted morphological measurements are then fed to a Machine learning algorithm to classify microscopy images as HPV+ or HPV-.

2. Methodology

The dataset consists of tissue micro-arrays (TMAs) slides containing 2,009 OPC cores plus associated clinical data. The TMAs were stained by ISH and prepared at the Institute of Cancer and Genomic Sciences, University of Birmingham, UK. OPC specimens were processed using the Ventana INFORM HPV III system (Roche), consisting of a mixture of HPV genomic probes of high-risk HPV strains which are labelled with an enzyme capable of precipitating a chromogen molecule on the tissues (in this case, the chormogen was nitroblue tretrazolium or NBT/BCIP, visible as a blue navy colour). This enables the visualisation of the hybridised genomes directly in the samples, in the case of OPC, in the nucleus of the infected epithelial cells. A counterstain (Red Counterstain II, pink in colour) is also used to facilitate identifying the general tissue morphology of the sample. Figure 1 shows core samples of HPV- (a, b) and HPV+ (c, d) OPC tumours processed by ISH, where the characteristic blue staining patterns in epithelial cell nuclei indicate positive viral infection (c and e). As illustrated in Figure 1(a), tissues may also include non-specific (i.e. non-HPV associated) artefacts which can mislead the HPV status assessment.

TMA slides were digitised using an Olympus BX50 microscope (Olympus Optical Co. Ltd, Tokyo, Japan) equipped with a ×20 magnification objective. The individual core images were approximately 3300×3300 pixels in size (inter-pixel distance = 0.367μm). The associated data included seven clinical measures: *patient gender, age at diagnosis, tumour size, lymph node status,*

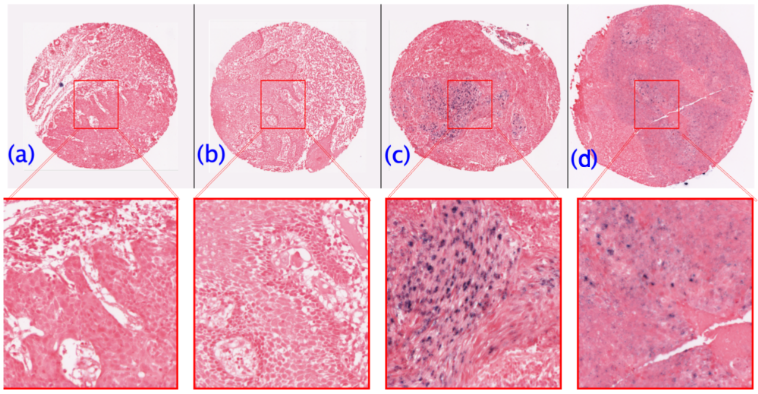


Figure 1: *In situ* hybridisation for high-risk HPV strains in oropharyngeal carcinoma tissue cores: (a) and (b) are HPV- tumour cases while (c) and (d) are HPV+ tumour cases (the blue stain in the epithelial tissue indicate the presence of the HPV genomes). Note that it can be difficult to assess whether the blue stain is localised in tumour epithelial regions or in other tissues (e.g. stroma).

smoking status, overall years of survival and recurrence free survival (in years). In addition, the data also included the P16 H-score [18]. P16 is a protein expressed in HPV+ tumours which has been called a 'surrogate marker' for HPV status, although the exact relation between P16 expression and HPV detection in tumours is still not completely resolved (e.g. HPV is not detectable in some proportion of P16 overexpressing tumours). While overexpression of P16 has been suggested to indicate favourable prognosis independent of HPV status [19], more recently it has been shown that P16 is not sufficiently accurate for prognosis and a HPV specific test is required [20]. The P16 H-score is assessed semi-quantitatively using immunohistochemistry: an expert histopathologist interprets the images and calculates a score given by the product of the percentage of P16+ cells and the intensity class they belong to using a categorical scale from 1 (weak) to 3 (strong) staining. In clinical practice, P16 expression in a tumour is considered as 'positive' using a cut-off value [21]: strong and diffuse nuclear and cytoplasmic staining present in $\geq 70\%$ of the tumour. That is an H-score equivalent to ≥ 2 intensity $\times \geq 70\% =$ H-score ≥ 140. For HPV status detection (the 'reference standard' in this paper), the histopathologist examined the TMA slides, recording them as either HPV+ or HPV- depending on the presence or absence of hybridisation products in epithelial regions of the samples[22]. In this study, 1355 cores were labelled as HPV- and 654 were labelled as HPV+.

The proposed HPV status detection algorithm is shown in Figure 2 and can be described in five main steps as follows:

Step 1 - Colour deconvolution of the original ISH image. This step is used to generate three 'staining' channel images based on the colours of the individual dyes. The procedure was originally introduced by Ruifrok and Johnston [23] and assumes that the dye colours mix subtractively. This type of colour separation is particularly useful in histopathology as it allows exploring different histological components based on dye uptake. We determined a set of colour vectors to perform colour deconvolution on ISH images [24] so the contribution of the ISH dyes is separated into three channels containing the contribution of: i) tissues counterstained with Red Counterstain II (pink channel) (Figure 2(b)), ii) blue stained regions (NBT/BCIP, blue channel) (Figure 2(c)) and iii) a 'residual' component (showed as the complementary of the other two colours, Figure 2(d)) which retains 'unexplained' colours (those which do not correspond to the subtractive mixing of the other two dyes).

Step 2 - Extraction of stained regions from the blue stain (NBT/BCIP) channel image. First, the empty unstained background is separated from the stained regions to ensure that the subsequent process is applied only to the stained components for segmentation consistency. This was achieved by (a) computing the minimum pixel intensity image between the blue and pink image channels to guarantee capturing all stained pixels, (b) applying Gaussian blur function to the generated image to reduce staining heterogeneity, (c) binarising the result using an auto threshold method to separate the core from its background and finally (d) removing isolated noisy particles from the binary image using an opening by reconstruction procedure (after five consecutive erosions, determined experimentally).

