A pathology-based machine learning method to assist in epithelial dysplasia diagnosis

Karoline da Rocha1 · José C. M. Bermudez1 · Elena R. C. Rivero2 · Márcio H. Costa1

Received: 8 December 2021 / Accepted: 16 July 2022 / Published online: 9 August 2022
© The Author(s), under exclusive licence to The Brazilian Society of Biomedical Engineering 2022

Abstract

Epithelial dysplasia (ED) is one of the most important factors in detecting the progression of an oral tissue alteration towards carcinoma. Its early detection is instrumental in avoiding the progression of the tumor. A major difficulty for detecting ED is the recognized variability of pathologist assessments. This study proposes a new method that leverages the pathologist expertise to design a simple and efficient classification system to support the detection of dysplastic epithelia. We employ a multilayer artificial neural network (MLP-ANN) and defining the regions of the epithelium to be assessed based on the knowledge of the pathologist. The performance of the proposed solution was statistically evaluated. The implemented MLP-ANN presented an average accuracy of 87%, with a variability much inferior to that obtained from three trained evaluators. Moreover, the proposed solution led to results which are very close to those obtained using a convolutional neural network (CNN) implemented by transfer learning, with 100 times less computational complexity. In conclusion, our results show that a simple neural network structure combined with the pathologist expertise can lead to a performance equivalent to that of much more complex structures, which are routinely used in the literature.

Keywords Epithelial Dysplasia · Nonlinear classifier · Neural networks · Computer-aided diagnosis · Assistive diagnosis

Introduction

The oral cavity Squamous Cell Carcinoma (SCC) is one of the most common cancers among men (Alabi et al. 2021). Survival rates for this cancer are low (Nag and Das 2018; El-Naggar et al. 2017). Thus, early diagnosis is essential for survival and for determining the most appropriate treatment. Epithelial Dysplasia (ED) is a very frequent tissue alteration in lesions that precede oral cancer. Its cause may be associated with etiological factors, mainly with the consumption of tabacco and alcohol (Mello et al. 2018).

The standard ED diagnosis is made by pathologists through analysis of histopathological slides in search of changes in the epithelium. These alterations may be architectural, such as basilar hyperplasia, droplet-shaped epithelial projections, increased number of mitoses or loss of cellular cohesion. They may also be cellular changes, such as enlarged nuclei, enlarged cells, increased nuclear-cytoplasmic ratio, or atypical mitotic figures (Ranganathan and Kavitha 2019).

The most widely employed criteria for grading oral ED are those defined in the World Health Organization (WHO) classification system, which considers the presence of certain architectural and cytological features. The more prominent and numerous these features are identified in the histopathological image, the more severe the grade of dysplasia (Warnakulasuriya et al. 2008; El-Naggar et al. 2017). However, the assessment of dysplasia is subjective and strongly dependent on the personal experience of the pathologist (Bubacar et al. 2021), and the diagnosis variability is well documented (Bouquot et al. 2006; Kujan et al. 2006; Warnakulasuriya et al. 2008; Mahmood et al. 2021).

The area of computer-aided diagnosis has been growing over the past decades (Talo 2019; Arevalo et al. 2015). Histopathological images are analyzed using quantitative measures of image characteristics. These measures are employed either by image processing algorithms or interpreted by machine learning systems to yield a suggested diagnosis. Such automated diagnostic tools can be an important aid
to pathologists, enabling the creation of large-scale decision support systems to identify potentially malignant lesions (Alabi et al. 2021; Pallua et al. 2020; Mahmood et al. 2021). Some recent works have proposed automated diagnostic solutions to assist in the evaluation of histopathological images for the detection of epithelial dysplasia (Krishnan et al. 2012; Venkatakrishnan et al. 2017; Almubarak et al. 2017; Gupta et al. 2019; Adel et al. 2019; Sami et al. 2009; Prabavathy et al. 2021; Baik et al. 2014; Krishnan et al. 2009).

In a previous work, texture characteristics of the epithelium have been used in Krishnan et al. (2012) to classify histopathological images into normal epithelia and oral submucous fibrosis (OSF) with or without dysplasia. In this article, textural characteristics of the epithelium were extracted using higher order spectra (HOS), local binary pattern (LBP), and laws texture energy (LTE). Five different classifiers were employed: Decision Tree (DT), Sugeno Fuzzy, Gaussian Mixture Model (GMM), K-Nearest Neighbor (K-NN), and Radial Basis Probabilistic Neural Network (RBPN). The results of this work indicate that combination of texture and HOS features coupled with a fuzzy classifier resulted in 95.7% accuracy. However, it is noted in the article that images need considerable processing to be used in the proposed classification method. In addition, the training and testing (evaluation) of the classifiers was performed using stratified 3-fold cross-validation, thus separating the data set into only three subsets, which may favor the occurrence of strong correlation between the training set and the test set. This evaluation approach leaves the performance results obtained (95.7% accuracy) open to questioning.

The Block Intensity Code Comparison (BICC) was used in Venkatakrishnan et al. (2017) to extract the characteristics of the epithelium. Each image was divided into blocks and the intensity of each block was calculated. Blocks of sizes 5 × 5, 10 × 10 and 15 × 15 have been tested. The classifier used was a Radial Basis Function Neural Network (RBFNN) with a single hidden layer, and Gaussian activation function, classifying epithelium into normal or dysplastic. Different values were tested for the centers (means) of the Gaussians. These centers were located using k-Means, and k = 6 led to the best result. The article fails to clearly explain how the final classification was made.

A convolutional neural network (CNN) has also been used in Almubarak et al. (2017) to extract characteristics of histological images of the uterine cervix epithelium. The authors classify the epithelium in four classes: normal, mild dysplasia, moderate dysplasia and severe dysplasia. Each of the 66 images available was subdivided into 10 vertical segments. The images had to be correctly positioned so that each segment had the three layers of the epithelium. Each of these segments was subdivided into three parts of the same size (top, middle, bottom). Finally, 32 × 32 pixel patches were extracted from each part of each segment using non-overlapping windows, which generated a total of 75,000 samples of size 32 × 32. Three CNNs were implemented, each of which classified the samples of each of the 3 parts into one of the 4 classes. These networks extract the characteristics of the samples and provide an initial classification. The authors did not inform the amount of samples used to train and test the networks. The extracted characteristics fed 5 different classifiers (SVM, LDA, MLP, logistic regression and random forest), which merged the characteristics of the 3 parts, obtained from the CNNs, and classified the entire epithelium image. The best performance (77.25% accuracy) was obtained with logistic regression and random forest, and using leave-one-out cross-validation performed only once to classify the 66 images.

