Diagnosis with Confidence: Deep Learning for Reliable Classification of Squamous Lesions of the Upper Aerodigestive Tract

Mélanie Lubrano*[1,2,3], Yaëlle Bellahsen-Harrar*[4,5], Sylvain Berlemont[3], Thomas Walter[1,6,7], Cécile Badoual[4,5]

**co-first authors: these authors contributed equally*

[1]Centre for Computational Biology (CBIO), Mines Paris, PSL University, 75006 Paris, France
[2]Keen Eye, 75012 Paris, France
[3]Tribun Health, 75015 Paris France
[4]Département de Pathologie, Hôpital Européen Georges-Pompidou, APHP, France
[5]Université Paris Cité, 75006 Paris, France
[6]Institut Curie, PSL University, 75005 Paris, France
[7]INSERM, U900, 75005 Paris, France

Correspondence:

Cécile Badoual
cecile.badoual@aphp.fr
ORCID: 0000-0002-1143-3085
Département de Pathologie
Hôpital Européen Georges-Pompidou
20 Rue Leblanc, 75015 Paris, France
Phone: 01 56 09 38 86

Abstract

Diagnosis of head and neck squamous dysplasia and carcinomas is critical for patient care, cure and follow-up. It can be challenging, especially for intraepithelial lesions. Even though the last WHO classification simplified the grading of dysplasia with only two grades (except for oral or oropharyngeal lesions), the inter and intra-observer variability remains substantial, especially for non-specialized pathologists. In this study we investigated the potential of deep learning to assist the pathologist with automatic and reliable classification of head and neck squamous lesions following the 2022 WHO classification system for the hypopharynx, larynx, trachea and parapharyngeal space. We created, for the first time, a large scale database of histological samples intended for developing an automatic diagnostic tool. We developed and trained a weakly supervised model performing classification from whole slides images. A dual blind review was carried out to define a gold standard test set on which our model was able to classify lesions with high accuracy on every class (average AUC: 0.878 (95% CI: [0.834-0.918])). Finally, we defined a confidence score for the model predictions, which can be used to identify ambiguous or difficult cases. When the algorithm is applied as a screening tool, such cases can then be submitted to pathologists in priority. Our results demonstrate that the model, associated with confidence measurements, can help in the difficult task of classifying head and neck squamous lesions.

Introduction

Head and neck squamous cell carcinomas (HNSCC), ranked 6th cancer worldwide, constitute a major public health issue because of their high mortality rate and the morbidity of their treatment regimens (1–3). These poor figures can be explained by a late diagnosis, usually at an advanced stage of the disease. However, it is estimated that early diagnosis of potentially malignant head and neck lesions could prevent almost 90% of cancers (4). The early detection of HNSCC could be allowed by a precise follow-up of precancerous lesions, or squamous dysplasias, depending on their potential to become invasive. The classification of head and neck dysplasias has been a highly controversial issue for many years. Indeed, since the first classification proposed by Kleinsasser in 1963 (5), many different classifications have been proposed by expert pathologists without achieving a strong consensus (6,7). Each of them had different terminologies and methods of grading, but their reproducibility was always low to moderate (6,8–12) (Table 1). Nevertheless, higher grades were associated with a higher risk of transformation into carcinoma, confirming their significance (13).

In order to improve inter-rater and intra-rater reliability, the World Health Organization (WHO) recommended to grade laryngeal squamous dysplasias with only two categories : low-grade and high-grade. The high-grade category encompasses moderate, severe dysplasia and carcinoma *in situ*. This system showed a large difference in terms of severity. In fact, Gale et al. observed in their work that high grade dysplasia were ten times more at risk to evolve to invasive carcinoma than low grade lesions (14). The last classification taken up by the (WHO) in 2022 (15,16) confirms the simplification of the grading proposed in 2017, even if the difference between high-grade dysplasia and *in situ* carcinoma is questioned.

For the oral cavity, the WHO grading system kept the distinction of moderate and severe lesions. Interestingly, this is the only grading system keeping three separate grades, even if oral squamous lesions share the same pathophysiology and are induced by the same carcinogens than in the larynx, hypopharynx, trachea and parapharyngeal space. Notably, moderate and severe lesions are reunited in a “high-grade dysplasia” category for squamous dysplasia grading in the uterine cervix (17), the anus (18) and the esophagus (19). Two-grades systems are also used for glandular dysplasias such as in colorectal adenomatous polyps or Barrett's esophagus (19). In light of these considerations, the evaluation of oral dysplastic lesions could benefit from a two-grades classification system.
Nevertheless, reproducibility between pathologists remains moderate for all grading methods (20). This difficulty to classify is due to the multiple elements to take into account both at cytological and architectural levels, on an epithelium that can have noticeable variations of thickness depending on the anatomical location, and inflammatory and dystrophic alterations that can sometimes be difficult to distinguish from true dysplasia (16). Finally, the arbitrary classification categories imposed on a continuous spectrum of lesions that have no absolute and clearly definable boundaries induces more subjectivity.

Given these considerations, the field is in need of new tools to help pathologists make robust and consistent classifications of squamous head and neck lesions, for better clinical management.

In order to help pathologists to perform precise and robust diagnoses, many artificial intelligence (AI) algorithms have been developed these last years (21). Deep learning models can learn meaningful patterns without explicit definition by an expert. Weak supervision has proven to be a very powerful strategy for many classification tasks in computational pathology, from cancer detection (22–24) to classifying carcinoma subtypes (25,26), grading (27), prognosis, prediction of molecular signatures (28,29) and primary origins (30). Yet, only a few papers about grading dysplasias have been published so far and studies applied to head and neck pathology are even more scarce. This can be explained by the lack of a public database including dysplastic lesion annotations and the difficulty to reach grading consensus, as shown by the low reproducibility even between expert pathologists. Interestingly, most studies used classical supervised machine learning methods rather than deep learning (4,31) and focused on the oral cavity, with no study about laryngeal lesions (31).
Classification of head and neck squamous lesions could similarly benefit from computer-assisted analysis by helping pathologists standardize and reduce bias of their grading.
To be integrated in pathologists workflows, AI models should provide an assessment of the models confidence for each prediction. Even though measuring AI model uncertainty has been studied frequently in the past years (32–34), it was rarely applied to computational pathology. Most works focused on segmentation (35–38) but less on diagnosis tasks (39,40). In a recent work, Poceviciute et al. (41) stressed out the need for an AI reliability measure for diagnosis and compared methods to assess it. Lu et al. (30) evaluated their prediction model by computing the top-k differential diagnostic accuracy to determine the primary origin of carcinomas. By integrating the prediction probabilities directly in their workflow, their model

helped reduce the potential primaries to investigate. Finally Dolezal et al. (42) proposed a thresholding paradigm not susceptible to domain shift for reliable use of uncertainty measure in clinical practice.
As the grading of laryngeal dysplasia is often controversial, even for experts, tempering a model's predictions with a measure of its confidence could help pathologists to integrate them into their grading choice.

The aim of this work was to develop a fully automated, weakly supervised model for the diagnosis of dysplasias and squamous cell carcinomas of the head and neck associated with a method to assess the models confidence for each prediction. The classification followed the WHO grading system for the hypopharynx, larynx, trachea and parapharyngeal space. We compared the models reproducibility to pathologists and evaluated its performances on a gold standard test set. We propose a measure of confidence of the model's predictions providing pathologists with a score indicating to which extent they can trust the prediction. We showed that this score was consistent with pathologists' hesitations when grading dysplasia, and we believe that it can greatly enhance acceptance of such an automatic grading system.