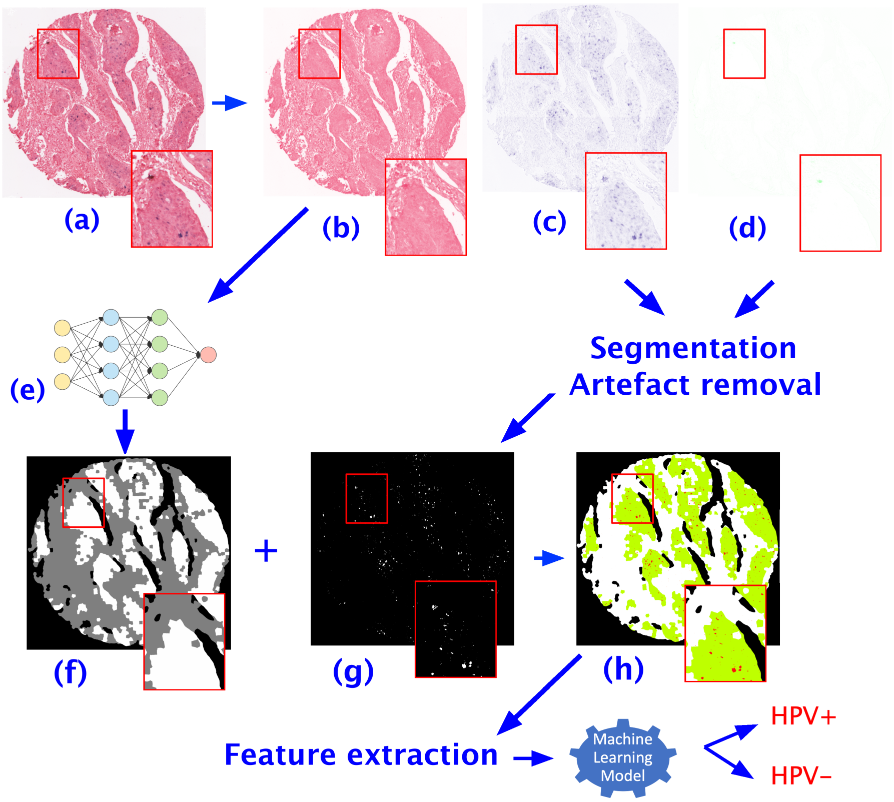


Figure 2: Overview of the proposed method. (a) RGB image of a core tissue sample processed by ISH, (b) 'pink' tissue channel (Red Counterstain II), (c) blue channel image (stained regions with BT/BCIP) and (d) residual colour component image obtained after colour deconvolution (see Figure 3). (e) The deep Central Attention Residual (CAR) learning network (see Figure 4) is used to identify epithelium regions from (b) (results shown in white in (f)). In (g) are shown the segmented regions from the blue stain (NBT/BCIP) channel after artefact removal (see Figure 3). In (h) is shown the the final segmentation result with the 'blue stained regions' (in red) located in the detected epithelial regions (in green). Morphological features are extracted from these and submitted to the machine learning model (see Figure 5).

Second, an auto threshold method is applied to the blue channel (Figure 2(c)) to extract binary objects representing the NBT/BCIP stained regions. Experimentally Renyi's entropy auto threshold [25] returned the best results when compared to other available methods. For simplicity, these binary objects shall be referred to as 'blue stained regions' for the remaining of the paper.

Third, 'artefact' regions are minimised in the blue stain image. Artefacts are extraneous (non-histological) features arising from a variety of causes, such as contamination, faulty tissue processing, etc., and which can lead to erroneous interpretation of the hybridisation results. Figure 3 top left shows an example with imaging artefacts (in this case large dark regions). A considerable number of such artefacts can be identified by shape and size (e.g. not matching expected histological structures), while others can be detected by their colour, i.e. they are also detected in the residual colour deconvolution channel. This is so because artefacts rarely have the same colour characteristics of true histological features, therefore appear prominently in the residual channel (green regions in Figure 3). To identify artefact regions in the blue channel we firstly binarise the residual channel and apply a morphological opening by reconstruction, using a copy of the blue stained regions image as mask and the binarised residual channel as seeds. The result retains the blue (binarised) regions that also overlap regions in the residual channel; those are subtracted from the original blue regions image. The complete process is illustrated in Figure 3.

Step 3 - Epithelial segmentation using a deep Central Attention Residual (CAR) network.

This step is applied to the counterstain (pink) channel (Figure 2(b)) to assess which blue stained regions are located in epithelial tissue, while ignoring those in others non-epithelial regions (e.g. connective stroma, background). This is necessary because the diagnostic value of the ISH products in OPC relies on identifying HPV genomes strictly in the epithelial cells. To perform this, we exploited our deep learning-based approach, presented in [17], to identify the epithelial component in the pink counterstained image (Figure 2(b)). This is a particularly difficult task even for expert human observers because all tissues

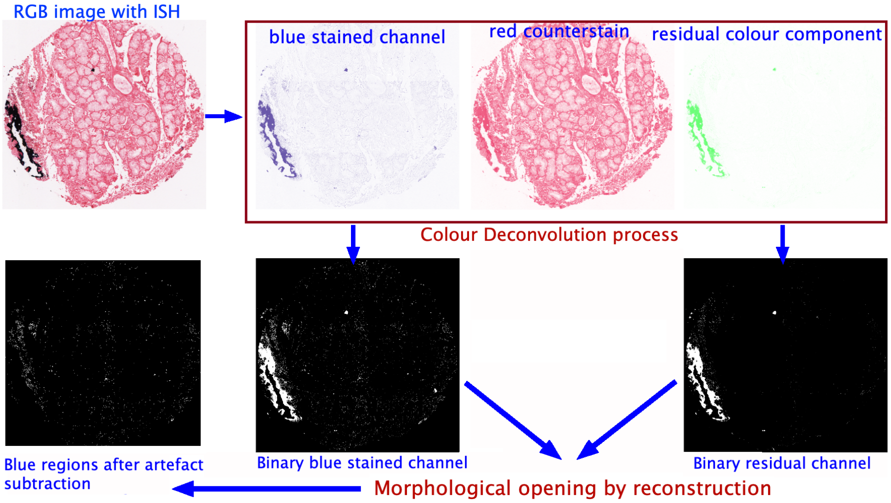

RGB image with ISH
blue stained channel
red counterstain
residual colour component
Colour Deconvolution process
Blue regions after artefact subtraction
Binary blue stained channel
Binary residual channel
Morphological opening by reconstruction

Figure 3: An example to illustrate the process of removing histological artefacts by colour deconvolution. The RGB image (containing large, dark arefacts) is colour-deconvolved into three channels. The artefacts appear in the residual colour channel as they are not purely stained with the blue chromogen. After binarisation, the artefacts in the blue channel are identified as those regions with a spatially related counterpart in the binarised residual channel.

are stained in the same colour and therefore a good understanding the fine tissue morphology is required. Despite this, the pink channel still retains a level of tissue morphological information, without the influence of ISH precipitation products or the presence of staining artefacts (also shown in Figure 2(b)).

To achieve this segmentation step, the pink channel is first partitioned into superpixels (2,500 pixels in size) using the SLIC algorithm [26]. The resulting superpixel regions are then framed within square patches of size 100×100 pixels, which are used as input images for a deep Central Attention Residual (CAR) network that is trained based on a gold standard produced by an experienced microscopist to discriminate superpixels belonging to epithelium from those belonging to non-epithelial (e.g. stroma) regions by considering the features of each superpixel and its surrounding area.

The network consists of a) four convolution layers that generate a number of feature maps and reduce their dimensions, b) four CAR blocks, c) an average pooling layer and d) a softmax classifier. Each CAR block utilizes the concept of residual network to efficiently learn specific features and achieve more accurate identifications than traditional CNNs [27]. In addition, the CAR blocks include three convolution layers, a central attention (CA) unit, which emphasizes the features information of the central area of the input image, batch normalization and a rectified linear unit (ReLu) activation function. Figure 4 shows the architecture of our CAR network and the whole framework for epithelium segmentation based on the counterstain (pink) channel. More details of the architecture of the CAR network can be found in [17].