Also using CNN, Gupta et al. (2019) classified epithelial images into 4 classes (normal, mild dysplasia, moderate dysplasia or severe dysplasia). The CNN received epithelial images as inputs. The classifier was trained using off-the-shelf packages. No details were given regarding the characteristics of the network nor the training algorithm. There were 2688 images taken from 672 tissue images of 52 patients, with approximately the same amount of images belonging to each of the four classes. The CNN training was performed from scratch using 70% of the available samples. The remaining 30% were used for testing. The CNN trained over 75 epochs presented an accuracy of 89.3% for the test set.

Classification based on extracted values of the 16 WHO defined features was proposed in Adel et al. (2019). No detail was provided on how the extraction was performed. An SVM and a K-nearest-neighbor classifiers were tested using the different feature sets. The reported results were based on a single classification run using 46 images, from which 32 were used for training an 14 for test in the SVM implementation. The accuracy results varied from 71.4% to 78.6% in most implementations, and reached 92.8% only for the SVM classifier operating on the features extracted by the Oriented FAST and Rotated BRIEF (ORB) algorithm.

Classification based on the similarity of neighboring rete ridges was proposed in Sami et al. (2009). Epithelium was classified as normal, dysplastic or carcinoma in-situ (CIS). The method was based on comparing the drop-shaped similarity level between the best matching pair of neighboring rete ridges. A contour extraction method was proposed, and the roundness of extracted twin contours was quantified. Clustering of the three classes was based on the roundness absolute values and differences. Method illustration was based on a set of 17 images. No statistical evaluation was presented regarding the accuracy of the proposed classification method.

In a recent publication, two characteristic extraction techniques were employed (Prabavathy et al. 2021), namely...
histogram oriented gradient (HOG) and local binary pattern (LBP), were applied to discrete wavelet transformed 512 × 512 epithelial images. A three-layer back propagation neural network (BPNN) classified the images into normal or dysplastic epithelium. The best results were obtained using the HOG features, yielding an accuracy of 85%. The article does not detail the methodology used for the classification.

A semi-automatic algorithm has been also proposed in Baik et al. (2014) to predict the progression of oral premalignant lesions (OPL) to invasive squamous cell carcinoma (SCC). The authors initially use two Random Forests, one to segment the nuclei of the histopathological images and the other to classify the nuclei found into normal and abnormal (cancerous). After the identification of the nuclei, a Nucleus Phenotype Score (NPS) was calculated based on the voting score that each nucleus received from the Random Forests classifier. Based on the average NPS of all image cores, an automated Tissue Nuclear Phenotype Score (aNPS) was assigned in order to identify OPL with high risk of progression. During the prediction of the progression of 71 lesions in the test set, a 78% sensitivity and a 71% specificity were obtained.

Segmentation and classification of sub-epithelial connective tissue cells into normal or with oral submucous fibrosis (OSF) was proposed in Krishnan et al. (2009). Limiarization with several levels was used in the segmentation. An SVM-based classifier was used to classify the cells as per their geometric shape features (eccentricity and compactness). The classifier presented an accuracy of 88.69% for the test set.

One common problem to these previous proposals is that they rely almost entirely on the information provided by the collected data. Despite the large popularity of machine learning (ML) algorithms, it is known that they tend to be data inefficient, and frequently generalize poorly to unseen cases. This characteristic is especially present in medical applications where the amount of available data is usually limited. Most ML algorithms that process raw data lead to classifications based on the correlation structure of the data presented to them during training. The success of this approach depends on a huge amount of data. Moreover, the diagnosis quality is also frequently dependent on factors independent of the data correlation, such as the amount and type of noise present in the data, the quality of data acquisition, and the amount of useful information embedded in the data. For instance, some of the presented solutions proposed to extract the characteristics of the epithelium by analyzing the entire image, which increases the possibility of using information that may hamper the classification process. This also tends to lead to a lack of consensus on the most relevant characteristics for classification. Another consideration usually made is that a binary classification system tends to be more helpful to the clinician for making critical clinical decisions in cases of high-risk epithelial dysplasia than the WHO three-level classification (Kujan et al. 2006; Bubacar et al. 2021).

A more sensible approach is to complement the information provided by the raw data with the knowledge of experts in the application field. This tends to lead to more robust performances using simpler algorithms and less data. This work is a contribution in such a direction. We combine the knowledge of the pathologist with the information provided by the data to generate a method of detecting oral dysplasia which performs on a smaller amount of data, and is more robust to changes in the statistical characteristics of the data. We use the expert’s knowledge to define the information to be delivered to an ML algorithm that aims at classifying histopathological images into dysplastic and non-dysplastic epithelia. The proposed approach results in a simple classifier design process, which does not require complex pre-processing procedures, and leads to much lower computational complexity solutions with performance comparable with that obtained using the typical image classification solutions employing convolutional neural networks.

**Methods**

In this section, we describe the database and the methods employed in this work.

**Database and sample preparation**

The database consists of histological images obtained from the Biobank Archives of the Oral Pathology Laboratory of UFSC (CONEP B-051; Process No. 25000.237810/2014-54). The research project was approved by the UFSC Human Beings Research Ethics Committee (Platform Brasil under number CAAE 15025319.3.0000.0121).

The images used were from histological slides stained with hematoxylin and eosin. These slides were photographed with an original magnification of 200x, using a digital camera coupled to an optical microscope. After being scanned, the images were duly anonymized to be used in this research.

The data set corresponded to 73 cases, totaling 172 images, each one with 2584 × 1936 pixels. Of these, 36 cases (88 images) were from oral potentially malignant disorders with dysplastic epithelia and 37 cases (84 images) were from fibrous hyperplasia with non-dysplastic epithelia.

As evidenced in the Introduction, there seems to be no consensus yet regarding a good set of automatically extracted image characteristics to be used for classification of dysplastic epithelia. In this work, we propose to use the pathologist expertise to define regions of the image to be presented to the ANN for classification purposes.
To take advantage of the existing expertise, we have chosen to extract cutouts from the image region containing the main visual characteristics used by the pathologist for diagnosing dysplastic epithelia. The cutouts were extracted from regions located close to the border that separates the epithelial tissue from the connective tissue. The red squares shown in Fig. 1 illustrate seven of the proposed cutouts. The reasoning for this choice was that the lower third of a dysplastic epithelium image will always present detectable changes, regardless of the dysplasia degree (mild, moderate or severe). Through this informed choice, only small image regions containing most of the necessary information can be fed to the ANN. This tends to simplify the processing without harming (most likely improving) the classification potential. The number of cutouts extracted from each image was not fixed, depending on the size of the epithelium area available in each image. The number of cutouts extracted from an image varied between 7 and 15.