Material and Method

Selection of patients

Patients were selected retrospectively from 2000 to 2013 from the Hôpital Européen Georges Pompidou (HEGP, Paris, France) clinical database (DxCare® software). Patients were at least 18 years old and diagnosed with head and neck squamous cell dysplasia or carcinoma, either of the larynx, pharynx, nasopharynx, hypopharynx, lateral edge of the tongue or oral cavity. Oropharyngeal cases were excluded from the study, since in this anatomical region most SCC are related to the *Human Papillomavirus*, have specific aspects (43), and the existence of dysplastic lesions is still debated with no diagnostic guidelines from the WHO. Carcinomas sampled after chemotherapy or radiotherapy were excluded since the treatment modifies the aspect of the lesion and the data do not fit the use case of the model.

Selection of samples and slides

The samples were identified in the HEGP Pathology Department database (Diamic® software). Both biopsies and surgical samples were included, but not all the surgical samples of carcinomas per patient that were available in the hospital's archives. This selection was made in order to keep a balance between classes in the dataset. When several slides of the same sample were available, one investigator (YBH) selected the slide where the lesion was most visible. Every pathology report was read to assess which types of lesions were present. Samples were excluded if the pathologist mentioned in the report that it was impossible for him to distinguish between high-grade dysplasia and invasive carcinoma because of tangent inclusion. Verrucous carcinomas were also excluded. The slides were stained at the time of the sampling with Hematoxylin Eosin and Saffron (HES) staining.

Digitized dataset constitution

The slides were digitized with a pathology slide scanner (Hamamatsu NanoZoomer® s360) at 20X magnification (pixel resolution of 0.45 µm). All the sections present on the slide were scanned. The quality of the digitization was checked for all the WSIs. Scans were excluded if the digitization failed after two attempts. During digitization, each slide was given an identification number for anonymization purposes.

After digitization, the WSIs were uploaded to the EyeDo© platform. Each slide was given a global label corresponding to the most severe lesion in the sample, following the WHO classification, according to the clinical report. These initial labels were thus provided by several pathologists between 2000 and 2013. Slides with no surface epithelium and slides with strong artifacts were excluded but slides with artifacts that did not impair a clear diagnosis were kept in the dataset.

Review of "mild to moderate" dysplasias

Because the samples were selected from 2000 to 2013, many dysplasias were diagnosed at this time with a "mild to moderate" grade. However, this grading does not follow the WHO classification anymore, since mild dysplasia is synonymous with low grade and moderate dysplasia with high grade. The 127 "mild to moderate" samples were reviewed jointly by the pathologist investigators (YBH, CB) in order to assign a grade compatible with the latest WHO grading system.

Blind review for assessment of a gold standard test set

Diagnosis of squamous dysplasias by pathologists lacks reproducibility (8). In order to assess the model's performance in the best conditions, pathologist investigators made a dual blind review on a selected portion of the dataset. This portion was considered as a gold standard test set. Only biopsies were selected for this test set, as they correspond to the use case of the model (for surgical samples, the diagnosis is usually already known at the time of the surgical resection). The data scientist investigators, not taking part in the grading, were in charge of selecting the slides. The samples were selected based on their labels, such that classes were balanced. To make sure the test samples were independent, slides were selected in the following way: all the patients with only one slide present in the initial dataset were attributed to the test set; then the patients associated with multiple slides (no more than two) were attributed to the test set only if the slides were acquired multiple years apart. Slides from different samples of the same patient but from the same year were discarded. Thus, all slides in the gold standard test set are considered independent from the rest of the cohort. The reference standard for the internal test dataset was determined in two rounds. Two raters, both with expertise in head and neck pathology, but at different career stages (reviewer 1-CB: international expert, reviewer 2-YBH: early-career pathologist) independently reviewed the slides of the test set. The review was carried out on the digitized slides, through the online viewer on the EyeDo© platform. The raters were blind to any clinical information, the initial diagnosis, and the rating of the other assessor. They were aware that the test set had been chosen to be balanced (according to the initial labels). To avoid bias resulting from this information, the raters could not change their diagnosis after a slide had been reviewed and assessed. The two raters finally met during a consensus meeting to discuss the slides on which they disagreed. If the disagreement

persisted, the slides were excluded. Label noise in the dataset was measured by comparing final consensus labels with initial diagnosis from patient records.

Data management

To assess the diagnostic capacity of the model, 128 slides were selected from the entire dataset and reviewed to constitute the gold standard test set as described above. The remaining slides (2121 WSIs from 498 patients) were split into 5 randomly sampled training and validation sets following the 80%-20% standard to train the deep learning model and fix hyperparameters according to best practices. To ensure class balance between training and validation sets, the splits were performed in a stratified manner by patients and grades using the Multilabel Stratified K-Fold algorithm from Sedichis (44) (worst grade was kept for each patient). In the training splits, low grades slides were upsampled (1.5 times) to reduce the effect of class imbalance. There was no patient overlap between training, validation and gold standard test sets.

Deep Learning Model

Implementation details are provided in the Supplementary Materials.

Tissue Selection

In order to train the WSI classification model, slides were divided into smaller images of 224x224 pixels at a resolution of 1 µm/pixel. When tiling the entire sample and removing only the white background, WSIs contained up to 8800 tiles (with an average of 1300 tiles per slide), giving a total of 3.9 millions of tiles across the dataset. To reduce the processing time, we removed any tiles not containing epithelial cells (from surface epithelium or tumor). To do so, we first trained a UNet (45) to perform binary segmentation between epithelium and carcinoma tissue (the “foreground”) versus any other type of tissue or background (the “background”). The UNet was trained at 10X resolution (1 µm/pixel) on 5439 annotated tiles of 512x512 pixels from 121 slides. The classification threshold was modified to 0.4 in order to minimize false negatives and thus make sure all the tissue of interest was selected. Non-overlapping tiles were then extracted from the selected tissue, resulting in 2.2 millions of tiles.

Multiple Instance Learning architecture

Our model was derived from the Attention-based Multiple Instance Learning (MIL) architecture proposed by Ilse et al (46). It consisted of a feature extractor, a scoring module and a classification module. Due to their size, WSIs have to be cut into smaller tiles. Features are extracted from each tile through a frozen convolutional neural network (DenseNet121, (47)), resulting in feature vectors of dimension 1024. A global label (the grade of the worst lesion on the slide) is associated with the bag of tiles. Thanks to an attention mechanism, the network learns which tiles within the bag are most important for the grading of the lesion and attributes a score. The classification module aggregates the attention scores and the extracted tile representations to obtain a slide representation (weighted sum of the tile representations, where the weights are the attention scores) from which the slide-level label is predicted.

Cost-sensitive training

Due to the ordinal nature of our classes, we used a cost-aware classification loss introduced in (48). The network was trained to predict a class-specific risk rather than a posterior probability. The predicted class corresponds to the class minimizing this risk. The risk was defined

according to the cost matrix, inspired by the TissueNet Challenge organized by the French Society of Pathology in 2020 (49) and penalized large errors (e.g. benign vs high grade dysplasia) more than small errors (e.g. low grade vs high grade). The cost matrix can be found in the Supplementary Material (Table A).