The segmentation result is shown in Figure 2(f), where the detected epithelial sections are shown in white and non-epithelial in grey. The blue regions in the epithelium are obtained by the intersection of the epithelial regions in Figure 2(f) with the segmented blue regions in Figure 2(g). The final segmentation results are shown in Figure 2(h), with epithelial regions labelled in green and the blue regions in red.

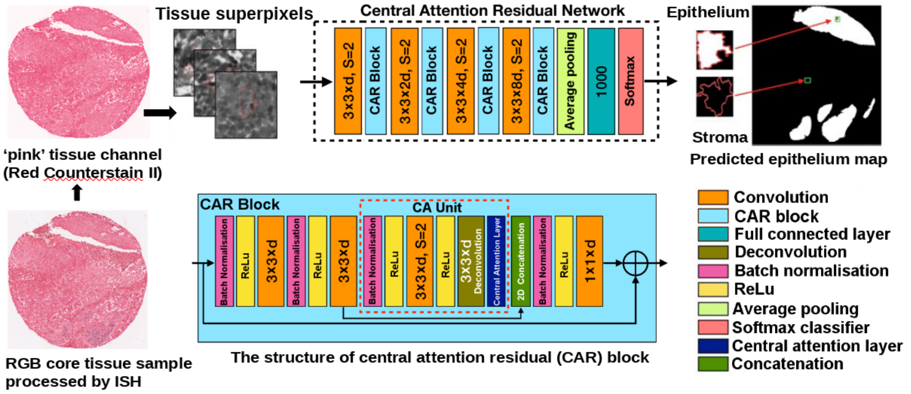
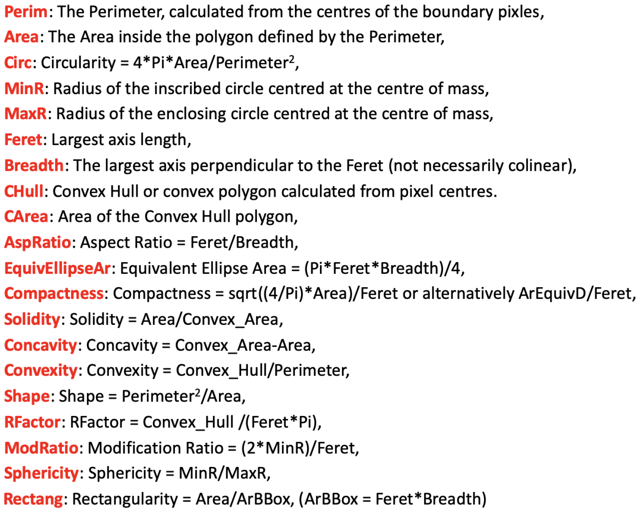


Figure 4: Overview of the Epithelial segmentation method using a deep Central Attention Residual (CAR) network . The parameters d and S denote the number of feature maps and the sliding pixel shift, respectively. The blue lower box shows the structure of the central attention residual (CAR) block.

Step 4 - Extraction of morphological features from blue stained regions.

A set of morphological features describing shape and size is extracted from the blue stained regions described previously, using an ImageJ plugin that computes geometrical properties of binary 8-connected regions (Particles8, [28]). The distribution characteristics of these features in HPV+ and HPV- images (as determined by the 'reference standard') were preliminary inspected using Kolmogorov–Smirnov tests to select those with distributions statistically different (i.e. p value$\leq$0.05) across HPV+ and HPV- images. This procedure identified 20 morphological descriptors (listed in Figure 5) to produce a feature matrix of size 20×m, where m is the total number of blue stained regions per image. To create a single vector summarising and describing the morphological characteristics of a single image, four different distribution statistics of the morphological parameters were computed (mean, minimum, maximum and standard deviation), bringing the total number of extracted features per image to (4×20=80), plus, the number of blue stained regions in the image.



Perim: The Perimeter, calculated from the centres of the boundary pixles,
Area: The Area inside the polygon defined by the Perimeter,
Circ: Circularity = 4*Pi*Area/Perimeter^2,
MinR: Radius of the inscribed circle centred at the centre of mass,
MaxR: Radius of the enclosing circle centred at the centre of mass,
Feret: Largest axis length,
Breadth: The largest axis perpendicular to the Feret (not necessarily colinear),
CHull: Convex Hull or convex polygon calculated from pixel centres.
CArea: Area of the Convex Hull polygon,
AspRatio: Aspect Ratio = Feret/Breadth,
EquivEllipseAr: Equivalent Ellipse Area = (Pi*Feret*Breadth)/4,
Compactness: Compactness = sqrt((4/Pi)*Area)/Feret or alternatively ArEquivD/Feret,
Solidity: Solidity = Area/Convex_Area,
Concavity: Concavity = Convex_Area-Area,
Convexity: Convexity = Convex_Hull/Perimeter,
Shape: Shape = Perimeter^2/Area,
RFactor: RFactor = Convex_Hull /(Feret*Pi),
ModRatio: Modification Ratio = (2*MinR)/Feret,
Sphericity: Sphericity = MinR/MaxR,
Rectang: Rectangularity = Area/ArBBox, (ArBBox = Feret*Breadth)

Figure 5: Description of the extracted morphological features.

Step 5 - Determining HPV status using machine learning algorithms.

The extracted morphological features are pre-processed, using standard procedures (normalization, noise removal) and passed to a supervised machine learning algorithm classifier which predicts the HPV status as + or -. The dataset (containing imaging and clinical data) is split into training and testing sets, so the classifier learns the characteristics of the HPV+ and HPV- features. After training, the classifier can be used to detect the HPV status of unseen images.

3. Experiments and Results

3.1. Evaluation of the proposed model and comparison against classical machine learning methods (handcrafted features)

This section evaluates the effectiveness of the proposed method in the context of HPV status as a binary classification (HPV+ or HPV-) and compares it against our preliminary approach in [16], which exploits classical machine learning methods (using handcrafted features only) and hence doesn't discriminate epithelial from non-epithelial regions. It also compares the results against other baseline methods using non-imaging features (clinical data, survival characteristics and P16 H-score) without any of the histological features derived programmatically from the image.

A description of the feature sets used in the experiments is provided in Table 1. Imaging features in **PECP16** and **PE** were extracted using the proposed framework while in **PWCP16** and **PW** they were extracted using the analysis described in [16]. The architecture of our deep central attention learning was the same as in [17]. Gaussian parameter (σ^2) of the central attention function was set to 0.4. The Adam optimiser was used with a learning rate of 10^{-3} to minimise the network loss and cross-entropy was used as the loss function.

