Journal of Pathology Informatics 13 (2022) 100149

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

Can AI predict epithelial lesion categories via automated analysis of cervical biopsies: The TissueNet challenge?

Nicolas Loménie [a,*], Capucine Bertrand [b], Rutger H.J. Fick [b], Saima Ben Hadj [b], Brice Tayart [c], Cyprien Tilmant [d], Isabelle Farré [e], Soufiane Z. Azdad [f], Samy Dahmani [f], Gilles Dequen [g], Ming Feng [h], Kele Xu [h], Zimu Li [h], Sophie Prevot [i], Christine Bergeron [j], Guillaume Bataillon [k], Mojgan Devouassoux-Shisheboran [l], Claire Glaser [m], Agathe Delaune [n], Séverine Valmary-Degano [o], Philippe Bertheau [p]

[a] *LIPADE, UFR Mathématiques-Informatiques, Université Paris Cité, 45 rue des Saints-Pères, 75006 Paris, France*
[b] *Tribun Health, Courbevoie, France*
[c] *Idemia, Courbevoie, France*
[d] *GHICL, Lille, France*
[e] *XPath Nord, Leulinghem, France*
[f] *Algoscope, 9 rue Gaspard Monge, 60200 Compiègne, France*
[g] *Laboratoire Modélisation, Information, Systèmes (MIS), Université de Picardie Jules Verne, 80080 Amiens, France*
[h] *Tongji University, Shanghai, China*
[i] *Pathologie, CHU Bicêtre, APHP, 78 Rue du Général Leclerc, 94270 Le Kremlin-Bicêtre, France*
[j] *CerbaPath, 16 bis rue Odessa, Paris 75014, France*
[k] *Pathologie, Institut Curie, 26 rue d'Ulm, 75005 Paris, France*
[l] *Centre de Pathologie Sud des Hospices Civils de Lyon, Centre Hospitalier Lyon Sud, 165 chemin du grand Revoyet, 69495 Pierre Bénite Cedex, France*
[m] *Pathologie, CHG Versailles, 177 Rue de Versailles, 78150 Le Chesnay-Rocquencourt, France*
[n] *Plateforme de données de santé - Health Data Hub, 9 rue Georges Pitard, 75015 Paris, France*
[o] *Pathologie, Université Grenoble Alpes, Inserm U1209, CNRS UMR5309, Institute for Advanced Biosciences, CHU, Grenoble 38000, France*
[p] *Pathologie, CHU Saint-Louis, APHP, Université Paris Cité, 1 avenue Claude Vellefaux, 75010 Paris, France*

ARTICLE INFO

Keywords:
Artificial intelligence
Data challenge
Whole slide images
Uterine cervix

ABSTRACT

The French Society of Pathology (SFP) organized its first data challenge in 2020 with the help of the Health Data Hub (HDH). The organization of this event first consisted of recruiting nearly 5000 cervical biopsy slides obtained from 20 pathology centers. After ensuring that patients did not refuse to include their slides in the project, the slides were anonymized, digitized, and annotated by expert pathologists, and finally uploaded to a data challenge platform for competitors from around the world. Competing teams had to develop algorithms that could distinguish 4 diagnostic classes in cervical epithelial lesions. Among the many submissions from competitors, the best algorithms achieved an overall score close to 95%. The final part of the competition lasted only 6 weeks, and the goal of SFP and HDH is now to allow for the collection to be published in open access for the scientific community. In this report, we have performed a "post-competition analysis" of the results. We first described the algorithmic pipelines of 3 top competitors. We then analyzed several difficult cases that even the top competitors could not predict correctly. A medical committee of several expert pathologists looked for possible explanations for these erroneous results by reviewing the images, and we present their findings here targeted for a large audience of pathologists and data scientists in the field of digital pathology.

Introduction

The French Society of Pathology and the Health Data Hub set up a challenge in digital histopathology published on DrivenData in 2020.[14] The data for this challenge includes thousands of microscopic slides of uterine cervical tissue from medical centers across France. The objective was to classify each image according to the most severe category of epithelial lesion present in the sample. The classes are defined as follows according to the WHO classification:

- 0: benign (normal or subnormal),
- 1: low malignant potential (low grade squamous intraepithelial lesion),
- 2: high malignant potential (high grade squamous intraepithelial lesion),
- 3: invasive cancer (invasive squamous carcinoma).