Self-Supervised pre training

We initialized the feature extractor with pretrained weights obtained with self-supervised training. SimCLR (50), a simple architecture relying on data augmentation, was trained on all the tiles in the dataset (3.5 Millions tiles of 336x336 pixels) for 300 hours (143 epochs) on 8 GPUs. During the training, each tile was randomly cropped and resized to 224x224 pixels, flipped and rotated. The H&E staining was modified with RGB to Haematoxylin-Eosin-DAB (HED) color space conversion and colors were randomly altered (contrast, hue, brightness, saturation).

Test and Evaluation

The deep learning model was trained on the five cross-validation training splits. Early stopping, monitoring the validation loss, was used to stop the training (with a patience of 60 epochs). The five models were assembled to make predictions on the gold standard test set. Overall predictions were obtained by averaging the posterior probabilities of the five models. Class-wise classification metrics were computed in a one-versus-all manner, the average of the class-wise scores was performed to compute overall performances. Performances comparing models predictions and consensus labels were measured. Metrics were reported with 95% confidence intervals (CI) using a bootstrapping method (10 000 iterations).

Confidence score

In order to assess the certainty of the model's predictions, a confidence score was computed for every prediction of the model on the gold standard test set. The confidence score was derived from the risk estimation output by the last layer of the network. The softmax of the inverted risk (- risk vector) was computed, turning cost estimation into probabilities. The confidence score was defined as the difference between the two highest risk probabilities: if the probabilities were close, the network was hesitating between two classes, if they were far, the network was considered more confident. As the confidence score is derived from the cost sensitive risk estimation, it takes into account the ordinal characteristic of the classes; smaller confidence values hence reflect pathologists hesitations.

This confidence score was designed for potential application of the model in the context of screening. To exclude predictions with poor confidence, a threshold was set to filter out uncertain predictions. The threshold was optimized on the validation set to reach an overall AUC > 0.9. The performance of the model after filtering uncertain predictions was assessed on the gold standard test set.

Analysis of misclassified slides

Attention scores learned by the MIL model reveal tiles that strongly influence the decision and are thus supposed to be of diagnostic relevance. Heatmaps of attention scores were overlaid on WSIs on the EyeDo© Platform in order to inspect important regions. Qualitative analysis of tiles with high predictive value was performed.

Results

Training dataset

The data used for training and validation of the model consisted of 2144 slides. The low-grade dysplasia class was significantly underrepresented (10.7% of the total, 229 slides). Other classes were present in balanced proportions. Patient characteristics and a summary of the cohort are presented in Table 2.

Slides reviews

"Mild to moderate" dysplasia

Almost a third of the "mild to moderate" dysplasias were reclassified during the review as not dysplastic. A summary of the review is shown in the Supplementary Materials.

Gold standard test set

After the first round of grading, the two graders independently agreed on 79 slides. The remaining 52 slides were discussed at the consensus meeting. Three slides were excluded: one slide for which it was impossible to distinguish high-grade dysplasia from invasive carcinoma; one slide for which it was impossible to choose between low-grade dysplasia and artifacts; and one slide for which there was a suspicion of carcinoma in the chorion but with no connection to the surface epithelium, which was normal. Consensus was achieved for the 49 other slides, resulting in 128 slides from 110 patients to be used as the gold standard test set. The labels of the gold standard test set before and after review are shown in Supplementary Materials (Table B).

Classification performance of the deep learning model

When considering the consensus labels as an absolute ground truth, the AI model achieved an average AUC on the 4 classes of 0.878 (95% CI: [0.834-0.918]). The AI model reached an AUC > 0.8 for all 4 classes. ROC AUC can be found in Figure 1. Average AUC dropped significantly when using the *initial labels* as the ground truth rather than the *consensus labels* to evaluate the AI model's predictions. (AUC=0.832 [0.787-0.875]). This shows that the review reduced the noise in the labeling, leading to better classification performances. Classification performances are summarized in Table 3 and confusion matrices are shown in Figure 2. The misclassified slides are listed in Table C in Supplementary Materials. The majority of the misclassifications came from the "low grade" dysplasia class, with seven slides misclassified as benign. The initial label of all of these slides was "benign", suggesting that there is some ambiguity in these cases. As many slides in the gold standard test set were classified as low grade dysplasias and had a "benign" initial label, these misclassifications could be explained by a significant difference of grading between the validation set and the test set.

Four slides of carcinoma were missed by the AI model and predicted as low-grade or high-grade dysplasia, but with a low confidence score beneath the threshold which would have filtered them out. Three of these slides had significant artifacts and the other showed carcinoma under a non-dysplastic epithelium, which could be more difficult for the model to identify because of the rarity of this presentation.

Analysis of misclassified slides

In the test set, three high grade dysplasias were misclassified as carcinomas, two of them with a high confidence score. For two of these slides, the tiles with the highest attention score showed similar aspects, with marked atypia restricted to the basal layers and a corrugated aspect of the lamina propria. This aspect was identified with a high attention score on a slide labeled low grade dysplasia and confidently classified as high grade dysplasia by the model (Figure 3). These lesions were sampled in the larynx. In the validation sets, analysis of five slides misclassified from high grade to carcinoma or from low grade to high grade revealed the same aspects on high attention tiles.

Assessment of inter-rater agreement

To assess the noise present in our ground truth labels we measured the agreement between the reviewers (reviewers 1 and 2), the initial labels and the AI model (Table 4). The main metric used to measure agreement was the linear Cohen's kappa. Agreement between the AI model and the initial labels was slightly lower (linear Cohen's kappa = 0.641 [0.546-0.726]) than the agreement between reviewer 1 or 2 versus the initial labels (linear Cohen's kappa = 0.689 [0.606-0.764] and 0.723 [0.634-0.803] respectively), however it remained substantial (51). Agreement between reviewer 1 and reviewer 2 led to a linear Cohen's kappa of 0.676 [0.592-0.753]. These figures were of the same order of magnitude as in the literature (8).

Confidence score assessment

For the correct predictions, the confidence score was on average 0.73 +/- 0.303 compared to 0.418 +/- 0.303 for incorrect predictions. The confidence threshold, optimized to reach an overall AUC > 0.9 on the 5 validations sets, was set to 0.5. On the gold standard test set, at this threshold, 52 slides (40.6%) were considered as uncertain, most of them being low grade dysplasias. On the remaining slides, the Invasive Carcinoma AUC was 0.987 [0.962-1.000]. No carcinoma slides were missed by the model (Negative Predictive Value of 1.000 [1.000-1.000]). The overall AUC improved by 4.5% (0.931 [0.892-0.965]) when removing slides with low confidence. Conversely, overall AUC computed on the uncertain slides was equal to 0.764 [0.672-0.848] (-12.2% compared to the overall AUC on the full gold standard test set). In Figure 1 (C and D) we see that the confusion matrix on the confident slides is almost diagonal. The confidence score being the difference between the 2 highest probabilities, we observed that the model was always hesitating between two adjacent classes. Additionally, in the Supplementary Figure D we show that when removing slides from the gold standard test set according to their confidence score, the metrics on the remaining slides were consistently increasing, while removing randomly picked slides led to erratic evolution of the performances. In Figure 4, we compared the confidence score distributions for slides dependent on whether the two reviewers agreed or not. The figure suggests that the confidence level of the model reflects the probability of disagreement between reviewers. We note that even for a confidence score of 0, we would expect disagreement in only 50% of the cases. This explains why the distributions are overlapping.