The definitions of location and size of each cutout were influenced by a combination of factors:

a) The use of small regions simplifies the data processing.
b) The cutout image should contain enough information for a classification by the pathologist.
c) All cutouts must have the same size, leading to a fixed number of neural network input nodes.
d) The cutouts of a given image should not overlap to avoid the use of redundant information.
e) A basic classifier should yield a reasonably low classification error using the cutouts.

Based on these criteria, we propose to use non-overlapping cutouts at the frontier between the epithelial and connective tissues, with most of their area in the epithelial region, as shown in Fig. 1. The choice of cutout dimensions is discussed in section “Definition of the classifier input”.

For processing by the classifier, each cutout image was converted from RGB to grayscale using the Matlab function rgb2gray(). This reduced by 2/3 the amount of data to be processed. Then, the cutout image values were normalized to increase robustness to intensity variations due to lighting effects in different acquisitions. Pixel intensities of each cutout were normalized to be in the range [0, 1] according to

\[ R_{\text{norm}} = \frac{R}{\max\{r_{ij}\}} \]  

(1)

where \( R \) is the cutout matrix in grayscale and \( \max\{r_{ij}\} \) is the maximum value of all the elements \( r_{ij} \) of \( R \).

Finally, each cutout used in the classifier training stage was rotated three times by 90°, yielding four images for each cutout. Rotations by non-integer multiples of 90° were not considered to avoid the need for further processing at the corners of the rotated cutouts. The use of four rotated positions was considered sufficient to provide the necessary robustness to image rotation.
Methodologies

This section describes the methods employed in the different steps of the classification task. The project was implemented in Matlab© (The MathWorks, Inc. Software) using a personal computer with Windows 10 operating system, a 1.8 GHz Intel Core i5 processor and 8 GBytes of RAM. The neural networks were implemented as Multilayer Neural Networks (MLP-ANN) using the “patternnet” function.

Definition of the classifier input

The classifiers were implemented using feed-forward multilayer neural networks in which each input layer neuron was fed with one pixel of the image cutout. To define the structures of these networks it was necessary to choose, besides the size of the input cutouts, the number of hidden layers, the number of neurons in each hidden layer, the cost function to be used, and other necessary parameters depending on the previous choices. Different structures were considered and evaluated in the training stage. This stage was subdivided into two phases.

The objective of the first phase was to determine the size of the input cutouts so that enough information is provided to the ANN for classification purposes. Visual inspection with the help of the pathologist (third author) determined a minimum dimension of 128 × 128 pixels for the cutouts. Hence, we decided to test cutouts with dimensions 128 × 128 and 256 × 256, the latter being considered a reasonable trade-off between quality and implementation complexity. To this end, in this first design phase we have implemented classifiers using very simple networks with only one hidden layer.

Considering cutout dimensions of 128 × 128 pixels and 256 × 256 pixels, the networks have been tested, respectively, with 16, 384 and 65,536 input neurons (one neuron for each cutout pixel). We first extracted cutouts of dimension 256 × 256 from the original images. The 128 × 128 cutouts were obtained from the 256 × 256 cutouts through resizing using Matlab function imresize(), which employs cubic interpolation. In addition, the number of neurons in the hidden layer was varied in the set {20, 50, 100}. The cost function was the Mean Squared Error (MSE), given that the data had no outliers and the objective was an initial evaluation of the classification performance as a function of cutout size. Each classifier performance was evaluated based on the classification of each cutout individually, independent of the case to which they belonged. A total number of 1840 cutouts were obtained from the 172 images. This number was the same for both cutout sizes tested. For the training, 1440 cutouts (720 with dysplasia and 720 without dysplasia) were randomly selected from the 1840 cutouts available. For testing, 160 randomly selected cutouts had been previously separated (being 80 with and 80 without dysplasia). A total of 100 classifications, comprising training and test, were carried out to evaluate the classifier performance. For each classification, 50 cutouts were randomly selected from each of the sets of 80 cutouts (with and without dysplasia). The results of the 100 classifications were averaged to determine the confusion matrix, specificity, sensitivity and average accuracy rate of the classifications. Based on these results, it was verified that the structures with 65,536 (256 × 256 cutouts) input neurons yielded better performance on the average. Thus, all the structures implemented in the second phase considered 256 × 256 cutouts, and thus had 65,536 neurons in the input layer.

The second phase of the design will be described next.

Definition of the classifier structure

The second design phase aimed at improving the performance of the classifier, given an input layer with 65,536 neurons. We have initially used networks with two hidden layers. The number of neurons per layer was varied within the set {20, 50, 100, 150}. The cost functions tested were MSE and the cross-entropy. Another difference from the first phase was that the data unit considered was cases rather than individual cutouts. As we had data for 42 cases, at each run the classifier was trained with the cutouts corresponding to 41 cases, and tested with the cutouts of the remaining case, thus using leave-one-out cross-validation. As the number of cutouts available was not exactly the same for all cases, care was taken to maintain the same number of dysplastic and non-dysplastic cutouts during training. Finally, recall that each cutout used to train the classifiers were rotated 3 times by 90 degrees for greater robustness to rotation and for increasing the amount of training data.

The classification decision (dysplastic or non-dysplastic) was made by majority of the cutout classifications for each case. For classifier performance evaluation, the confusion matrices, sensitivities, specificities, accuracies, and a novel figure of merit D to be defined later were evaluated for each of the cases after test. For statistical evaluation purposes, the complete process described above was repeated 50 times (50 runs) with randomly selected network initialization. Then, the confusion matrices, sensitivities, specificities, accuracies, and the values of the figure of merit D reported correspond to averages over 50 runs for each of the 42 cases.

After evaluating the performance of the structures with two hidden layers, those which yielded better accuracy results had their number of hidden layers increased to verify if this increase would lead to an improvement of the average accuracy for the 42 cases under test. For each structure, the number of hidden layers was incremented until a drop in
classification performance was verified. The training and test were performed exactly as done for the two-layer structures.

In both design phases, the classifiers were trained using the scaled conjugate gradient algorithm. The stop criterion was early stopping, in which the training was stopped when any of the following conditions occurred:

a) The maximum number of 1000 iterations was reached.
b) The error was equal to zero.
c) The gradient was less than or equal to $10^{-6}$.
d) The validation set error increased in 6 consecutive epochs.

For each training performed, the cutouts in the training set were randomly subdivided into three subsets: 70% for training, 15% for validation and 15% for testing. The test subset was not used at this stage.