To assess the predictive ability of the feature sets provided in Table 1, five well-known classifier algorithms were used, namely (1) *Support Vector Machine (SVM)*[29, 30], a procedure that finds a hyperplane in the feature space that maximizes the margin (distance) between data points of classes, (2) *k-Nearest Neighbours (KNN)*[31, 30], which classifies data points by a majority vote of its k nearest neighbors, (3) *Random Forests (RF)* [32, 30], based on the construcion

Table 1: Description of the feature sets used in the experiments.

Feature set	**Description**
Pathology-**E**pithelial-**C**linical-**P16** (**PECP16**)	Histopathology imaging features obtained from the proposed approach *(morphology of blue stained regions in epithelial regions)* + clinical features *(listed in section 2)* + P16 H-score
Pathology-**W**holeimage-**C**linical-**P16** (**PWCP16**)	Histopathology imaging features obtained from the analysis in [16] *(morphology of blue stained regions anywhere in the slide image)* + clinical features *(listed in section 2)* + P16 H-score
Pathology-**E**pithelial (**PE**)	Histopathology imaging features only obtained from the proposed approach *(morphology of blue stained regions in epithelial regions)*
Pathology-**W**holeimage (**PW**)	Histopathology imaging features only obtained from the analysis in [16] *(morphology of blue stained regions anywhere in the slide image)*
Clinical-**P16** (**CP16**)	Clinical features *(listed in section 2)* + P16 H-score
Clinical (**C**)	Clinical features only *(listed in section 2)*

of multiple decision trees with a final classification decision made based on the majority of the trees, (4) *Logistic Regression (LR)*[30], a statistical model that uses a logistic function to estimate the probability of a certain class and (5) *Multilayer Perceptron (MLP)* [30] that consists of a feed-forward supervised learning network with up to two hidden layers. In all experiments, the (hyper-)parameters of the classification algorithms were tuned using a randomized the grid-search technique.

To build the HPV status classifiers, the dataset was randomly partitioned into 70%, 10%, and 20% for training, validation, and testing, respectively. Data was normalized to zero mean and unit variance to ensure that all features contributed to the classification comparably. The unequal distribution of classes within the dataset (1355 HPV- *vs.* 654 HPV+) was addressed using the SMOTE technique [33], which looks at the feature space for the minority class data points and oversamples it by considering k nearest neighbours.

HPV status classification results were evaluated using the two following performance measures:

- **The classification accuracy**, measures all of the correctly identified cases,

$$\frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

  where TP, FP, FN and TN denotes true positives, false positives, false negatives and true negatives, respectively.

- **The F1-score**,

$$\frac{2 \times (Recall \times Precision)}{(Recall + Precision)} \tag{2}$$

  where $Precision$ is the ratio of correctly predicted positive records to the total predicted positive records

$$\frac{TP}{TP+FP} \tag{3}$$

  and $Recall$ is the ratio of correctly predicted positive records to the all data records in a class,

$$\frac{TP}{TP+FN} \tag{4}$$

For fair comparisons, the experiments were run five times for each classical classifier and we reported the average classification accuracy and F1-score over five runs in Tables 2 and 3, respectively.

Tables 2 and 3 show that feature sets containing histology imaging features perform generally better than those with only clinical features or P16 H-score across all learning algorithms. The results also reveal that the classification using histological features (**PECP16** and **PE**) outperforms our preliminary analysis in [16] (**PWCP16** and **PW**), across all learning algorithms. We believe this is because the morphology of the blue stained regions in epithelium is more representative of how pathologists assess visually HPV status unlike in [16], where there was no distinction made of where the ISH precipitation

Table 2: Comparative analysis against classical (using handcrafted features only) Average classification accuracy (along with standard deviation over five runs) obtained with five well-known supervised learning algorithms using the features in table 1.

Learning Algorithms	**PECP16**	**PWCP16**	**PE**	**PW**	**CP16**	**C**
KNN	**0.89** (±0.09)	0.85 (±0.14)	0.87 (±0.10)	0.83 (±0.07)	0.75 (±0.07)	0.74 (±0.07)
SVM	**0.91** (±0.07)	0.89 (±0.15)	0.90 (±0.08)	0.88 (±0.07)	0.79 (±0.09)	0.74 (±0.07)
MLP	**0.89** (±0.09)	0.85 (±0.11)	0.87 (±0.07)	0.85 (±0.05)	0.74 (±0.09)	0.72 (±0.09)
LR	**0.89** (±0.10)	0.86 (±0.10)	0.88 (±0.09)	0.86 (±0.09)	0.76 (±0.08)	0.72 (±0.12)
RF	**0.90** (±0.12)	0.88 (±0.11)	0.89 (±0.09)	0.88 (±0.09)	0.79 (±0.09)	0.77 (±0.08)

Best results are marked with **bold** font.

products were localised. Furthermore, results obtained using **PECP16** (combination of histology imaging features, clinical data and P16 score) is better than the ones obtained using **PE** (histology imaging features only), across all learning algorithms. This result suggests that clinical features also play a role in determining HPV status.

The best result was obtained by the SVM algorithm using **PECP16** feature set with an accuracy and F1-score of (91%) and (89%), respectively. The SVM was used with the following parameters: 'C'= 10, 'gamma'= 75 and 'kernel'= 'rbf'. By contrast, clinical (clinical data *without* P16 H-score) provided an accuracy and F1-score of (74%) and (65%), respectively, much inferior results using the same algorithm (SVM), which highlights the importance of the histological imaging features extracted using the proposed imaging workflow.

3.2. Comparison with end-to-end deep learning classification methods

In order to further evaluate the efficiency of our proposed approach in detecting HPV status we compared it against end-to-end deep learning classification

Table 3: Average F1-scores (along with standard deviation over five runs) obtained with five well-known supervised learning algorithms using the features in table 1.

Learning Algorithms	**PECP16**	**PWCP16**	**PE**	**PW**	**CP16**	**C**
KNN	**0.87** (±0.14)	0.80 (±0.15)	0.85 (±0.07)	0.79 (±0.06)	0.72 (±0.08)	0.73 (±0.07)
SVM	**0.89** (±0.10)	0.85 (±0.11)	0.88 (±0.09)	0.83 (±0.10)	0.74 (±0.10)	0.65 (±0.13)
MLP	**0.86** (±0.11)	0.84 (±0.09)	0.85 (±0.07)	0.80 (±0.11)	0.72 (±0.12)	0.69 (±0.09)
LR	**0.88** (±0.11)	0.85 (±0.07)	0.86 (±0.1)	0.81 (±0.05)	0.72 (±0.09)	0.65 (±0.07)
RF	**0.88** (±0.11)	0.87 (±0.15)	0.87 (±0.05)	0.86 (±0.06)	0.77 (±0.09)	0.76 (±0.12)

Best results are marked with **bold** font.

for HPV status in ISH processed images. While our hybrid method combines CAR deep learning network, classical machine leaning algorithms, handcrafted morphology features of blue stained regions in epithelium, clinical measurements and P16 scores, the end-to-end deep learning classification determines the HPV status by learning directly from the whole ISH processed images and don't focus on the biologically important regions (epithelial). Furthermore, it doesn't have the capacity to encounter and learn from other relevant information that might boost the learning process such as those provided by clinical measurements or P16 scores.