Discussion

To the best of our knowledge, we propose the first deep learning model for the grading of head and neck squamous lesions following the WHO classification system. In the literature, studies about grading dysplasia with deep learning are scarce. Most of them focus on cervix Pap Smears (52,53), which does not take into account the epithelial architecture. A model for classification of esophageal lesions was proposed in Tomita et al (54) but didn't differentiate low-grade from high-grade dysplasias.
Because no public database with annotated head and neck dysplasias nor "benign" epithelium was available, we collected a large scale dataset of head and neck samples and clinical data from the HEGP, a renown center for head and neck diagnosis and medical care in France.

As dysplasia grading is difficult and reproducibility between pathologists is low to moderate, a blind reviewed test set was generated in order to properly assess the AI model performance. We observed that the agreement between the two reviewers on this test set was in line with previous reports (6,8–11), illustrating once more the difficulty to obtain objective and robust grading of head and neck dysplasia and the need for new tools to help pathologists make reliable diagnoses. We developed and trained a weakly supervised deep learning model that was able to accurately grade head and neck dysplasia, offering a first tool for assisted diagnosis.
Even though measuring a diagnostic test reliability is mandatory in pathology (for example, controls for immunomarkings), very few studies in computational pathology developed a confidence score to accompany the model's predictions. In this work, we propose a novel confidence score that is defined as the difference between the posterior probability of the two top classes, ranging between 0 if the two top classes were equally likely for the network, to 1 if the network had maximal confidence in the top class.
Our confidence score showed that the model, when uncertain, was always hesitating between adjacent classes, mimicking pathologists doubts. This observation is in line with the fact that epithelial dysplasia belongs to a continuous spectrum of lesions. Our confidence score can be useful to mitigate artificial sharp borders imposed by the WHO grading system and help pathologists decide on the grade with more objectivity. Moreover, we believe that such confidence scores can greatly improve real-world applicability and acceptance: for screening purposes, the application of the tool can be restricted to cases with high confidence, and thus identify the slides for which review by a pathologist would be recommended in priority.
Misclassified slides were analyzed regarding their confidence scores. We observed that for slides with low confidence predictions pathologists were more likely to disagree.
The analysis of the four false negative carcinoma samples showed that severe technical artifacts (tangent cuts, staining artifacts…) have the potential to negatively impact the classification results. But we also found that they ultimately also lead to low confidence scores and would have been filtered out according to our procedure.
Misclassified slides associated with high confidence scores were mostly upgraded by one class (low grade classified as high grade, high grade classified as carcinoma). The analysis of the tile attention scores revealed that the model placed strong attention on severe atypia located in the lower half of the epithelium, with a corrugated lamina propria. The two pathologists reviewed

these highlighted tiles together and agreed that these aspects are challenging to interpret and were present only focally in the slides. Since pathologists analyze epithelial surface in its globality, focal severe atypia can be missed. The attention heatmap analysis, associated with the confidence score, could be of great use by pathologists as a tool to focus on the most severe dysplastic aspects, guiding their final grading.

The presented study is monocentric and retrospective. For this reason, it will be necessary to establish an external cohort for further validation. However, we note that the large time range of patient recruitment (13 years) conveys the dataset already a high degree of heterogeneity and thus limits the risk of overfitting.

To conclude, we propose a reliable and powerful deep learning model for the classification of head and neck squamous cell lesions, especially helpful in a context of lack of available experts to screen enough samples or give a second read on difficult interpretations. The confidence score is an original and efficient way to assess the reliability of the model's predictions, making it closer to medical tests standards of use. We believe this method is a significant milestone towards reliable AI assisted diagnosis in digital pathology workflow, especially for subjective tasks such as grading of head and neck squamous lesions.

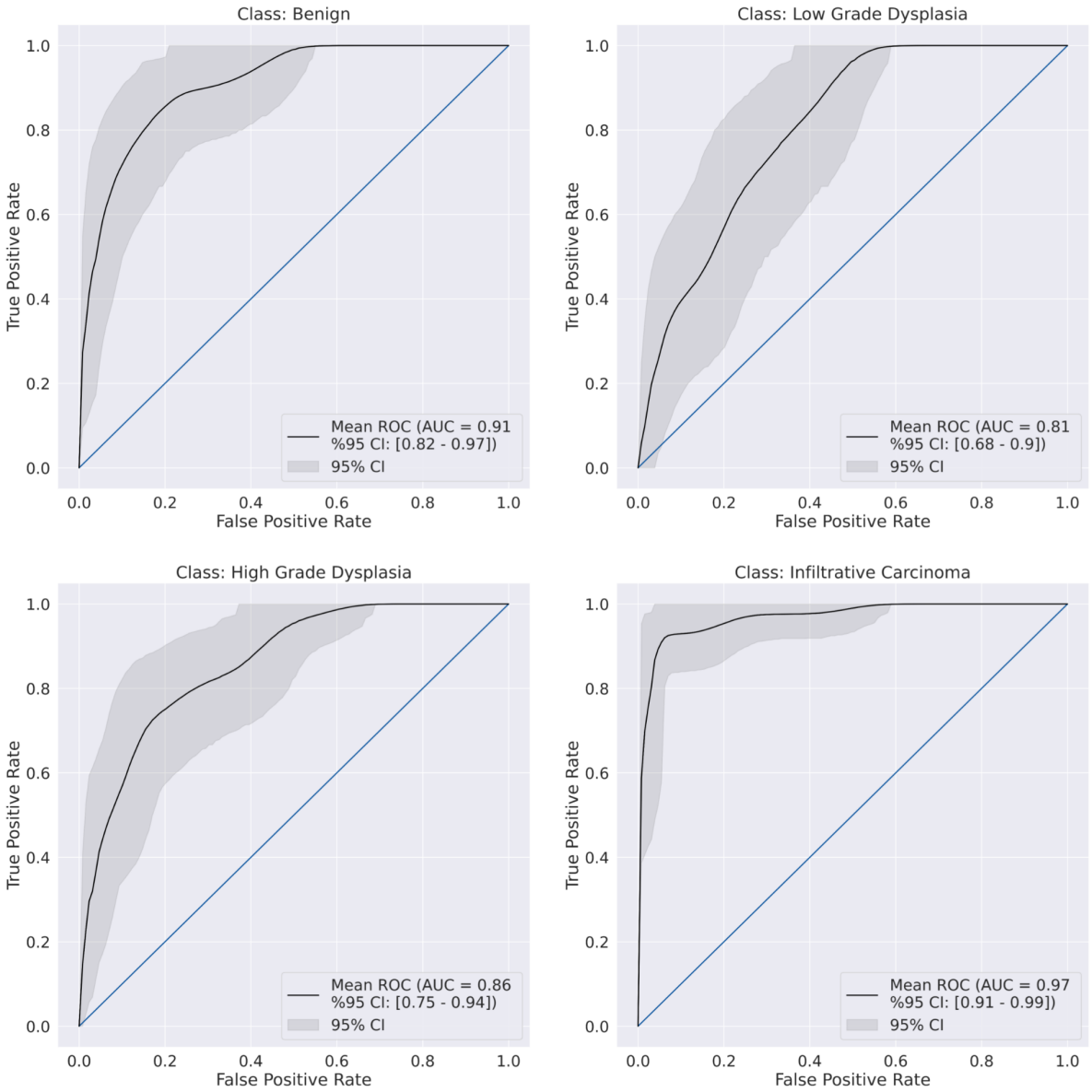
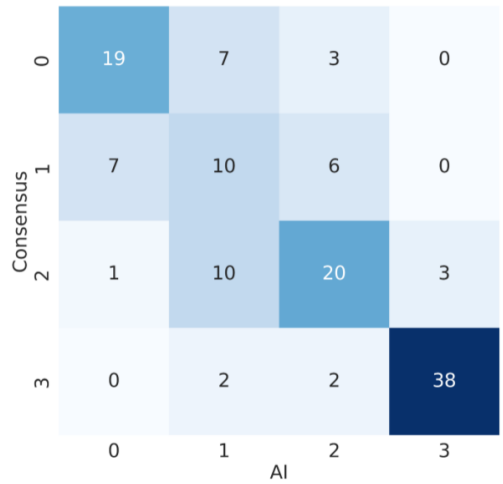
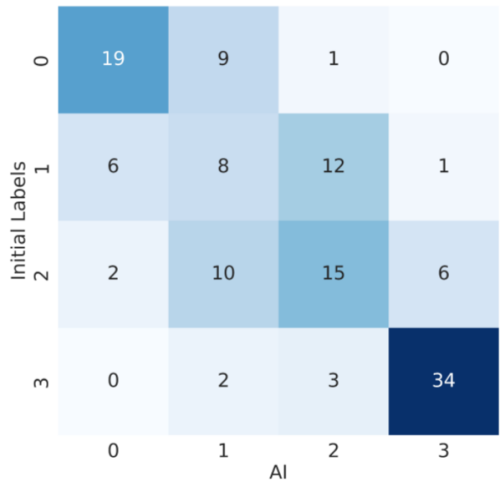