[*] Corresponding author.
E-mail address: nicolas.lomenie@u-paris.fr (N. Loménie).

http://dx.doi.org/10.1016/j.jpi.2022.100149
Received 28 July 2022; Received in revised form 22 September 2022; Accepted 23 September 2022
Available online 05 October 2022
2153-3539/© 2022 The Author(s). Published by Elsevier Inc. on behalf of Association for Pathology Informatics. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Data

This dataset consists of high resolution images of microscopic slides created from cervical biopsies. Additionally, the competitors were given slide metadata as well as annotations for the training set that outlined some (but not necessarily all) of the lesions present on a slide.

The training set images were provided in 3 formats: native whole slide image formats (720 GB), pyramidal TIFs (928 GB), and down-sampled JPEGs (752 MB). The organizers provided just the annotated regions of each slide at full resolution to further support prototyping. The training folder contained images in their native digital pathology format, known as whole slide images (WSIs used as digital representations of a microscopic slide at high levels of magnification and at extremely high resolution in size (e.g., 150 000 × 85 000 pixels)). There are 4 native WSI image formats in the dataset across the different centers: .mrxs (3D Histech)/.svs (Leica)/.ndpi (Hamamatsu)/.tif (pyramidal TIF). The tif images folder contained the training images as a standardized set of compressed images in pyramidal TIF format. These images are compressed using JPEG Q = 75. The pyramidal TIF format maintained a sufficient level of detail for pathologists to perform diagnoses while enabling smaller file sizes and easier loading with actively developed Python libraries such as *PyVips*.

The downsampled images folder contained JPEG versions of the images that have been downsampled by 32x. While these images likely do not contain sufficient information for diagnosis at this resolution, they may be particularly helpful for prototyping model pipelines. The annotated regions folder contained JPEG versions of the annotated portions of the slides at *full resolution*. The annotated regions were defined by geometry in a specific csv file.

Annotations

The images are very large and contain a mixture of pathological and otherwise normal tissue. Pathologists have annotated images to point out regions that represent lesions. The train_annotations.csv contained the following information: a unique identifier for each annotation, the slide image filename corresponding to each annotation (each image containing multiple annotations), (x, y) coordinates of the annotation in WKT format (all geometries being closed rectangles and assuming origin is bottom left corner of the image), the class of the annotated tissue, the file locations of the annotated region jpeg in the public S3 (Simple Storage Service on the cloud) bucket in different regions of the world in the USA, Europe, and Asia.

Annotations were only provided for the training set and were not provided for the test set. When working with the annotations, it was important to keep in mind the following points:

- The annotated regions *do not necessarily* include all lesions in the slide. An unannotated region is not necessarily normal tissue.
- The *whole image* class label and the *annotation* class label do not necessarily match. The annotated regions may be the image's labeled class or below. For instance, an image labeled as a class 2 lesion could have annotations representing class 0, 1, or 2. At least some of the annotated regions represent the most severe/labeled class. All annotations on a slide with label 0 is considered as normal tissue.
- The lesion may fall entirely within the square, or may extend beyond the annotation boundaries.
- All annotations are a fixed size of 300 x 300 micrometers. As images have different resolutions in pixels/micrometer, annotations will have different dimensions in terms of pixels.

Aims and objectives of the challenge

The fields of application of artificial intelligence (AI) are expanding every day, particularly in the world of healthcare, a promising field given the mass of data generated for diagnosis and care.[1,2] Whether it is for intelligent patient monitoring using connected devices, for choosing the best treatment for a patient based on several sources of diagnostic data, or for optimizing analytical processes in biology or imaging, applications are multiplying, driven by multiple players, academic, and/or industrial. Many medical devices using AI have already been approved in the United States and Europe.[4]

Advances in the field of image analysis in general and in particular in medical imaging[5] are driven by the ability of machine neural networks to classify images after a more or less supervised learning phase, but sometimes also without prior learning, highlighting links between images and the prognosis of a disease or the response to a treatment. In pathology, promising applications[6] and guidelines on AI in pathology[7,8] are starting to emerge.

Developing an AI algorithm in healthcare requires strong computational and statistical expertise, but also quantity and quality of data, accounting for as many as possible of all real-life situations.[1] Access to representative and richly annotated health data is currently the main limiting factor for the development of AI algorithms.