In this experiment, the ISH images were randomly partitioned into groups containing 70%, 10%, and 20% of the images for training, validation, and testing, respectively. Images were then resized and normalized to a zero mean and unit variance to assist in faster convergence. Different image sizes were tried (200×200, 400×400, 600×600, 800×800) to determine the best performing image size, which was 400×400. As discussed earlier, our data set suffers from an unequal distribution of classes (1355 HPV- vs. 654 HPV+). This imbalance in the training set could affect the performance of the CNN as it would not get the optimized results for the less popular class (HPV+). Hence, data augmentation to handle the cardinality of the training set for all classes was achieved by oversampling the HPV+ images in the training set. Data augmentation purposely changes the appearance of some images in the training examples, before passing them into the network for training. We applied image rotations by 90°, 180° and 270° as well as horizontal flips but no adjustments of contrast or intensity were applied, in order to preserve the color and morphological properties of the histological images.

3.2.1. Comparative Analysis of the Proposed Framework with sequential Convolutional Neural Network (CNN) architectures

The performance of several Convolutional Neural Network (CNN) architectures to classify HPV status by learning directly from the original ISH processed images was examined. CNN is one of the most powerful and successful deep learning approaches used in the analysis of cancer images (e.g. [34], [4],[35], [36],[37]). CNN models exploit local feature detectors or filters over the whole image to measure the correspondence between individual image and class label within the training set. Then, the dimensionality of the feature space is reduced using an aggregation or pooling function. In this experiment, we empirically evaluated several possible CNN architectures but we report results of the most successful ones in Table 4. Other CNN architectures are possible, however performing an exhaustive evaluation is highly time consuming and out of the scope of this research.

Inspired by [37], we examined the performance of multiple sequential CNN architectures that were used in similar pathology image classification problems (e.g. [34], [35], [36]). In particular, we changed various settings in CNN model characteristics including network depth, layer properties, kernel sizes and number of filters as described later. The attempted networks were composed of multiple blocks of convolution layer followed by (RELU) rectified linear activation function, Batch Normalization, Max pooling and Dropout, as shown in

Figure 6. ReLU helps to speed up the convergence learning and introduce the non-linearity [38] [3]. Batch Normalization aims to stabilize training and make tuning hyperparameters [39]. The max pooling layer reduces the spatial dimension with a filter of 2×2 and a stride of length equal to 2. The pooled output of the last convolutional layer is fed to the one fully-connected layer that has 256 neurons across all attempted networks. Dropout layer was applied on convolutional layers as well as after the fully connected layer with a keep probability of 0.25 and 0.5, respectively. Dropout is commonly used to regularize deep neural networks and it also helps the network to generalize and not overfit. The output of the fully-connected layer was fed into the Softmax classifier to predict if an image was HPV+ or HPV-.

As shown in Table 4, different numbers of feature maps were attempted such that layers early in the network architecture learn fewer convolutional filters (32), and going deeper in the network, the number of filters were increased to 64, 128 and 256 (a common practice when designing CNN networks). We also attempted different kernel sizes in convolutional layers; these determine the number of kernels to convolve with the input volume. Kernel sizes used were in the range of 3, 5, 7 to help learn larger spatial filters and reduce volume size. Same padding was applied, meaning that the size of output feature-maps are the same as the input feature-maps. Each of the attempted models was trained on the training set the performance evaluated using the validation set.

For all the attempted CNN, a parameter exploration was performed using training and the validation sets. The parameter selection was done according to validation accuracy. As illustrated in 4, Model no. 3 returned the best validation accuracy. For this model, optimal CNN parameter values for number of epochs, initial learning rate and learning rate decay were found to be 50, 1e-2 and 1e-3 respectively.

[3] The RELU is a piecewise linear function that outputs the input directly if is positive, or zero otherwise, following function: f(x) = max (0, x). It is one of the most common and best performing activation functions for many types of neural networks.

Table 4: Details of most successful empirically evaluated CNN architectures for HPV status (HPV+ or HPV-) prediction applied on core tissue samples processed by ISH.

Model No.	**Total No. of layers**	**No. of convolution layers blocks**	**Kernel sizes in convolution layers**	**No. of feature maps**	**Validation accuracy**
1	6	3	3,3,3	32, 64, 128	0.70
2	6	3	7,5,3	32, 64, 128	0.72
3	9	4	3,3,3,3	32, 64, 128, 256	**0.75**
4	9	4	7,5,3,3	32, 64, 128, 256	0.69
5	12	5	7,5,3,3,3	32, 64, 128, 128, 256	0.62
6	12	5	7,5,5,3,3	32, 64, 128, 128, 256	0.63

Best results are marked with **bold** font.

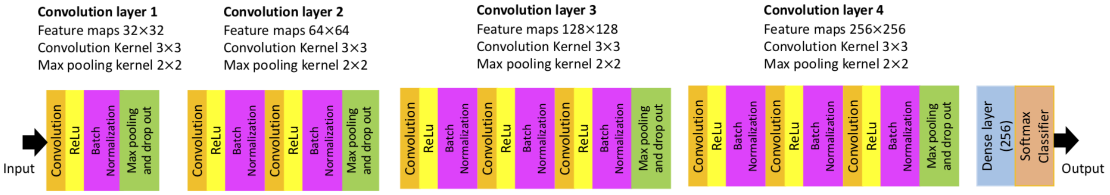


Figure 6: Architecture of the most successful empirically evaluated CNN (Model No. 3 in Table 4) for HPV status prediction applied on core tissue samples processed by ISH.

3.2.2. Comparative Analysis of the Proposed Framework with three popular CNN Architectures (VGG-16, Resnet and Inception V3)

In this experiment we compare our work against three popular CNN architectures (VGG-16 [40], Resnet [41] and Inception V3[42]). These architectures were specifically chosen for comparison because they have been widely adopted for image recognition tasks and their performances are commonly used as benchmarks for other architectures. In addition, they have been shown to perform well in pathology image classification tasks (e.g. [43]). Furthermore, pre-trained versions of these models are available so that a transfer learning approach can be adopted.

VGG16 is a CNN architecture used to win ILSVR (Imagenet) competition in 2014. One main feature of VGG-16 is the use of convolution layers of 3x3 filters with a stride 1 while using the same padding and maxpool layer of a 2x2 filters of stride 1. This arrangement of convolution and max pool layers was consistently used throughout the whole architecture. Our VGG-16 network had two fully connected layers followed by a softmax function for classification. The network was composed of 16 convolutional layers belonging to five convolution blocks. The small kernel sizes allow for training a deeper CNN while still reserving the fine-grained information in the network. Here, VGG-16 was trained from scratch using the formerly described architecture.