The structures of the trained networks used the SoftMax activation function for the output layer and the hyperbolic tangent function (Tanh) for the hidden layers.

**Decision rule**

All neural networks were designed to classify individual cutouts, and not the epithelium image as a whole. To classify a given case (one given patient), we first classify each of the cutouts corresponding to that case, assigning each cutout either to class $\omega_1$ (dysplastic) or to class $\omega_2$ (non-dysplastic). The final diagnosis for the case (patient) is determined by the majority of classifications obtained for all the cutouts for that patient.

Each cutout classification is done using the following Bayes risk rule:

Assign $x$ to $\omega_1$ ($\omega_2$) if: $\lambda_{21}P(\omega_1|x) > (<)\lambda_{12}P(\omega_2|x)$

in which $x$ is the input vector (cutout image); $P(\omega_i|x)$ is the posterior conditional probability (as estimated by the neural network) that $x$ is from class $\omega_i$, given the observed cutout image $x$. The loss parameters $\lambda_{ij}$ are defined as the loss associated to assigning a cutout to class $\omega_i$ when it actually belongs to class $\omega_j$.

Two different situations were considered for the choice of loss values $\lambda_{ij}$ in (2). The first one employs $\lambda_{12} = \lambda_{21} = 1$, which applies the same weight to the loss in having a false negative (erroneously classifying a cutout as non-dysplastic) or a false positive (erroneously classifying a cutout as dysplastic). The second situation considers the loss in having a false negative to be twice as large as the loss associated to a false positive. In this case, $\lambda_{12} = 1$ and $\lambda_{21} = 2$. This latter consideration (maybe with different relative weights) is common in medical applications.

From (2), the decision rule actually implemented is then “classify $x$ in class $\omega_1$ if”

$$
P(\omega_1|x) \geq \frac{\lambda_{12}}{\lambda_{21}}$$

Otherwise, classify $x$ in class $\omega_2$.

**Performance metrics**

Different metrics were used to evaluate the performance of the classifiers. The main metric was the confusion matrix, which reports the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Using these results we evaluated the sensitivity ($S_e$) and the specificity ($S_p$) of the classifier, given by, respectively

$$
S_e = \frac{TP}{TP + FN}
$$

$$
S_p = \frac{TN}{TN + FP}
$$

Another metric employed was the accuracy, evaluated as

$$
A_{cc} = \frac{TP + TN}{N}
$$

where $N$ stands for the total number of data classified.

Another way to evaluate the performance of binary classifiers is through ROC (Receiver Operating Characteristics) curves. This curve plots the True Positive Rate (TPR) ($S_e$, Eq. (4)) against False Positive Rate (FPR) ($1 - S_p$, Eq. (5)) for different classification thresholds. The performance of the classifier is measured by the Area Under the ROC Curve (AUC). An ideal ROC curve will touch the top left corner of the graph, so the higher the value of the AUC the better the classifier (James et al. 2013; Fawcett 2006). An AUC of 1 would indicate that the classifier made all predictions correctly, resulting in a sensitivity and specificity of 100%.

Finally, we propose in this study a new figure of merit $D$ to evaluate the deviation from an ideal classifier performance. Figure 2 shows a graphical interpretation of the new figure of merit $D$. The horizontal axis shows the Positive Predictive Value (PPV = TP/total number of positives), and the vertical axis shows the Negative Predictive Value (NPV = TN/total number of negatives) in the training set. Vector $R_{ef}$ represents the composite relative accuracy. In the ideal classifier (gold standard), $R_{ef} = R_{ef_0}$ has magnitude $\sqrt{2}$ and 45° phase. Deviations from this ideal situation represent loss in classification performance.

To this end, we define $D$ as follows:
In (7)–(10), $D_1$ corresponds to the magnitude of vector $R_{ef}$, and $D_2$ corresponds to the angular deviation from the ideal $45^\circ$ reference. Hence, $0 \leq D \leq 1$, and the larger the value of $D$ the better the classifier performance. One particularly interesting aspect of the performance criterion $D$ is that it provides an objective way to rank the tested network structures for relative performance. Criterion $D$ is used in the next design phase to estimate the best trained networks.

**Definition of the best designs**

At this design stage we had already defined the input dimension (phase 1) and a set of network structures leading to the best performance results, as measured by the average accuracy obtained in design phase 2. The next stage was a statistical performance evaluation, in which we used 100 runs of training and test for each structure defined in phase 2 with random initialization in search for the best training for each of them. In each run, the networks were trained with all the cutouts of the 42 cases used in the first and second phases. The tests were performed using cutouts of 31 new cases that had not been used in any previous training or test. These 31 cases provided a total of 826 cutouts. The difference in the performances obtained at each of the 100 runs was due to a different initialization of the network coefficients. For each of the network structures, the coefficients resulting from the training cycle that yielded the highest value of criterion $D$ (out of the 100 runs) for the test set were stored. After these steps, the trained networks, with fixed structures and their optimal coefficients, were used for the classifications, and the results were compared with other classifiers.

**Performance comparisons**

We had no access to implementations of any of the previous works, and most of them were not described with enough detail to be faithfully implemented. Hence, the performances of the networks trained during the classification of the 31 test cases were compared with the performances in these same cases provided by a convolutional neural network pre-trained by learning transfer (one of the most popular solutions for image classification), and by trained evaluators (oral pathologists).

The pre-trained CNN used in our comparison was the Resnet-18 network (Mathworks 2019). Some adaptation of the original code was necessary in order to replace the fully connected layer and the sorting layer so that the resulting network output had only two classes as required by the application. In addition, all cutouts were modified to the RGB scale, normalized, and had their sizes readjusted from $256 \times 256$ to $224 \times 224$ pixels, to be compatible with the input dimensions of the pre-trained network.

The main CNN training options used were the following:

a) Convolutional layer learning rate equal to $10^{-4}$. It was chosen very low so that the filter coefficients and the previously trained weights were not lost.

b) Mini-batch size equal to 10.

c) Fully connected layer learning rate equal to 10. It was chosen high so that the learning for these layers was faster than for the convolutional layers.

d) Maximum number of epochs equal to 8.

The training and performance evaluation stages of the CNN network were evaluated using the same methodologies described in sections “Definition of the classifier structure” and “Definition of the best designs”.