A data challenge is a real opportunity to bring together both multiple data science expertise and a large collection of quality data. On the one hand, organizing a competition with a reward can mobilize a whole community of developers and data scientists. On the other hand, this large mobilization of a scientific community stimulates upstream the involvement of data producers. It is the conjunction of these 2 dynamics that might lead to the best algorithms. Even if the algorithms developed during a data challenge still have many steps to go through before being validated for use in medical practice, they already provide the first proof of concept of their potential interest.

Data challenges on medical and clinical images have already been organized in recent years. We can mention those of French radiologists since 2018,[9,10] and the Camelyon 16 and 17 in pathology.[11] A recent review paper reported several data challenges in pathology and what these competitions brought to this speciality.[12]

For the 2020 Data Challenge organized by the French Society of Pathology (SFP) and the Health Data Hub (HDH), the prospect of an international competition with winner announcement at the French National Congress of Pathologists "Carrefour Pathologie" in 2020 helped mobilize the pathology community to produce a large database of cervical lesions. In this article, we detail the construction and progress of this competition and above all discuss the results by a retro-analysis of the errors made by the winners and its prospects.

For this first edition of the SFP-HDH data challenge, a classification question was chosen to assist the pathologist in the diagnosis of cervical epithelial lesions. This disease area was chosen because its high frequency in diagnostic practice and because it has not yet been studied from the point of view of a diagnostic approach by AI.

The question submitted to the algorithms for this data challenge was therefore their ability to classify cervical biopsy slides into one of the four WHO diagnostic categories,[13] from "normal" to invasive carcinoma, through low- or high grade lesions. The algorithms were asked to identify on each slide the most severe diagnostic class.

Cases and methods

Expert annotation of cases

The virtual slides were uploaded on a server of the HDH in order to be annotated by the expert pathologists. The software used to visualize and annotate the virtual slides was the open-source software Cytomine.org (https://cytomine.be/) which was installed on the HDH servers, after some developments by Cytomine to adapt it to the annotation process by the pathologists.

4934 slides were included in the project and put online on the HDH servers, including 3709 cervical biopsies and 1225 surgical samples limited to a cone-shaped portion of the cervix called conizations. These conization slides were not included in the final competition but might be used later to enrich the image bank available to researchers. Among the 3709 biopsy slides, 2542 were reviewed by the 5 expert pathologists (SP, CB, MD, GB, and CG) who indicated the most severe diagnostic class present on each slide.

Following this diagnostic annotation phase, the 2542 slides were distributed in 3 distinct groups: 1015 slides for the learning set, 513 slides for the test set, and 1014 slides for the final validation set, thus avoiding that competitors overfit on a single test set after multiple submission. For each set of slides, the proportion of each of the 4 diagnostic classes had to be as close as possible to 25% in order to balance the classes.

On the 1015 slides of the training set, the experts then added more precise visual annotations: 300-micron squares were placed not only on the most severe diagnostic class lesions, but also on less severe diagnostic class lesions if present, with different colored squares used to distinguish the 4 diagnostic classes (Fig. 1). A maximum of 10 squares for a single class were deposited on each slide, summing up to 5941 visual annotations on these 1015 slides.

All 2542 slides annotated for the competition were reviewed successively by 2 expert pathologists. Some cases required the 2 pathologists to agree during consensus meetings, leading to the exclusion of some cases considered too difficult or of insufficient technical quality. The 1015 finely annotated slides constituted the training set of the data challenge, the visual annotations deposited by the experts being used as references of the 4 lesion classes by the competitors. The 2048 unannotated slides were also made available to the competitors as additional learning data if they wished. Some slides were deleted during the annotation phase because they had defects that made their analysis impossible.

The training slides were analyzed and exploited to feed the algorithms with the competitors' computing resources. These algorithms could then be tested on the set of 513 test slides, allowing each competitor to have an idea of its degree of progress compared to the other competitors thanks to a ranking on an "intermediate leader-board" available during the whole challenge timeline. The final ranking, displayed on the "final leader board", was calculated on the set of 1014 validation slides. Competitors submitted their algorithm on the Driven Data platform as a package and the calculations for the test and final validation set were performed directly with the platform's graphics processors (GPU, Graphics Processing Unit).