VGG16 networks exploit deeper networks to improve classification performance, however, they are often more difficult to train. The performance of a ResNet (Deep Residual Learning for Image Recognition) was also examined. This procedure skips the connection in convolution blocks by sending the previous feature map to the next convolution for a smoother training process. Here a compact version of ResNet was trained from scratch. ResNet performed (3, 4, 6) stacking with (64, 128, 256, 512) convolutional layers. This implies, the first convolutional layer in ResNet had a total of 64 filters. Then, we stacked 3 sets of residual modules. Each residual module had three convolutional layers which learned 32, 32 and 128 filters respectively. This is followed by another stack of four sets of residual modules, each had three convolutional layers which learned 64, 64, and 256 filters. The final stack had six sets of residual modules, where each convolutional layer learned 128, 128, and 512 filters. Spatial dimensions were reduced between first, second and third convolutional block layers and a softmax classifier applied.

The Inception model, also known as GoogleNet, has been developed by Google during the ImageNet Recognition Challenge. Inception network has a lower computational cost and memory requirement when compared to the VGG and ResNet, which makes it more suitable for large data classification tasks. The Inception network consists of a collection of Inception modules, which is a block of parallel convolutional layers with different kernel-sized filters (e.g. 3×3) and a 3×3 max pooling layer, the results of which are then concatenated. Our network included five convolutional layers, each followed by a batch normalization layer, 2 pooling layers and 11 inception modules.

VGG16, ResNet and Inception were optimized for hyper-parameters by the randomized grid search method. For VGG16 and Inception, optimal parameter values for number of epochs, initial learning rate and learning rate decay were 50, 1e-2 and 1e-2 respectively. For ResNet, optimal parameter values for number of epochs, initial learning rate, learning rate decay and classification threshold of stochastic gradient descent algorithm were found to be 50, 1e-2, 1e-5, and 0.9 respectively. The best performing CNN model (model 3 in Table 4 and Figure 6), VGG16, ResNet and Inception networks were trained from scratch and deployed on the entire data set using the optimized parameters.

In this experiment we compare our work against VGG16, ResNet and Inception-V3 networks using a transfer learning process, where the parameters are initially trained on the ImageNet [44] dataset. ImageNet is a dataset containing thousands of images of different objects and scenes used to train and evaluate image classification models. Transfer learning is a well-known machine learning technique where a model developed for a learning task is reused as the starting point for another relevant learning task. This allows for faster and more accurate training by transferring the knowledge from very large public image data sets to the studied problem. Transfer learning has previously been used to train models to detect various diseases from pathology images [45]. Here, the transfer learning approach was applied by fine-tuning three pre-trained models (VGG16, ResNet and Inception-V3) on our imaging dataset. Model weights were initialized based on pre-training on the ImageNet dataset, except for the final, fully connected layers which were randomly initialized. Additional training was performed using our training set images to fine-tune these networks for detecting the HPV status. We applied the three pre-trained models on the previously partitioned (training, validation and testing) data sets. We used the optimizer Adam, batch size 32, with 100 epochs using a learning rate of 10^{-3}. The model that achieved the highest performance on the validation dataset was selected for evaluation on the testing dataset. This process was repeated for each architecture (VGG16, ResNet and Inception-V3).

3.2.3. Experimental Results

For fair comparisons with the above experiments (section 3.1), all deep learning models were run five times on the randomly partitioned data and we report the average sensitivity, specificity, accuracy and F1-score for the testing data over five runs in Tables 5, respectively.

In terms of classification accuracy, VGG-16 (trained from scratch) returned the lowest F1-score for all evaluation measures, CNN (Model No.3 in Table 4) performed relatively better than VGG-16 and ResNet was deemed as the best performing end-to-end deep learning network with classification accuracy of 78%. On the other hand, our imaging framework (hybrid method), which combines CAR deep learning network with classical machine leaning algorithms, returned the best results across all evaluation measures when acting on **PECP16** feature set (handcrafted morphology features of blue stained regions in epithelium, clinical measurements and P16 scores) as well as **PE** feature set (handcrafted morphology features of blue stained regions in epithelium).

Table 5: Comparative analysis of proposed framework with other CNN architectures in terms of average Sensitivity, Specificity, Accuracy and F1-score.

Learning Algorithms	**Sensitivity**	**Specificity**	**Accuracy**	**F1-score**
VGG-16 (trained from scratch using the studied dataset)	0.70 (±0.17)	0.65 (±0.15)	0.68 (±0.17)	0.66 (±0.15)
VGG-16 (pre-trained on imagenet dataset)	0.54 (±0.20)	0.86 (±0.18)	0.68 (±0.19)	0.71 (±0.18)
Inception-V3 (trained from scratch using the studied dataset)	0.65 (±0.14)	0.64 (±0.17)	0.68 (±0.15)	0.68 (±0.15)
Inception-V3 (pre-trained on imagenet dataset)	0.70 (±0.18)	0.53 (±0.18)	0.62 (±0.19)	0.69 (±0.18)
Resnet (trained from scratch using the studied dataset)	0.77 (±0.13)	0.74 (±0.10)	0.78 (±0.14)	0.77 (±0.13)
Resnet (pre-trained on imagenet dataset)	0.76 (±0.19)	0.42 (±0.18)	0.61 (±0.18)	0.72 (±0.19)
Best performing CNN (Model No.3 in Table 4)	0.69 (±0.15)	0.72 (±0.13)	0.71 (±0.12)	0.70 (±0.15)
Proposed framework applied on imaging information, clinical measurements and P16 scores (**PECP16** feature set - Table 1)	**0.89** (±0.07)	**0.91** (±0.08)	**0.91** (±0.07)	**0.89** (±0.10)
Proposed framework applied on imaging information only (**PE** feature set - Table 1)	0.88 (±0.07)	0.90 (±0.08)	0.90 (±0.09)	0.89 (±0.05)

Values in brackets indicate the standard deviation over five runs. Best results are marked with **bold** font.

4. Discussion

The first experiment revealed that classifying HPV status using the morphology of blue stained regions in the epithelium tissue compartment is more accurate than when considering whole tissue regions. This was perhaps expected because the virus infects epithelial cells and excluding non-epithelial regions from the analysis should in reduce the possibility of false positive detection. It also highlights the importance of preliminary tissue segmentation before extracting ISH features. However, the detection accuracy also improves further when including clinical data (**PECP16** feature set). The second experiment indicated that the proposed handcrafted feature technique outperforms the end-to-end deep learning classification methods (including the pre-trained models where the parameters were pre-trained on the ImageNet dataset) in this particular problem. There are a number of possible reasons for this: **(1)** Deep learning-based methods learn the HPV status using the whole core tissue image, which contains a mixture of epithelial and stroma tissues, i.e. not focusing on features in epithelial cells. In contrast, the deep central attention learning technique that identifies epithelial regions (step 3 in the imaging workflow), allows the assessment of ISH products in the target tissue, **(2)** Deep learning-based methods seem to be unable to capture critical imaging information such as that provided by the morphological descriptors (step 4 in our imaging workflow), **(3)** The ISH images pose a particularly challenging problem due to the complexity of staining patterns and the presence of staining artefacts. Unlike the proposed workflow which removes certain type of staining artefacts (step 2 in our imaging workflow), deep learning methods might be unable to avoid the effect of such artefacts resulting in false positive results. **(4)** It is harder to incorporate the clinical features in a deep learning image classification frameworks. They mainly rely on imaging features without considering important clinical handcrafted features such as P16 score, age at diagnosis or smoking status. However, our future studies will examine the integration of such clinical features to deep learning image classification frameworks.