Finally, to evaluate the usefulness of a classifier in this case, we compared the performances of the best neural classifiers designed with the classifications provided by three trained

---

**Fig. 2** Two-dimensional descriptive space of figure of merit $D$. In red: the best classification scenario. In blue: Example of classifier performance

$$D = \frac{|D_1| + |D_2|}{2}$$

(7)

$$D_1 = \frac{\sqrt{PPV^2 + NPV^2}}{\sqrt{2}}$$

(8)

$$D_2 = \cos(45^\circ - \theta)$$

(9)

$$\theta = \arctan\left(\frac{NPV}{PPV}\right)$$

(10)
evaluators. To this end, we asked the trained evaluators to evaluate the original images of each of the 31 cases used as the test set. The evaluation was made using the whole images (not the cutouts), as would be the case in a real diagnosis situation.

**Results**

In this section we present the performances obtained from the networks designed in the second phase of the project. As mentioned in section “Definition of the classifier input”, it was determined that the structures with 65,536 neurons in the input layer had the best performances on average. Hence, all the structures implemented in the second phase had this same input layer size.

**Definition of the best classifier structures**

Structures with 20, 50, 100, and 150 neurons per hidden layer, and the cost functions mean-square error (MSE) and cross-entropy (CE) were designed. Table 1 shows the average classification performances (50 runs) obtained for the 42 cases available for \( \lambda_{12} = \lambda_{21} = 1 \) in Eq. (3). Each run was characterized by a random initialization of the neural network weights. The majority of structures presented average accuracy higher than 0.80 with small standard deviations, what is a good indicator of design reliability. Structure 8 presented the highest average accuracy of 0.8205, and its best result reached a 0.8810 accuracy. The results demonstrate that all structures presented high averages for the proposed figure of merit D (Eq. (7)). However, the best performance was that of structure 8.

The same eight structures in Table 1 were trained using \( \lambda_{12} = 1 \) and \( \lambda_{21} = 2 \) in (3) to test the design performance with a conservative risk function. It was observed that the average accuracy and mean values of D did not present a significant increase when compared to the values in Table 1. Figures 3 and 4 show, respectively, comparisons of the sensitivity and specificity values obtained using the two different risk functions. As expected, the sensitivity increased and the specificity reduced when \( \lambda_{21} = 2 \) was used. This is because, in this case, a greater importance was given to the occurrence of false negatives (erroneous classification of dysplastic epithelia). As a consequence, a greater number of cases were classified by the network as dysplastic, increasing the number of true positives and the average sensitivity.

**Table 1** Results — second design phase for \( \lambda_{12} = \lambda_{21} = 1 \)

| Structures | Accuracy ±SD | Max. value | Figure of merit |
|------------|--------------|------------|----------------|
| (NHL\(^1\)/NNPHL\(^2\))/Cost function | Average  | Max.  | D(%)       |
| 1 (2/20/MSE) | 0.7881±0.0334 | 0.8571 | 89.25     |
| 2 (2/20/CE\(^6\)) | 0.8019±0.0386 | 0.8810 | 89.93     |
| 3 (2/50/MSE) | 0.8014±0.0421 | 0.9048 | 89.94     |
| 4 (2/50/CE\(^6\)) | 0.8143±0.0340 | 0.8810 | 90.58     |
| 5 (2/100/MSE) | 0.8105±0.0391 | 0.8810 | 90.36     |
| 6 (2/100/CE\(^6\)) | 0.8086±0.0376 | 0.8810 | 90.26     |
| 7 (2/150/MSE) | 0.8043±0.0507 | 0.9048 | 90.05     |
| 8 (2/150/CE\(^6\)) | 0.8205±0.0382 | 0.8810 | 90.88     |

\(^1\)NHL, number of hidden layers
\(^2\)NNPHL, number of neurons per hidden layer
\(^3\)SD, standard deviation
\(^4\)Max. value, maximum value
\(^5\)%Percentage of maximum possible value in (7)
\(^6\)CE, cross-entropy

**Fig. 3** Average sensitivities of network structures with two hidden layers. Black: \( \lambda_{12} = \lambda_{21} = 1 \). Gray: \( \lambda_{12} = 1 \), \( \lambda_{21} = 2 \).
On the other hand, the number of false positives increased, decreasing the average specificity.

For the next design step we selected the three structures in Table 1 which yielded the largest average accuracies, namely structures 4, 5, and 8. For these three structures we proceeded to increase the number of hidden layers while such increase resulted in clear possibility of performance improvement.

Table 2 Performance of structure 4 with 2 and 3 hidden layers (\(\lambda_{12} = \lambda_{21} = 1\))

| Structures (NHL/NNPHL) | Accuracy Average±SD | Max. (%) | \(S_e\) \(^1\) | \(S_p\) \(^2\) | Figure |
|------------------------|---------------------|---------|-----|-----|--------|
| Cost function          | value               |         |     |     |        |
| 4 (2/50/CE)            | 0.8143±0.0340       | 0.8810  | 76.70 | 85.73 | 90.58  |
| 9 (3/50/CE)            | 0.8105±0.0363       | 0.9048  | 76.10 | 85.55 | 90.38  |

\(^1\) \(S_e\), sensitivity
\(^2\) \(S_p\), specificity

Table 3 Performance of structure 5 with 2–5 hidden layers (\(\lambda_{12} = \lambda_{21} = 1\))

| Structures (NHL/NNPHL) | Accuracy Average±SD | Max. (%) | \(S_e\) | \(S_p\) | Figure |
|------------------------|---------------------|---------|-----|-----|--------|
| Cost function          | value               |         |     |     |        |
| 5 (2/100/MSE)          | 0.8105±0.0391       | 0.8810  | 75.60 | 86.00 | 90.36  |
| 10 (3/100/MSE)         | 0.8100±0.0515       | 0.8810  | 77.20 | 84.45 | 90.39  |
| 11 (4/100/MSE)         | 0.8195±0.0411       | 0.9048  | 78.80 | 84.82 | 90.88  |
| 12 (5/100/MSE)         | 0.8071±0.0394       | 0.8810  | 77.20 | 83.91 | 90.25  |

Table 4 Performance of structure 8 with 2–4 hidden layers (\(\lambda_{12} = \lambda_{21} = 1\))

| Structures (NHL/NNPHL) | Accuracy Average±SD | Max. (%) | \(S_e\) | \(S_p\) | Figure |
|------------------------|---------------------|---------|-----|-----|--------|
| Cost function          | value               |         |     |     |        |
| 8 (2/150/CE)           | 0.8205±0.0382       | 0.8810  | 77.10 | 86.55 | 90.88  |
| 13 (3/150/CE)          | 0.8271±0.0343       | 0.8571  | 78.80 | 86.27 | 91.25  |
| 14 (4/150/CE)          | 0.8110±0.0313       | 0.8571  | 75.40 | 86.27 | 90.38  |

On the other hand, the number of false positives increased, decreasing the average specificity.