Three of the competitors put on the challenge website some elements describing their solution, accessible at.[14] Their algorithms and software have been opened in open source and deposited on GitHub at.[15]

We present the algorithmic pipelines and results of the 3 most promising contestant's pipelines according to clinical criteria of explainability and reproducibility in the context of clinical routine. Here unfolds a description of each of these 3 contestant's methods:

1. Lifeis2Short from Tongji University, Shanghai, China
2. Algoscope, Compiègne, France
3. Tribun Health, Paris, France

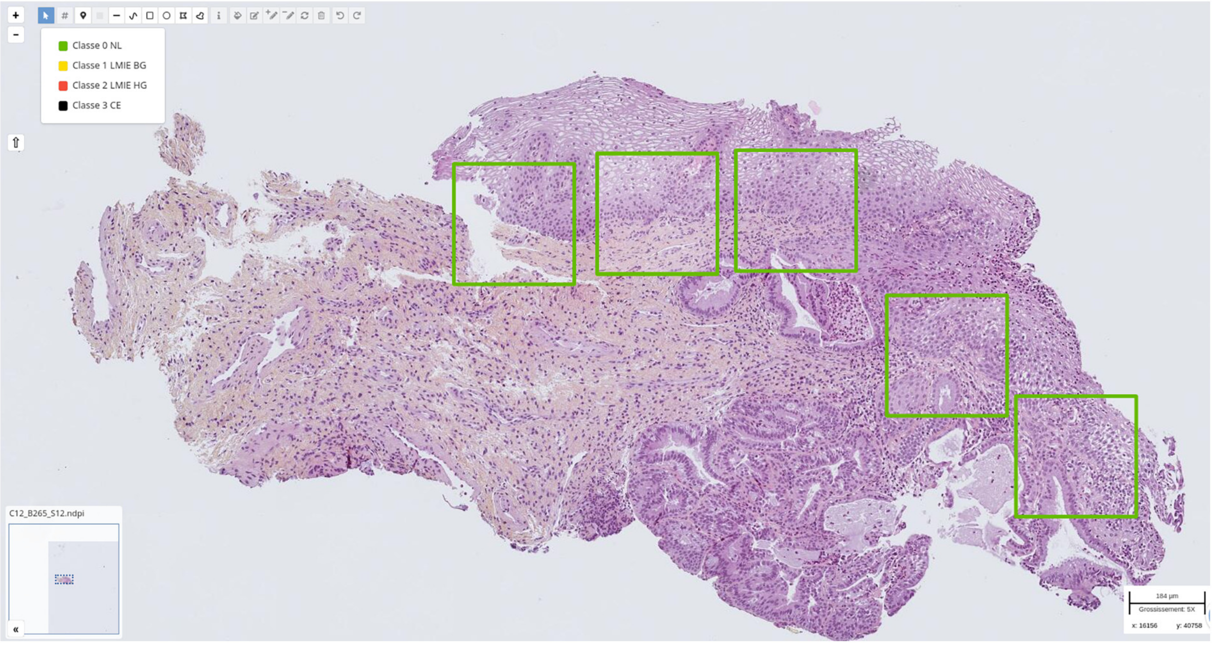
From a legal and ethical perspective, this data challenge falls within the French regulatory framework of the Reference Methodology n°004, which has been declared to the French Data Protection Authority (CNIL). The entire visualization of certain slides for the unique needs of the publication of the results of the research project is carried out in compliance with this regulatory framework and the de-identification of patients.

Lifeis2Short processing pipeline and WSI analysis results

Methodology

Fig. 2 shows the outline of the method pipeline. First, they trained a patch level classification neural network. The official pre-extracted patches have different sizes, considering the effect of different scales, they extracted a 320*320 size patch on level 2 (¼ size of level 0) according to the coordinates. After training, they used the Otsu thresholding algorithm to extract the tissue area, then extracted the patches with a sliding window with a step size of 256 on the Level 2 tissue area and generated the probability map of each category. They extracted the designed features from each probability map separately, concatenated them together, and used machine learning models such as Xgboost, LightGBM, and Random Forest for classification. The main contribution of the pipeline is the optimization of every step in the algorithm.