We believe that the method presented allows more explainability in the analysis of ISH images when compared with deep learning alone. Understanding the actual model and exploiting handcrafted features appears to be intuitive and straightforward and less ambiguous for applications such as diagnosis where the understanding of tissue changes in terms of known morphological architectural features is most desirable.

(5) Although transfer learning has proven to be successful in various image classification tasks, its application to medical images is still debatable because the pre-training of models is done using large datasets that might not be directly relevant to medical images (e.g., CT, MRI, microscopy images). For example, ImageNet pre-training is done using natural images (e.g. plants, sports, people, animals, etc.), which are different in content from medical images. Recent research suggested that ImageNet pre-trained models are of limited help for some tasks, including medical imaging [46]. Our results show that the application of transfer learning to our images resulted in an inferior detection performance when compared to the proposed approach.

5. Conclusion

The incidence of HPV-related oropharyngeal cancer has been reported to be in the increase in the Western world. Determining the HPV status in histopathology is therefore essential for accurate patient diagnosis, prognosis as well as for epidemiological studies. Unfortunately the task of assessing HPV status in ISH slides is both challenging and time-consuming for pathologists and therefore would significantly benefit from automation. In addition, defining formal numerical methods to diagnosis is likely to improve its reproducibility by reducing the level of subjectivity inherent in perceptual tasks.

We presented an intelligent technique for the detection of HPV status in tumours from digitized samples using ISH. The framework consists of a segmentation algorithm, based on mathematical morphology to identify ISH products, plus a deep learning network that identifies epithelium. ISH by being a technique based on hybridisation of complementary strands of DNA and RNA sequences has many uses in molecular biology, apart from the viral genome detection shown here, therefore the principles presented are likely to be transferable to a variety of other applications.

Our experimental results from the analysis of 2,009 TMA images predicted the HPV status with 91% accuracy. This outperformed baseline methods which exploit other learning predictors, including clinical data, P16 H-score and their combinations. The results also revealed that the morphology of the blue stained regions in the epithelium are better HPV status predictors than the morphology of blue stained regions in the whole core tissue.

The work also showed that the results obtained from the handcrafted feature set, compared favourably with popular end-to-end deep learning networks including CNN, VGG-16, ResNet and Inception-V3 which learn directly from the colour images. Interestingly, our results outperformed those obtained from deep learning architectures pre-trained with the ImageNet dataset. Approaches based on morphologically-relevant data representing biological levels of structure and organisation might be preferable to the 'black box' approach of deep learning because of the 'explainability', in biological terms, of the results, specially in this type of life-critical applications. In addition, unavoidable artefacts in histological preparations might be better detected and controlled by procedures like those described here to prevent them acting as adversarial examples that might lead to misinterpretation of the histological scenes. At the same time, unlike deep learning methods, our approach uses an 'interpretable' machine learning model enabling pathologists to understand why the model has taken such decision and check the plausibility of computer-based image classification.

Acknowledgment

The research reported in this paper was supported by the Engineering and Physical Sciences Research Council (EPSRC), UK through funding under grant EP/M023869/1 'Novel context-based segmentation algorithms for intelligent mi-

croscopy'.

References

[1] H. Mehanna, M. Evans, M. Beasley, S. Chatterjee, M. Dilkes, J. Homer, J. O'Hara, M. Robinson, R. Shaw, P. Sloan, Oropharyngeal cancer: United Kingdom National Multidisciplinary Guidelines, The Journal of Laryngology & Otology 130 (2016) S90–S96. doi:10.1017/S0022215116000505.

[2] S. Elrefaey, M. A. Massaro, S. Chiocca, F. Chiesa, M. Ansarin, HPV in oropharyngeal cancer: The basics to know in clinical practice, Acta Otorhinolaryngologica Italica 34 (2014) 299–309.

[3] A. Schache, J. Croud, M. Robinson, S. Thavaraj, Human papillomavirus testing in head and neck squamous cell carcinoma: Best practice for diagnosis, Methods in Molecular Biology (Clifton, N.J.) 1180 (2014) 237–55. doi:10.1007/978-1-4939-1050-2_13.

[4] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, A. Campilho, Classification of breast cancer histology images using convolutional neural networks, PLOS ONE 12 (2017) e0177544. doi:10.1371/journal.pone.0177544.

[5] S. Fouad, D. Randell, A. Galton, H. Mehanna, G. Landini, Unsupervised morphological segmentation of tissue compartments in histopathological images, PLOS ONE 12 (2017) e0188717. doi:10.1371/journal.pone.0188717.

[6] N. Bayramoglu, J. Kannala, J. Heikkilä, Deep learning for magnification independent breast cancer histopathology image classification, in: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 2440–2445.

[7] N. Dimitriou, O. Arandjelović, P. D. Caie, Deep learning for whole slide image analysis: An overview, Frontiers in Medicine 6 (2019) 264. doi:10.3389/fmed.2019.00264.

[8] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[9] M. N. Kashif, S. E. A. Raza, K. Sirinukunwattana, M. Arif, N. Rajpoot, Handcrafted features with convolutional neural networks for detection of tumor cells in histology images, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), IEEE, 2016, pp. 1029–1032.

[10] H. Wang, A. C. Roa, A. N. Basavanhally, H. L. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, A. Madabhushi, Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features, Journal of Medical Imaging 1 (3) (2014) 034003.

[11] K. Yu, Y. Zhang, C. Huang, R. Liu, T. Li, L. Yang, J. Morris, V. Baladandayuthapani, H. Zhu, Radiomic analysis in prediction of human papilloma virus status, Clinical and Translational Radiation Oncology 7 (2017) 49–54. doi:10.1016/j.ctro.2017.10.001.

[12] M. Bogowicz, O. Riesterer, K. Ikenberg, S. Stieb, H. Moch, G. Studer, M. Guckenberger, S. Tanadini-Lang, CT radiomics predicts HPV status and local tumor control after definitive radiochemotherapy in head and neck squamous cell carcinoma, International Journal of Radiation Oncology Biology Physics 99 (06 2017). doi:10.1016/j.ijrobp.2017.06.002.