For the next design step we selected the three structures in Table 1 which yielded the largest average accuracies, namely structures 4, 5, and 8. For these three structures we proceeded to increase the number of hidden layers while such increase resulted in clear possibility of performance improvement.

Table 2 shows that the performance of structure 4 degraded when a third hidden layer was added. Table 3 shows the performance of structure 5. It is noted that a performance increase was obtained with 4 hidden layers (structure 11). Table 4 shows the performance of structure 8. In this case, an improved performance was verified as the number of hidden layers increased to 3 (structure 13). Comparing the results for structures 4, 11 and 13, structure 13 was the best one, with an average accuracy of 0.8271, and an average figure of merit D above 91%. Given their comparable performances, we considered these three structures (4, 11, and 13) in the following design and comparison stages.

The performances of structures 4, 5 and 8 obtained with the increase of the hidden layers for \(\lambda_{12} = 1\) and \(\lambda_{21} = 2\) were also verified. The relative performances were very similar to those shown in Figs. 3 and 4, namely, the average sensitivities increased while the average specificities reduced. The average accuracies and figure of merit D had no significant changes when compared with the results obtained from the classifications using \(\lambda_{12} = \lambda_{21} = 1\) (Tables 2, 3 and 4). Hence, the results are presented only for the latter case.
Performance evaluation stage

The figure of merit $D$ was computed for the test cases using (7) after each training. The trained networks with each structure yielding the highest value of $D$ had their coefficients stored to be used in future comparisons. Table 5 shows the performance of the best design for each of the three selected structures.

Table 6 shows the average performances of all four classifiers in the training stage. The pre-trained CNN yielded a slightly better average accuracy and the best performance when compared to the MLP networks. However, its accuracy had the largest standard deviation (about 10% higher than structure 11, and 30% higher than the best MLP structure 13).

Table 7 shows the obtained classification results for the test set. The best performance was obtained from the CNN, which led to an accuracy only 3.7% better than the result provided by structures 11 and 13.

Comparison with trained evaluators

Table 8 compares the performances of the neural classifiers and the three evaluators in the classification of the 31 cases of the test set. This table shows that neural classifiers generated an accuracy equal to or even higher than the three trained evaluators. The three evaluators correctly classified all dysplastic cases (sensitivities were 100%). However, they presented specificities equal or inferior to the neural classifiers.

Discussion

The diagnosis of epithelial dysplasia is provided by oral pathologists based on visual analysis of histopathological images. However, this diagnosis procedure tends to be highly subjective and dependent on the professional experience of the pathologist. In this work, a diagnostic aid system was proposed to help to reduce this subjectivity and the variability in diagnoses. The objective was to develop an epithelial classifier that leverages the expertise of pathologists to avoid excessive complexity without compromising quality. A methodology was proposed for the design of a simple MLP neural network classifier. Two of the resulting classifiers (11 and 13 in Table 5) yielded an average accuracy of 87% in the performance assessment stage.

It has been recognized in the technical literature that Convolutional Neural Networks (CNNs) tend to yield excellent performance in image classification, especially in the medical field (Almubarak et al. 2017). Hence, CNNs qualify as the best candidates for performance comparison with the oral lesion classifiers designed using the proposed approach. However, training a CNN from scratch requires a huge amount of data, in the

---

**Table 5** Performance of the three best structures ($\lambda_{12} = \lambda_{21} = 1$)

| Structures (NHL/NNPHL/ Cost function) | Accuracy | $S_e$ (%) | $S_p$ (%) | Figure $D$ (%) |
|----------------------------------------|----------|-----------|-----------|---------------|
| 4 (2/50/CE)                            | 0.8065   | 81.25     | 80        | 90.31         |
| 11 (4/100/MSE)                         | 0.8710   | 81.25     | 93.33     | 93.63         |
| 13 (3/150/CE)                          | 0.8710   | 81.25     | 93.33     | 93.63         |

**Table 6** Performance comparison with pre-trained CNN — training

| Structures (NHL/NNPHL/ Cost function) | Accuracy | $S_e$ (%) | $S_p$ (%) | Figure $D$ (%) |
|----------------------------------------|----------|-----------|-----------|---------------|
| 4 (2/50/CE)                            | 0.8143±0.0340 | 80.81    | 85.73     | 90.58         |
| 11 (4/100/MSE)                         | 0.8195±0.0411 | 84.82    | 84.82     | 90.88         |
| 13 (3/150/CE)                          | 0.8271±0.0343 | 86.27    | 91.25     |               |
| Pre-trained CNN                        | 0.8505±0.0449 | 86.64    | 92.47     |               |

**Table 7** Performance comparison with pre-trained CNN — test

| Structures (NHL/NNPHL/ Cost function) | Accuracy | $S_e$ (%) | $S_p$ (%) | Figure $D$ (%) |
|----------------------------------------|----------|-----------|-----------|---------------|
| 4 (2/50/CE)                            | 0.8065   | 81.25     | 80        | 90.31         |
| 11 (4/100/MSE)                         | 0.8710   | 81.25     | 93.33     | 93.63         |
| 13 (3/150/CE)                          | 0.8710   | 81.25     | 93.33     | 93.63         |
| Pre-trained CNN                        | 0.9032   | 93.75     | 86.67     | 95.10         |
order of millions of samples. Such amount of data is hardly available in the medical area, and such a solution (if ever feasible) would result in a time-consuming and very costly training process. A typical solution to circumvent this limitation is to apply transfer learning. In this method, a fine-tuning is done in a pre-trained network, where the weights of the initial layers are frozen to values obtained from training with general images, and only the last layers are trained with the application-specific data (Zhang et al. 2017). This was the possible solution in our case, due to the limited amount of data available (a typical situation for the application at hand).

The performances of the best classifier designs were compared with that of a pre-trained CNN, the state of the art in image classifiers for medical applications. The results (Table 7) show an accuracy of more than 87%, only 3.7% smaller than that obtained from the CNN solution. The sensitivity of the CNN was 12.5% better (93.75% versus 81.25%), but at the cost of a 7% worse specificity (86.67% versus 93.33% of the proposed solution). The difference in the value of the newly proposed figure of merit was only 1.6% in favor of the CNN solution (95.1% versus 93.63%). We have also constructed the ROCs for each classifier so that the corresponding areas under the curve (AUC) could be compared. It is noted that structure 11 presented an AUC of 0.8854, which is 7.6% smaller than that of CNN. This result, however, cannot be considered very precise given the step-like ROCs due to the small number of cases available. It should be noted that this good performance of the proposed solution, which is quite comparable to that of the pre-trained CNN, comes at a significant advantage in complexity.