Figure 1 - AUC ROC for each class on the reviewed gold standard test set - ROC curves were obtained by bootstrapping of the AI model predictions (10 000 bootstrap samples). They were computed for each class in a One vs Rest manner using consensus labels as a ground truth. ROC = receiver operator characteristic. AUC area under the curve. ROC AUC of the Carcinoma class is better than for the other classes, certainly because the diagnosis of this class is often less ambiguous than for the other grades. Thus, the training data contains less noise on this class, as well as the test data. Misclassification on Carcinoma class concerned microinvasive lesions.

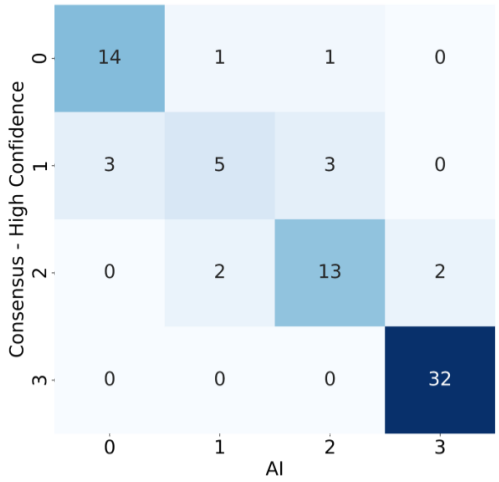
A. IA vs Consensus
(overall AUC= 0.886 [0.822-0.939])



B. IA vs Initial Labels
(overall AUC= 0.866 [0.796-0.929])



C. IA vs Consensus - High Confidence Slides (threshold = 0.51) *(overall AUC=0.931 [0.892-0.965])*



D. IA vs Consensus - Low Confidence Slides (threshold = 0.51) *(overall AUC= 0.764 [0.672-0.848])*

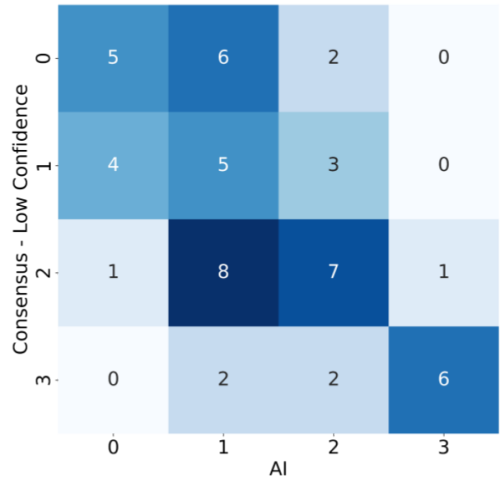


Figure 2 - Confusion Matrices - AI model's performances are evaluated on the gold standard test set on the reviewed labels (A) and the initial labels (from patient's records) (B). Numbers 0, 1, 2, 3 corresponds respectively to classes Benign, Low Grade, High Grade and Carcinoma. Classification performances are superior when using reviewed labels indicating that the review helped reduce noise in the labels. Confusion matrices show that the model is more confused on the Low Grade (1) and High Grade (2) classes, rather than the Carcinoma class (3) for instance which is justified by the ambiguity carried by this classes, on which even pathologist can struggle. Matrix C corresponds to the confusion matrix on the high confident slides at threshold=0.5, matrix D corresponds to the low confident slides. Matrix C is almost diagonal,

and the overall AUC on the confident slides subset is higher by more than 10% than on the unconfident subset. Additionally, we see that most of the Carcinoma slides are considered confident by the model.