First of all, as shown in Fig. 3, they found that the original Otsu-like tissue region extraction algorithm is not applicable to all data. Firstly, the blank areas on the whole slide image may affect the results. Therefore, they added some judgments and filtering thresholds when generating the tissue map, and achieved good tissue segmentation results on all the


🟩 Classe 0 NL
🟨 Classe 1 LMIE BG
🟥 Classe 2 LMIE HG
⬛ Classe 3 CE

C12_B265_S12.ndpi

184 µm
Grossissement: 5X
x: 16156 y: 40758

Fig. 1. An example of a slide with annotations for the training on the Cytomine Server (available at https://cytomine.be/ as an open-source rich internet application for collaborative analysis of multi-gigapixel images). (All figures in this article must be seen in color).

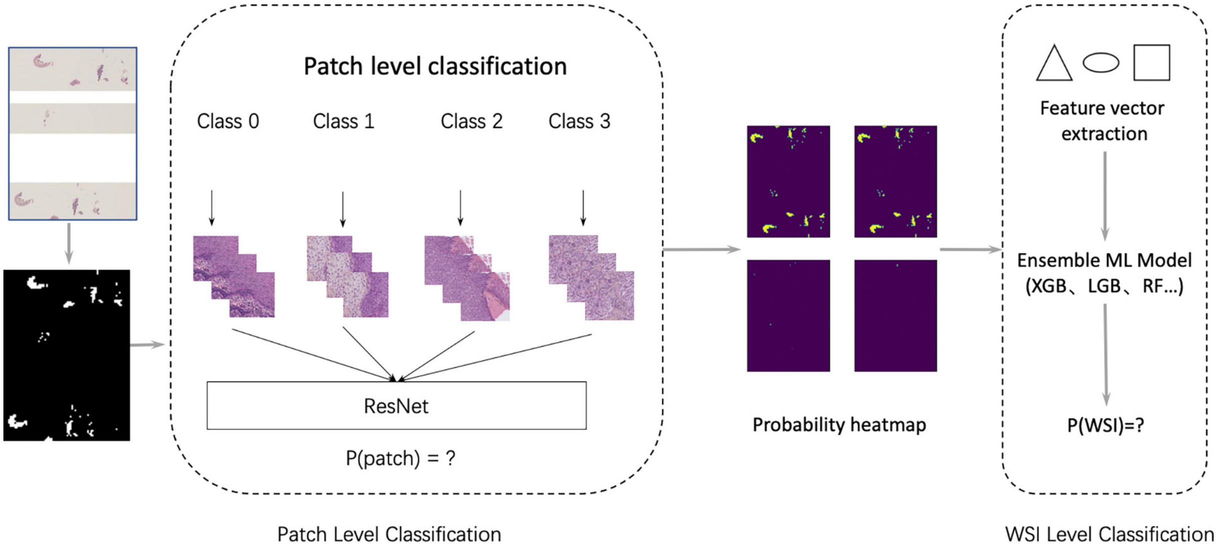

Patch level classification
Class 0
Class 1
Class 2
Class 3
ResNet
P(patch) = ?
Patch Level Classification
Probability heatmap
Feature vector extraction
Ensemble ML Model (XGB、LGB、RF...)
P(WSI)=?
WSI Level Classification

Fig. 2. Lifeis2Short team methodology.

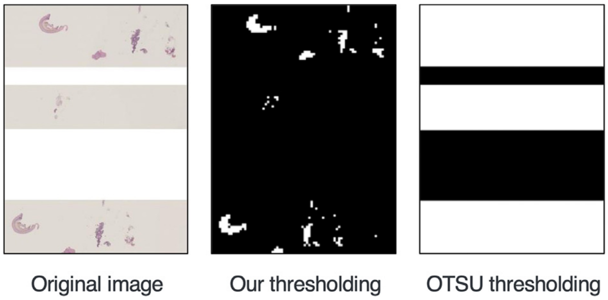


Fig. 3. Thresholding.

images. Secondly, the tissue area might be too small, causing the algorithm to extract the tissue mask on the low-level image and fail to recognize it. They made a comparison on different levels, and finally designed an adaptive level region extraction algorithm. The effect is shown in the Fig. 4.