Table 8 Comparison including trained evaluators — test set

| Structures | Accuracy | $S_c$ (%) | $S_p$ (%) | Figure of merit $D$ (%) |
|------------|----------|-----------|-----------|-------------------------|
| 4 (2/50/CE) | 0.8065   | 81.25     | 80        | 90.31                   |
| 11 (4/100/MSE) | 0.8710  | 81.25     | 93.33     | 93.63                   |
| 13 (3/150/CE) | 0.8710  | 81.25     | 93.33     | 93.63                   |
| Pre-trained CNN | 0.9032  | 93.75     | 86.67     | 95.10                   |
| Evaluator 1 | 0.9032   | 100       | 80        | 94.97                   |
| Evaluator 2 | 0.6451   | 100       | 26.67     | 79.86                   |
| Evaluator 3 | 0.8710   | 100       | 73.34     | 93.26                   |

Table 9 Comparison of computational cost

| Structures | Computational cost (FLOPs) |
|------------|---------------------------|
| 4 (2/50/Cross-Entropy) | 6.58 M |
| 11 (4/100/MSE) | 13.20 M |
| 13 (3/150/Cross-Entropy) | 19.79 M |
| Pre-trained CNN ResNet-18 | ≅ 2 G |
Table 9 compares the computational cost in Floating-point Operations (FLOPs) of the three classifiers designed using the proposed methodology with that of the pre-trained CNN. To measure the FLOPs required by the MLP-ANN structures, the number of operations necessary for the classification of one cutout were accounted for in FLOPs, estimated as reported in Hunger (2005) and Carmo (2017). For instance, for the network structure number 4 we obtained 2 NHL, 50 NNPHL plus the cost of the nonlinear activation function.

The number of FLOPs required for an exponential operation depends strongly on the implementation, including software and hardware employed. Also, the computational complexity in FLOPs is a nonlinear function of the input signal value. Considering the average implementation information [see, for instance, Stackoverflow (2017)] that the calculation of the $\tanh(x)$ has a cost between 10 and 100 FLOPs, and that this cost is approximately twice the calculated cost of an exponential function, we estimated the cost of an exponential operation to be between 5 and 50 FLOPs. Then, for our calculation we assumed the worst case of 50 FLOPs for each exponential operation.

Defining $N_i$ as the number of neurons in the network input, $M_1$ as the number of neurons in the hidden layer, $C$ as the number of hidden layers, and $M_2$ as the number of network outputs, we arrived at the following estimator for the number of FLOPs required for a classification:

$$N_{\text{Flops}} = (2M_1N_1 - M_1) + C(85M_1) + (C - 1)(2M_1^2 - M_1) + (2M_2M_1 - M_2) + 109M_2$$

Equation 11 was then used to estimate the computational cost for the three MLP-ANN structures, leading to the values shown in Table 9.

Regarding the 72-layer pre-trained Resnet-18 CNN, the computational complexity for the classification of an image size $224 \times 224 \times 3$ was calculated in Ferrari et al. (2018); OpenGenus (2019) and Albanie (2017) to be approximately $2 \ G$ FLOPs, as shown in Table 9.

It is noted that the pre-trained CNN has an operation complexity at least 100 times greater than the most complex network among networks 4, 11 and 13. This increase in complexity is also accompanied by a significant increase in the amount of memory required for the CNN, when compared to the other networks. This much higher complexity does not justify the corresponding modest increase in performance, showing that the contribution of theoretical expertise to the design of the classifier can easily surpass the advantages of using very sophisticated neural network structures to classify raw data.

We have also verified the performance of the classifiers when using different number of cases during their training. Our results (Fig. 6) showed that the performance of the proposed MLP classifiers converges to that of the pre-trained CNN as more cases are used for training. Hence, weighting implementation cost and classification performance, the generally accepted superiority of a CNN solution for any image classification application is clearly open to question.

Another important aspect of the specific application is to address the value of the proposed classifiers as a supporting tool for the pathologist in reaching the correct diagnosis. The value of such support should be evaluated considering...
the fact that a diagnosis of epithelial dysplasia is rarely made by a large number of experts. Also, even a group of highly trained professionals may reach diverse conclusions, especially in less obvious cases. Mathematically, these two aspects indicate a tendency of large variability in the diagnosis made by a small set of pathologists. Recognizing such a tendency, it is of interest of the pathologist to have the support of a well designed classifier when analyzing oral lesions for detecting dysplasia.

This tendency of high variability among trained evaluators was observed in Table 8, where average accuracy of evaluators was 0.8064, but with standard deviation of 0.1406, while the average accuracy of the three MLP networks was 0.8495, with standard deviation of 0.0372. Although these results are not statistically significant due to the small number of trained evaluators, they correspond to a practical situation, as the diagnosis will rarely be made by considering a large number of pathologists opinions. These results corroborate the expected tendency of having a large variability in classifications by few trained evaluators, which suggests that the help of a well trained algorithmic classifier should be welcome as a support for decision.

Another interesting observation in Table 8 is that the three evaluators correctly classified all dysplastic cases (sensitivities were 100%). However, the trained evaluators yielded specificities equal to or lower than those obtained with neural classifiers. This fact suggests that the trained evaluators tended to classify cases as dysplastic when there was doubt in the classification. This tendency to be “on the safe side” leads to an increase in the number of false positives, which corresponds to using a risk function with $\lambda_{21} > 1$ for reaching the decision.

Conclusions

This study showed that the multilayered network structures combined with the pathologists’ knowledge to choose the cutout region that would be delivered to the network, presented performances similar to those of the network that is considered the state of the art in image classifications (pre-trained CNN), with considerable less operational complexity. In addition, it was analyzed that increasing the number of cases used in the training of MLP networks would bring this performances even closer.

Finally, the average performance of three trained evaluators was compared with the average performance of the three MLP networks. We observed that they resulted in very close averages of accuracy, but the standard deviation of neural structures was approximately 74% lower than the standard deviation of trained evaluators. This high variability in the diagnoses of trained evaluators may be associated with their emotional state during the classification of cases. Thus, using a well trained classifier to aid in the diagnosis, could be welcome to reduce this high variability found.

Funding This study was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), by CNPq grants 304250/2017-1 and 409044/2018-0, and by FAPESC grant 201TR1460.

Declarations

Conflicts of interest The authors declare no competing interests.

References

Adel D, Mounir J, El-Shafey M, Eldin A, Masry N, AbdelRaouf A, Ibrahim I. Oral epithelial dysplasia computer aided diagnostic approach. 2019.