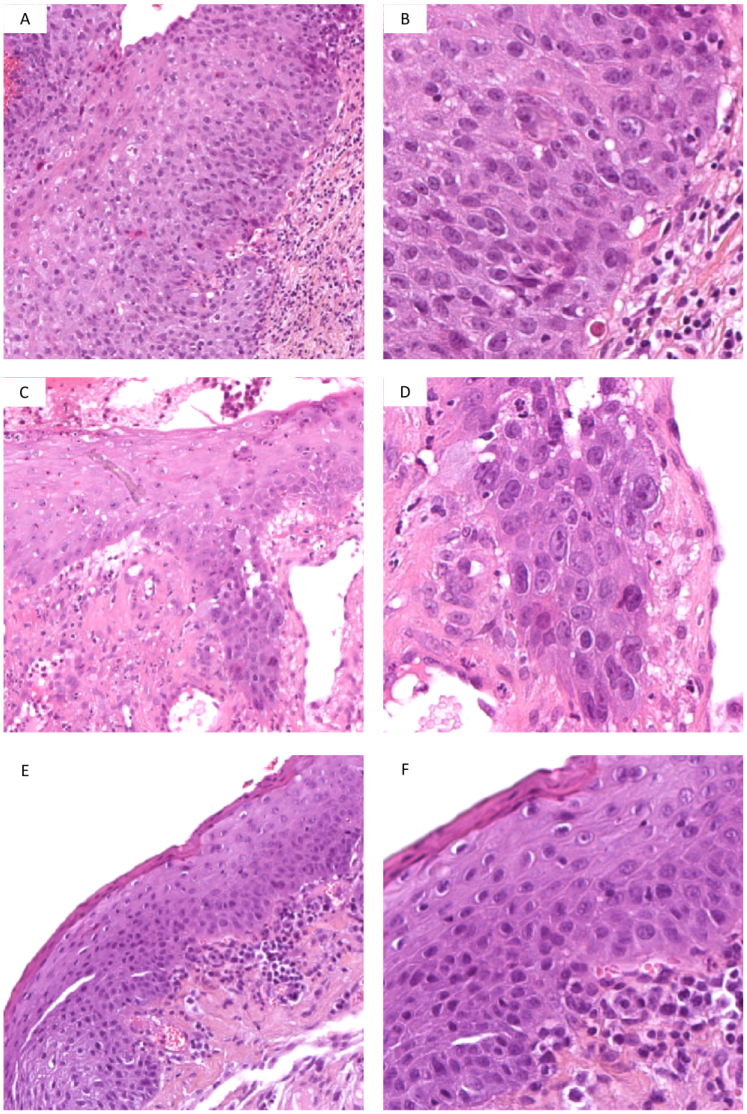


Figure 3 - Misclassified slides: attention score analysis - Tiles with high attention scores attributed by the MIL model. Left column: 20X magnification, right column: 10X magnification. A. and B. : slide_1597. High grade dysplasia predicted as invasive carcinoma. The model focused on marked basal atypia with a corrugated lamina propria C. and D.: slide_237. High grade dysplasia predicted as invasive carcinoma. The model focused on marked basal atypia and bulky rete ridges. E. and F.: slide_2712. Low grade dysplasia predicted as high grade. The model focused on marked basal atypia with a corrugated lamina propria. The three lesions are located in the larynx.

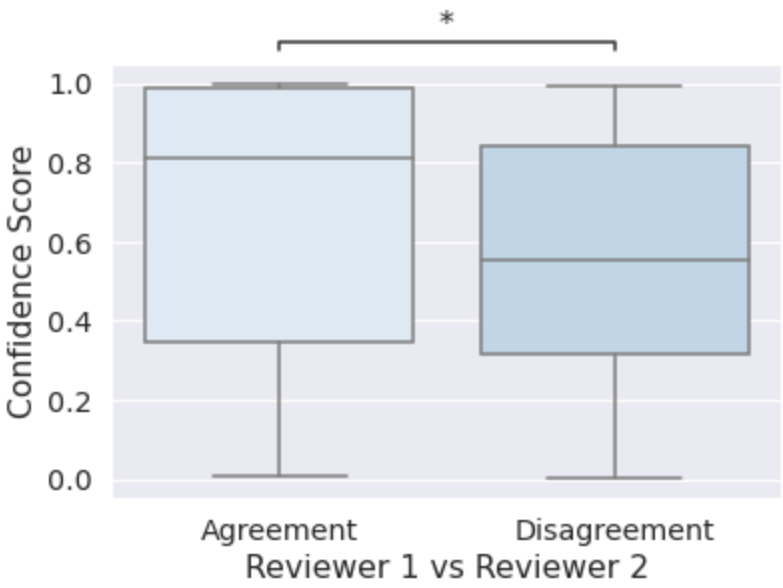


Figure 4 - Confidence level distributions - Comparison of confidence level and agreement between Reviewer 1 and Reviewer 2. The model is more confident on slides on which reviewers agreed during the dual blind review and is less confident on slides on which they disagreed. This suggests that the confidence score reflects the difficulties inherent to the slides, as pathologists would experiment with it. Significance: Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction, p-value=1.095e-02 U-stat=2.455e+03.

Table 1 - History of dysplasia classification - Overview of the different grading systems for intraepithelial head and neck lesions proposed over the years in the litterature. Table inspired from WHO *Blue Book*

Level of abnormal maturation	Ljubljana (2000)	SIN (2001)	WHO (2005)	LIN (2012)	Ljubljana (2014)	WHO (2017)
	Squamous hyperplasia	Squamous hyperplasia	Squamous hyperplasia		Low-grade SIL	Low-grade dysplasia
Lower 1/3	Basal/parabasal hyperplasia	SIN 1	Mild dysplasia	LIN 1		
Lower 1/3 - 1/2	Atypical hyperplasia	SIN 1 or SIN 2	Moderate dysplasia		High-grade SIL	
Lower 2/3				LIN 2		
1/2 - 3/4		SIN 2				High-grade dysplasia
More than 2/3				LIN 3		
All thickness	Carcinoma *in situ*		Severe dysplasia			
			Carcinoma *in situ*		Carcinoma *in situ*	

SIN: Squamous Intraepithelial Neoplasia
LIN: Laryngeal Intraepithelial Neoplasia
SIL: Squamous Intraepithelial Lesions

Table 1 - History of dysplasia classification

Table 2 - Cohort Description		
Characteristics	**Training and Validation sets**	**Gold Standard Sets (after consensus)**
Number of patients	498	110
Male	406 (81.5%)	86 (78.2%)
Female	92 (18.5%)	24 (21.8%)
Number of samples	740	128
Biopsies	485 (65.5%)	128 (100%)
Surgical resections	255 (34.5%)	0 (0%)
Anatomical localization		
Larynx	570	68
Pharynx	55	3
Nasopharynx	3	3
Hypopharynx	87	18
Lateral side of the tongue or oral cavity	63	7
Number of slides	2121	128
Biospies	1436 (67.7%)	128 (100%)
Surgical resections	685 (32.3%)	0 (0%)
Diagnosis (worst lesions on the slide)		
Benign (Negative for dysplasia or carcinoma)	545 (25.7%)	28 (21.8%)
Low Grade Dysplasia	229 (10.8%)	27 (21.1%)
High Grade Dysplasia	690 (32.5%)	33 (25.8%)
Invasive Carcinoma	657 (31.0%)	40 (31.2%)
Vital Status		
Alive	98 (19.7%)	20 (18.2%)
Dead	154 (30.9%)	36 (32.7%)
Default	246 (49.4%)	54 (49.1%)

Table 2 - Cohort description

Table 3 - Classification Performances - For each class, all the metrics are computed in a "one vs rest" manner: slides from the class are considered positives and slides from other classes are considered negatives. The average corresponds to the average over the 4 classes. Confusion matrices are shown in Figure 2. Confidence intervals are computed with bootstrapping (10 000 bootstraps). NPV corresponds to the Negative Predictive Value.