Besides, they found out that the difficulty of the task was mainly to distinguish between 0 and 1 subtype. The information that can be obtained from labeled patches of these 2 subtypes is limited, in order to ensure that the model was correct, they needed to learn the discriminative features of the 2 subtypes. They extracted a large number of 0 patches from the labeled 0 WSI to make the distribution of 0 subtype closer to the true distribution. After adding a large number of class 0 type patch to the training data, the result has been greatly improved, from 89% to 92%. In addition, they found out that using random sampling patch as shown in Fig. 5 provided better results than grid sampling, and the local verification score of random cropping was 95%. The local verification score for grid clipping was 93%.

They also tried different classification network architectures. Eventually, DenseNet201 achieved good results, possibly because DenseNet has fewer parameters and can avoid overfitting. In addition, they also tried the more recent EfficientNet, and they obtained better results than Densenet201, but due to virtual environment problems, DenseNet201 was still used in the final result.

After integrating all the above improvements, the classification model score reached 0.9306 on the public leader-board and 0.9332 on the private leaderboard. It is worth noting that there was an error in one of their submissions, i.e., different image sizes were used for inference and training. As a result, the 0 categories of the training set on whole slide image level were all classified as category 1 in their local verification. This is obviously a bad model. It has also achieved very poor results on the public list, but it has greatly improved on the private list.

In addition, they also tried some new improvements, such as pseudo-label to use unlabeled data. However, due to the limited number of submissions, they did not submit this to the system. But they believe that it can yields better results.

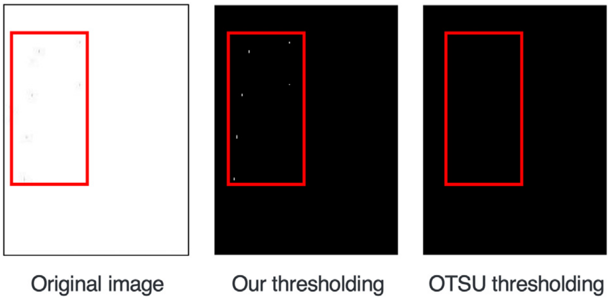


Fig. 4. Comparison results.

WSI analysis results

Their algorithm is hereby explained over a few slides according to the contestant's choice:

- C12_B091_S12,
- C12_B108_S12 (from the class 3 training set),
- C12_B098_S12 (from the class 0 test set predicted class 1),
- C11_C046_S11 (with a reference to C08_B023_S08 further discussed in the Discussion section).

- Incorrect labeling. There are mislabels in the labeled patch dataset. For example, a patch which is actually 0 (white area) and labeled as 2 (for instance in the slide C12_B528_S12). This one is obviously an error of annotation. Being not pathologists, it was difficult for them to assess how many incorrectly labeled data in the dataset.
- Data distribution. There are 2 steps in the method, the first step is patch level classification, and the second step is WSI level classification. According to the experimental results, the first step of patch classification is particularly important. They found out that if some 0-category patches are sampled from the labeled 0 WSI data as a supplement to the original patch annotation data, the results will be greatly improved. It shows that it is difficult for a given 6000 labeled patches to cover the actual

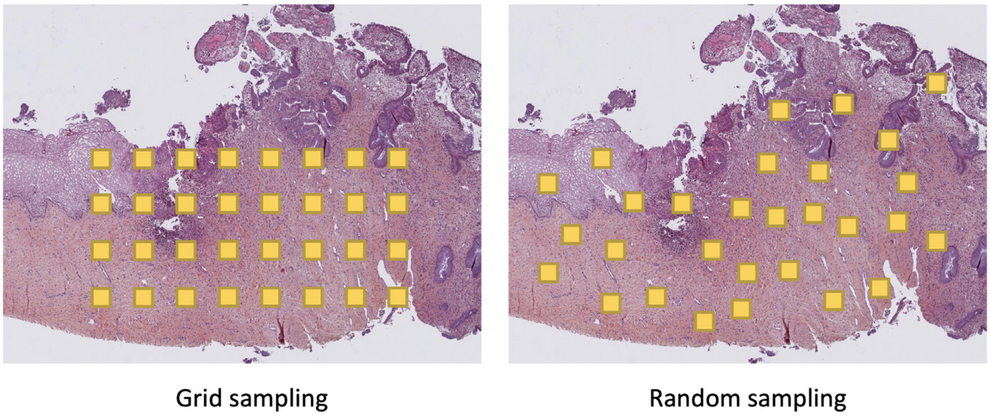


Fig. 5. Grid sampling.

data distribution of each category. This is the main reason for the poor classification effect. According to our experimental results, class 0 and class 1 have similar data distributions and they are particularly difficult to distinguish. They hence think that this is the difficulty of this competition.