Alabi RO, Youssef O, Pirinen M, Elmusrati M, Mäkitie AA, Leivo I, Almangush A, Mäkitie M. Machine-learning in oral squamous cell carcinoma: current status, clinical concerns and prospects for future—a systematic review. Artif Intell Med. 2021:115:102060. https://doi.org/10.1016/j.artintmed.2021.102060.

Albanie S. Memory consumption and flop count estimates for convnets, 2017.

Almubarak HA, Stanley RJ, Long R, Antani S, Thoma G, Zuna R, Frazier SR. Convolutional neural network based localized classification of uterine cervical cancer digital histology images. Procedia Comput Sci. 2017:114:281–7. https://doi.org/10.1016/j.procs.2017.09.044.

Arevalo J, Cruz-Roa A, Arias V, Romero E, Gonzalez F. An unsupervised feature learning framework for basal cell carcinoma image analysis. Artif Intell Med. 2015:64:131–45.

Baik J, Ye Q, Zhang L, Poh C, Rosin M, MacAulay C, Guillaud M. Automated classification of oral premalignant lesions using image cytometry and random forests-based algorithms. Cell Oncol. 2014:37(3):193–202. https://link.springer.com/article/10.1007/s13402-014-0172-x.

Bouquot J, Speight P, Farthing P. Epithelial dysplasia of the oral mucosa-diagnostic problems and prognostic features. Curr Diagn Pathol. 2006:12:11–21.

Bubacar E, Miguel A, Konrath A, Modolo F, Rivero E. Evaluation of two classification systems for oral epithelial dysplasia. Oral Dis. 2021:1–5. https://doi.org/10.1111/odi.13867.

Carmo DM. Algoritmo adaptativo de redução de ruído e preservação das diferenças de nível interurral para aparelhos auditivos, M.S. thesis, Universidade Federal de Santa Catarina. 2017.

El-Naggar AK, Chan JKC, Takata T, Grandis JR, Slootweg PJ. WHO classification of tumours. Pathology and genetics of head and neck tumours. 4th ed. Lyon, France: IARC Press; 2017.

Fawcett T. Introduction to roc analysis. Pattern Recogn Lett. 2006:27:861–74.

Ferrari V, Hebert M, Sminchisescu C, Weiss Y. Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part V, vol. 11209, Springer, 2018.

Gupta RK, Kaur M, Manhas J. Tissue level based deep learning framework for early detection of dysplasia in oral squamous epithelium. J Multimed Inf Syst. 2019;6:81–6.

Hunger R. Floating point operations in matrix-vector calculus, Technical Report, Associate Institute for Signal Processing, Technische Universität München, version 1.3, 2007. https://mediatum.ub.tum.de/doc/625604/625604.
James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. New York: Springer; 2013.

Krishnan MMR, Pal M, Bomminayuni SK, Chakraborty C, Paul RR, Chatterjee J, Ray AK. Automated classification of cells in subepithelial connective tissue of oral sub-mucous fibrosis—an SVM based approach. Comput Biol Med. 2009;39(12):1096–104. https://doi.org/10.1016/j.compbiomed.2009.09.004.

Krishnan MMR, Venkatraghavan V, Acharya UR, Pal M, Paul RR, Min LC, Ray AK, Chatterjee J, Chakraborty C. Automated oral cancer identification using histopathological images: a hybrid feature extraction paradigm. Micron. 2012;43(2–3):352–64. https://doi.org/10.1016/j.micron.2011.09.016.

Kujan O, Oliver RJ, Khattab A, Roberts S, Thakker N, Sloan P. Evaluation of a new binary system of grading oral epithelial dysplasia for prediction of malignant transformation. Oral Oncol. 2006;42:987–93. https://doi.org/10.1016/j.oraloncology.2005.12.014.

Mahmood H, Shaban M, Rajpoot N, Khurrum SA. Artificial intelligence-based methods in head and neck cancer diagnosis: an overview. Br J Cancer. 2021;124(12):1934–40. https://doi.org/10.1038/s41416-021-01386-x.

Mathworks. Resnet18, 2019.

Mello FW, Miguel AFP, Dutra KL, Porporatti AL, Warnakulasuriya S, Guerra ENS, Rivero ERC. Prevalence of oral potentially malignant disorders: A systematic review and meta-analysis. Oral Pathol Med. 2018;47(7):633–40. https://doi.org/10.1111/jop.12726.

Nag R, Das RK. Analysis of images for detection of oral epithelial dysplasia: A review. Oral Oncol. 2018;78:8–15. https://doi.org/10.1016/j.oraloncology.2018.01.005.

OpenGenius. Floating point operations per second (flops) of machine learning models, 2019.

Pallua JD, Brunner A, Zelget B, Schirmer M, Haybaeck J. The future of pathology is digital. Pathol Res Pract. 2020;216(9):153040–8. https://doi.org/10.1016/j.prp.2020.153040.

Prabavathy S, et al. Analysis of oral epithelial dysplasia using machine learning technique. IT in Industry. 2021;9(3):618–21.

Ranganathan K, Kavitha L. Oral epithelial dysplasia: Classifications and clinical relevance in risk assessment of oral potentially malignant disorders. J Oral Maxillofac Pathol. 2019;23(1):19–27. https://doi.org/10.4103/jomfp.JOMFP_13_19.

Sami M, Saito M, Kikuchi H, Saku T. A computer-aided distinction of borderline grades of oral cancer. In: Proc. 16th IEEE International Conference on Image Processing (ICIP). 2009. pp. 4205–4208.

Stackoverflow. How many flops does tanh need? 2017.

Talo M. Automated classification of histopathology images using transfer learning. Artif Intell Med. 2019;101:101743. https://doi.org/10.1016/j.artmed.2019.101743.

Venkatakrishnan S, Kalyani C, Dhanalakshmi P. Analysis of leukoplakia disease classification using radial basis function neural network. SSRG Int J Med Sci. 2017;4:7–11. https://doi.org/10.14445/23939117/IJMS-V4I3P102.

Warnakulasuriya S, Reibel J, Bouquot J, Dabelsteen E. Oral epithelial dysplasia classification systems: predictive value, utility, weaknesses and scope for improvement. Oral Pathol Med. 2008;37:127–33. https://doi.org/10.1111/j.1600-0714.2007.00584.x.

Zhang L, Lu L, Nogues I, Summers RM, Liu S, Yao J. DeepPap: Deep convolutional networks for cervical cell classification. IEEE J Biomed Health Inform. 2017;21(6):1633–43. https://doi.org/10.1109/JBHI.2017.2705583.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.