		AUC [95% CI]	NPV [95% CI]	Precision [95% CI]	Recall [95% CI]	Accuracy [95% CI]	Specificity [95% CI]	AUC (Precision/Recall) [95% CI]
AI vs Consensus Labels	**Average**	**0.886 [0.822-0.939]**	**0.895 [0.832-0.949]**	**0.655 [0.501-0.8]**	**0.646 [0.482-0.799]**	**0.84 [0.777-0.896]**	**0.897 [0.835-0.951]**	**0.713 [0.564-0.84]**
	Normal (0)	0.909 [0.848-0.958]	0.901 [0.84-0.954]	0.704 [0.52-0.867]	0.655 [0.476-0.826]	0.859 [0.797-0.914]	0.919 [0.863-0.969]	0.765 [0.6-0.889]
	Low Grade (1)	0.807 [0.718-0.883]	0.869 [0.798-0.931]	0.345 [0.171-0.524]	0.435 [0.227-0.643]	0.75 [0.672-0.82]	0.819 [0.743-0.89]	0.447 [0.252-0.648]
	High Grade (2)	0.859 [0.786-0.922]	0.856 [0.784-0.921]	0.645 [0.471-0.81]	0.588 [0.417-0.75]	0.805 [0.734-0.867]	0.883 [0.814-0.944]	0.69 [0.516-0.833]
	Carcinoma (3)	0.97 [0.936-0.995]	0.954 [0.907-0.989]	0.927 [0.841-1]	0.905 [0.806-0.979]	0.945 [0.906-0.984]	0.965 [0.92-1]	0.949 [0.89-0.99]
	Dysplasia (1+2)	0.889 [0.829-0.939]	0.838 [0.746-0.92]	0.767 [0.655-0.871]	0.807 [0.7-0.902]	0.805 [0.734-0.867]	0.803 [0.707-0.892]	0.872 [0.786-0.935]
AI vs Initial Labels	**Average**	**0.832 [0.763-0.893]**	**0.867 [0.798-0.928]**	**0.573 [0.413-0.731]**	**0.569 [0.411-0.726]**	**0.797 [0.729-0.861]**	**0.866 [0.796-0.929]**	**0.617 [0.48-0.743]**
	Normal (0)	0.916 [0.86-0.962]	0.901 [0.84-0.953]	0.704 [0.52-0.875]	0.655 [0.476-0.824]	0.859 [0.797-0.914]	0.919 [0.862-0.969]	0.753 [0.568-0.892]
	Low Grade (1)	0.706 [0.612-0.794]	0.808 [0.727-0.881]	0.276 [0.115-0.444]	0.296 [0.125-0.478]	0.688 [0.609-0.766]	0.792 [0.709-0.869]	0.316 [0.197-0.461]
	High Grade (2)	0.749 [0.66-0.831]	0.814 [0.735-0.888]	0.484 [0.31-0.667]	0.455 [0.286-0.629]	0.734 [0.656-0.813]	0.832 [0.755-0.903]	0.478 [0.312-0.647]
	Carcinoma (3)	0.958 [0.92-0.986]	0.943 [0.889-0.988]	0.829 [0.706-0.936]	0.872 [0.757-0.972]	0.906 [0.852-0.953]	0.921 [0.86-0.975]	0.92 [0.844-0.972]
	Dysplasia (1+2)	0.849 [0.779-0.91]	0.779 [0.676-0.875]	0.75 [0.635-0.855]	0.75 [0.635-0.857]	0.766 [0.695-0.836]	0.779 [0.677-0.873]	0.801 [0.678-0.908]

Table 3 - Classification performances

Table 4 - Inter rater agreements: Intrinsic noise in the dataset was assessed comparing agreements between different reviewers: reviewer 1, reviewer 2, the initial labels, and the AI model**.**

	Unweighted kappa [95% CI]	**Linear kappa [95% CI]**
AI vs Initial Labels	0.454 [0.339-0.561]	0.641 [0.546-0.726]
AI vs Reviewer 2	0.414 [0.308-0.52]	0.606 [0.511-0.692]
AI vs Reviewer 1	0.537 [0.427-0.642]	0.676 [0.585-0.757]
Reviewer 2 vs Reviewer 1	0.493 [0.394-0.599]	0.676 [0.592-0.753]
Reviewer 1 vs Initial Labels	0.518 [0.406-0.622]	0.689 [0.606-0.764]
Reviewer 2 vs Initial Labels	0.6 [0.498-0.701]	0.723 [0.634-0.803]

Table 4 - Inter rater agreement

Acknowledgements

Keen Eye team for thoughtful discussions and technical support. Yan Petit for code implementation assistance.

Conflict of interest

The authors declare no competing interests.

Ethics Approval and Consent to Participate

Our study was approved by the ethics committee of Assistance Publique - Hôpitaux de Paris Centre (CERAPHP. Centre - Institutional Review Board registration #00011928). All the patients were informed by a notification letter of the study and the possibility to refuse the use of their medical data, in line with current legislation. The study was performed in accordance with the Declaration of Helsinki.

Author Contributions

Concept and design: CB, TW, SB. Ethical approvals processes : YBH, CB, SB. Creation of the clinical and pathology database, slides selection, virtual slides quality control: YBH. Data management and processing, software implementation, experiments realization and statistical analysis: ML. Choices on experiments : ML and YBH. Design of AI methods: ML, TW. Results analysis: ML and YBH. Discussion of results: ML, YBH, TW, CB. Manuscript writing and review: all authors. All authors read and approved the final paper.

Funding

ML was supported by a CIFRE PhD fellowship founded by Keen Eye, Paris, France and ANRT (CIFRE 2019/1905). Furthermore, this work was supported by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

Data Availability Statement

The WSI dataset described in the manuscript were subject to hospital regulations and could not be publicly released.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. nov 2018;68(6):394-424.
2. Johnson DE, Burtness B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. Nat Rev Dis Primer. 26 nov 2020;6(1):1-22.
3. Liao LJ, Hsu WL, Lo WC, Cheng PW, Shueng PW, Hsieh CH. Health-related quality of life and utility in head and neck cancer survivors. BMC Cancer. 2019;19(1):1-10.
4. Mahmood H, Shaban M, Indave BI, Santos-Silva AR, Rajpoot N, Khurram SA. Use of artificial intelligence in diagnosis of head and neck precancerous and cancerous lesions: A systematic review. Oral Oncol. nov 2020;110:104885.
5. Kleinsasser O. The classification and differential diagnosis of epithelial hyperplasia of the laryngeal mucosa on the basis of histomorphological features. Z Laryngol Rhinol Otol. mai 1963;42:339-62.
6. Gale N, Cardesa A, Hernandez-Prera JC, Slootweg PJ, Wenig BM, Zidar N. Laryngeal dysplasia: persisting dilemmas, disagreements and unsolved problems—a short review. Head Neck Pathol. 2020;14(4):1046-51.
7. Hellquist H, Ferlito A, Mäkitie AA, Thompson LDR, Bishop JA, Agaimy A, et al. Developing Classifications of Laryngeal Dysplasia: The Historical Basis. Adv Ther. 1 juin 2020;37(6):2667-77.
8. Mehlum CS, Larsen SR, Kiss K, Groentved AM, Kjaergaard T, Möller S, et al. Laryngeal precursor lesions: Interrater and intrarater reliability of histopathological assessment. The Laryngoscope. oct 2018;128(10):2375-9.
9. Sarioglu S, Cakalagaoglu F, Elagoz S, Ersoy U, Etit D, Hucumenoglu S, et al. Inter-observer Agreement in Laryngeal Pre-neoplastic Lesions. Head Neck Pathol. 21 sept 2010;4(4):276-80.
10. Fleskens SAJHM, Bergshoeff VE, Voogd AC, van Velthuysen MLF, Bot FJ, Speel EJM, et al. Interobserver variability of laryngeal mucosal premalignant lesions: a histopathological evaluation. Mod Pathol. juill 2011;24(7):892-8.
11. Hu Y, Liu H. Diagnostic variability of laryngeal premalignant lesions: histological evaluation and carcinoma transformation. Otolaryngol Neck Surg. 2014;150(3):401-6.
12. Krishnan L, Karpagaselvi K, Kumarswamy J, Sudheendra U, Santosh K, Patil A. Inter-and intra-observer variability in three grading systems for oral epithelial dysplasia. J Oral Maxillofac Pathol JOMFP. 2016;20(2):261.
13. Van Hulst AM, Kroon W, van der Linden ES, Nagtzaam L, Ottenhof SR, Wegner I, et al. Grade of dysplasia and malignant transformation in adults with premalignant laryngeal lesions. Head Neck. 2016;38(S1):E2284-90.
14. Gale N, Blagus R, El-Mofty SK, Helliwell T, Prasad ML, Sandison A, et al. Evaluation of a new grading system for laryngeal squamous intraepithelial lesions—a proposed unified classification. Histopathology. 2014;65(4):456-64.
15. El-Naggar AK, Chan JKC, Grandis JR, Takata T, Slootweg PJ. WHO Classification of Head and Neck Tumours [Internet]. 4th éd. Vol. 9. 2017 [cité 18 mai 2021]. Disponible sur: https://publications.iarc.fr/Book-And-Report-Series/Who-Classification-Of-Tumours/WHO-Cl