- Picture granularity. Take a specific patch in WSI C12_B091_S12 as an example (see Fig. 6). In our re-test experiment, the label is 2 and the prediction is 0. We can see in Fig. 6 that the left subgraph should be category 2, and the right subgraph should be category 0. They hypothesize that the classification model considers the main features, i.e., there are more areas in class 0 than in class 2 on this image. But if images are annotated on 2 finer sub-images on the left and right the ambiguity of the classification can be greatly reduced.
- Hard mining with difficult classification data. Their model does not perform hard mining on the data, i.e., mining hard-to-divide samples. They that add weight to difficult samples that are incorrectly classified by the model will result in better results.

They also found out that the patch classification model predicted some patches of category 3 as category 0, such as C12_B108_S12 from the training set (See Fig. 7 (left) and (right)). For C08_B023_S08 and C11_C046_S11, they have also extracted images with similar surface features (See Fig. 8 (a) and (b) localized in the slide (c) and (d)). The method predicts some patches of category 3 as category 0 in C08_B023_S08 and C11_C046_S11. This led to wrong predictions at the WSI level. In short, the reason is that their model did not learn and cover the feature of the misclassified image.

On the whole, this competitor believes that the classification error is due to the limited and sometimes not completely correct annotations that cannot cover the entire data distribution, making the model prone to over-fitting on the limited annotation data. Therefore, the team extracted a large number of patches of category 0 in order to improve the classification performance, stressing out that it is very important to dig out the distribution patterns of each category from original data.

Algoscope processing pipeline and WSI analysis results

The team led by a pathologist tried to tackle the challenge by following a “business-oriented” approach. They developed a solution that tries to mimic the pathologist approach when diagnosing a cervical biopsy. By doing this way, it was hence possible to explain precisely to the pathologist the final score of the algorithm. In a few words, the algorithm extracted all regions of interest from the whole slide that characterize a lesion (Low grade, High grade, or Invasive carcinoma), assigned them to 1 of these 3 classes with a probability score, and then return a global score to the whole slide image.

A set of annotations was provided with the training set, but unfortunately, the size of the batch was not suitable. As mentioned in the problem description: “The annotated regions may be the image’s labeled class or below”. For example, annotations associated with a high grade lesion slide could contain characteristics of low- and high grade lesion. So, the first thing they did was to ask the pathologist to make a new set of annotations that were divided into the 4 mentioned classes (hundreds of image tiles for each class). Then, they trained this set of images on several

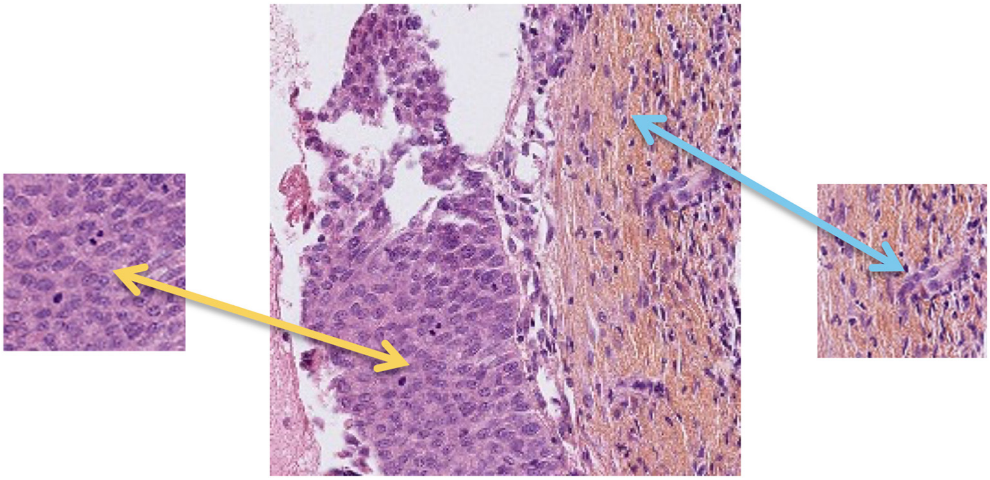


Fig. 6. A specific patch image in the slide C12_B091_S12 labeled 2 but predicted 0.