[13] N. Fujima, V. C. Andreu-Arasa, S. K. Meibom, G. A. Mercier, M. T. Truong, O. Sakai, Prediction of the human papillomavirus status in patients with oropharyngeal squamous cell carcinoma by fdg-pet imaging dataset using deep learning analysis: A hypothesis-generating study, European Journal of Radiology (2020) 108936.

[14] S. Ranjbar, S. Ning, C. Zwart, C. Wood, S. Weindling, T. Wu, J. Mitchell, J. Li, J. Hoxworth, Computed tomography-based texture analysis to determine human papillomavirus status of oropharyngeal squamous cell carcinoma, Journal of Computer Assisted Tomography 42 (2017) 1. doi:10.1097/RCT.0000000000000682.

[15] M. Guillaud, K. Adler-Storthz, A. Malpica, G. Staerkel, J. Matisic, D. Van Niekirk, D. Cox, N. Poulin, M. Follen, C. MacAulay, Subvisual chromatin changes in cervical epithelium measured by texture image analysis and correlated with hpv, Gynecologic oncology 99 (3) (2005) S16–S23.

[16] S. Fouad, G. Landini, M. Robinson, H. Mehanna, D. A. Randell, Imaging and machine learning methods for assessing HPV *in situ* hybridisation patterns in oropharyngeal carcinomas., in: 14th European Congress on Digital Pathology, 2018.

[17] T.-H. Song, G. Landini, S. Fouad, H. Mehanna, Epithelial segmentation from *in situ* hybridisation histological samples using a deep central attention learning approach, in: IEEE International Symposium on Biomedical Imaging (ISBI 2019), 2019, pp. 1527–1531. doi:10.1109/ISBI.2019.8759384.

[18] M. Robinson, J. James, G. Thomas, N. West, L. Jones, J. Lee, K. Oien, A. Freeman, C. Craig, P. Sloan, P. Elliot, M. Cheang, M. Rodriguez-Justo, C. Verrill, Quality assurance guidance for scoring and reporting for pathologists and laboratories undertaking clinical trial work, The Journal of Pathology: Clinical Research 5 (2019) 91–99. doi:10.1002/cjp2.121.

[19] J. S. J. Lewis, W. L. Thorstad, R. D. Chernock, B. H. Haughey, J. H. Yip, Q. Zhang, S. K. El-Mofty, P16 positive oropharyngeal squamous cell carcinoma: an entity with a favorable prognosis regardless of tumor HPV status, The American Journal of Surgical Pathology 34 (2010) 1088–1096. doi:10.1097/PAS.0b013e3181e84652.

[20] S. G. Craig, L. A. Anderson, A. G. Schache, M. Moran, L. Graham, K. Currie, K. Rooney, M. Robinson, N. S. Upile, R. Brooker, M. Mesri, V. Bingham, S. McQuaid, T. Jones, D. J. McCance, M. Salto-Tellez, S. S. McDade, J. A. James, Recommendations for determining HPV status in patients with oropharyngeal cancers under TNM8 guidelines: a two-tier approach, British Journal of Cancer 120 (2019) 827–833. doi:10.1038/s41416-019-0414-9.

[21] A. D. Singhi, W. H. Westra, Comparison of human papillomavirus in situ hybridization and p16 immunohistochemistry in the detection of human papillomavirus- associated head and neck cancer based on a prospective clinical experience, Cancer 1 (2010) 2166–2173. doi:10.1002/cncr.25033.

[22] T. Kelesidis, L. Aish, M. Steller, I. Aish, J. Shen, P. Foukas, J. Panayiotides, G. Petrikkos, P. Karakitsos, S. Tsiodras, Human papillomavirus (HPV) detection using *in situ* hybridization in histologic samples correlations with cytologic changes and polymerase chain reaction hpv detection, American Journal of Clinical Pathology 136 (2011) 119–27. doi:10.1309/AJCP03HUQYZMWATP.

[23] A. Ruifrok, D. Johnston, Quantification of histochemical staining by color deconvolution, Analytical and Quantitative Cytology and Histology 23 (2001) 291–299.

[24] G. Landini, G. Martinelli, F. Piccinini, Colour Deconvolution – stain unmixing in histological imaging, Bioinformatics (09 2020). doi:10.1093/bioinformatics/btaa847.

[25] P. Sahoo, C. Wilkins, J. Yeager, Threshold selection using Rényi's entropy, Pattern Recognition 1 (1997) 71–84. doi:10.1016/S0031-3203(96)00065-9.

[26] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Trans-

actions on Pattern Analysis and Machine Intelligence 34 (05 2012). doi:10.1109/TPAMI.2012.120.

[27] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[28] G. Landini, Advanced shape analysis with ImageJ, in: Proceedings of the Second ImageJ User and Developer Conference, 2008, pp. 116–121. doi:https://blog.bham.ac.uk/intellimic/g-landini-software/.

[29] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (3) (1995) 273–297.

[30] C. M. Bishop, Pattern recognition and machine learning, springer, 2006.

[31] N. S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, The American Statistician 46 (3) (1992) 175–185.

[32] T. K. Ho, The random subspace method for constructing decision forests, IEEE transactions on pattern analysis and machine intelligence 20 (8) (1998) 832–844.

[33] N. Chawla, K. Bowyer, L. Hall, P. W. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intellingent Research 16 (2002) 321–357. doi:10.1613/jair.953.

[34] A. Cruz-Roa, A. Basavanhally, F. González, H. Gilmore, M. Feldman, S. Ganesan, N. Shih, J. Tomaszewski, A. Madabhushi, Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks, Progress in Biomedical Optics and Imaging - Proceedings of SPIE 9041 (02 2014). doi:10.1117/12.2043872.

[35] K. Nazeri, A. Aminpour, M. Ebrahimi, Two-stage convolutional neural network for breast cancer histology image classification, in: International Conference Image Analysis and Recognition, Springer, 2018, pp. 717–726.

[36] D. Bardou, K. Zhang, S. M. Ahmad, Classification of breast cancer based on histology images using convolutional neural networks, IEEE Access 6 (2018) 24680–24693.

[37] H. Sharma, N. Zerbe, I. Klempert, O. Hellwich, P. Hufnagl, Deep convolutional neural networks for automatic classification of gastric carcinoma using whole slide images in digital histopathology, Computerized Medical Imaging and Graphics 61 (2017) 2–13.

[38] A. L. Maas, A. Y. Hannun, A. Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proc. icml, Vol. 30, 2013, p. 3.

[39] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167 (2015).

[40] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[43] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical image analysis 42 (2017) 60–88.

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale vi-

sual recognition challenge, International journal of computer vision 115 (3) (2015) 211–252.

[45] S. Khan, N. Islam, Z. Jan, I. U. Din, J. J. C. Rodrigues, A novel deep learning based framework for the detection and classification of breast cancer using transfer learning, Pattern Recognition Letters 125 (2019) 1–6.

[46] K. He, R. Girshick, P. Dollár, Rethinking imagenet pre-training, in: Proceedings of the IEEE international conference on computer vision, 2019, pp. 4918–4927.