assification-Of-Head-And-Neck-Tumours-2017
16. Zidar N, Gale N. Update from the 5th Edition of the World Health Organization Classification of Head and Neck Tumors: Hypopharynx, Larynx, Trachea and Parapharyngeal Space. Head Neck Pathol. 2022;16(1):31-9.
17. WHO. Female Genital Tumours WHO Classification of Tumours. 5th Edition. Vol. 4. 2020.
18. Roberts JR, Siekas LL, Kaz AM. Anal intraepithelial neoplasia: A review of diagnosis and management. World J Gastrointest Oncol. 2017;9(2):50.
19. WHO. Digestive System Tumours WHO Classification of Tumours. 5th edition. 2019.
20. Sathasivam HP, Sloan P, Thomson PJ, Robinson M. The clinical utility of contemporary oral epithelial dysplasia grading systems. J Oral Pathol Med. 2022;51(2):180-7.
21. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019;16(11):703-15.
22. Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama. 2017;318(22):2199-210.
23. Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol. 2018;42(12):1636.
24. Raciti P, Sue J, Ceballos R, Godrich R, Kunz JD, Kapur S, et al. Novel artificial intelligence system increases the detection of prostate cancer in whole slide images of core needle biopsies. Mod Pathol. 2020;33(10):2058-66.
25. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med. oct 2018;24(10):1559-67.
26. Courtiol P, Tramel EW, Sanselme M, Wainrib G. Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach. ArXiv180202212 Cs Stat [Internet]. 20 févr 2020 [cité 28 mai 2021]; Disponible sur: http://arxiv.org/abs/1802.02212
27. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol. 2020;21(2):233-41.
28. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat Commun. 2020;11(1):1-15.
29. Coudray N, Tsirigos A. Deep learning links histology, molecular signatures and prognosis in cancer. Nat Cancer. 2020;1(8):755-7.
30. Lu MY, Chen TY, Williamson DF, Zhao M, Shady M, Lipkova J, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature. 2021;594(7861):106-10.
31. Mahmood H, Shaban M, Rajpoot N, Khurram SA. Artificial Intelligence-based methods in head and neck cancer diagnosis: An overview. Br J Cancer. 2021;124(12):1934-40.
32. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. PMLR; 2016. p. 1050-9.
33. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inf Process Syst. 2017;30.
34. Osband I, Blundell C, Pritzel A, Van Roy B. Deep exploration via bootstrapped DQN. Adv Neural Inf Process Syst. 2016;29.
35. Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. Med Image Anal. 2020;59:101557.
36. Camarasa R, Bos D, Hendrikse J, Nederkoorn P, Kooi E, Lugt A van der, et al. Quantitative

comparison of monte-carlo dropout uncertainty measures for multi-class segmentation. In: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis. Springer; 2020. p. 32-41.
37. Di Scandalea ML, Perone CS, Boudreau M, Cohen-Adad J. Deep active learning for axon-myelin segmentation on histology data. ArXiv Prepr ArXiv190705143. 2019;
38. Kohl S, Romera-Paredes B, Meyer C, De Fauw J, Ledsam JR, Maier-Hein K, et al. A probabilistic u-net for segmentation of ambiguous images. Adv Neural Inf Process Syst. 2018;31.
39. Thagaard J, Hauberg S, Vegt B van der, Ebstrup T, Hansen JD, Dahl AB. Can you trust predictive uncertainty under real dataset shifts in digital pathology? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer; 2020. p. 824-33.
40. Senousy Z, Abdelsamea MM, Gaber MM, Abdar M, Acharya UR, Khosravi A, et al. MCUa: Multi-level context and uncertainty aware dynamic deep ensemble for breast cancer histology image classification. IEEE Trans Biomed Eng. 2021;69(2):818-29.
41. Pocevičiūtė M, Eilertsen G, Jarkman S, Lundström C. Generalisation effects of predictive uncertainty estimation in deep learning for digital pathology. Sci Rep. 2022;12(1):1-15.
42. Dolezal JM, Srisuwananukorn A, Karpeyev D, Ramesh S, Kochanny S, Cody B, et al. Uncertainty-Informed Deep Learning Models Enable High-Confidence Predictions for Digital Histopathology. ArXiv Prepr ArXiv220404516. 2022;
43. Westra WH. The morphologic profile of HPV-related head and neck squamous carcinoma: implications for diagnosis, prognosis, and clinical management. Head Neck Pathol. 2012;6(1):48-54.
44. Sechidis K, Tsoumakas G, Vlahavas I. On the stratification of multi-label data. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer; 2011. p. 145-58.
45. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234-41.
46. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. In: International conference on machine learning. PMLR; 2018. p. 2127-36.
47. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 4700-8.
48. Chung YA, Lin HT, Yang SW. Cost-Aware Pre-Training for Multiclass Cost-Sensitive Deep Learning. IJCAI. 2016;
49. Delaune A, Valmary-Degano S, Loménie N, Zryouil K, Benyahia N, Trassard O, et al. Le premier data challenge organisé par la Société Française de Pathologie: une compétition internationale en 2020, un outil de recherche en intelligence artificielle pour l'avenir? In: Annales de Pathologie. Elsevier; 2022. p. 119-28.
50. Chen T, Kornblith S, Norouzi M, Hinton G. A Simple Framework for Contrastive Learning of Visual Representations. 13 févr 2020; Disponible sur: https://arxiv.org/abs/2002.05709v3
51. Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977;159-74.
52. Wang CW, Liou YA, Lin YJ, Chang CC, Chu PH, Lee YC, et al. Artificial intelligence-assisted fast screening cervical high grade squamous intraepithelial lesion and squamous cell carcinoma diagnosis and treatment planning. Sci Rep. 2021;11(1):1-14.
53. Bao H, Bi H, Zhang X, Zhao Y, Dong Y, Luo X, et al. Artificial intelligence-assisted cytology for detection of cervical intraepithelial neoplasia or invasive cancer: a multicenter, clinical-based, observational study. Gynecol Oncol. 2020;159(1):171-8.
54. Tomita N, Abdollahi B, Wei J, Ren B, Suriawinata A, Hassanpour S. Attention-based deep

neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. JAMA Netw Open. 2019;2(11):e1914645-e1914645.