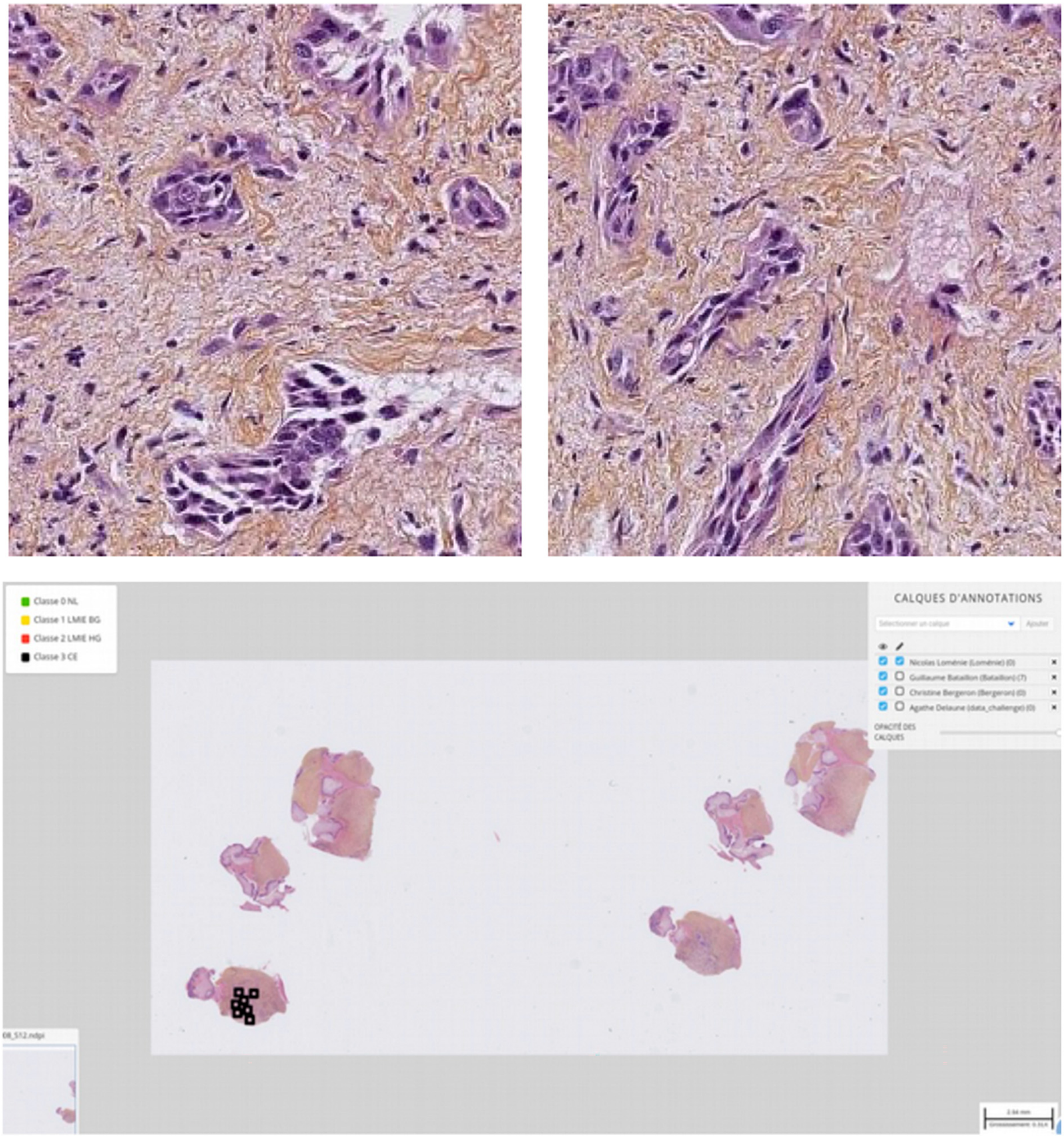


Fig. 7. Patches from slide C12_B108_S12 (bottom) with annotated patches from the experts in the training set (top left) C12_B108_S12_2 and (top right) C12_B108_S12_3.

open-source pre-trained Deep Neural Networks (Resnet-50, Google Inception, etc). Finally, they used the best output model for inference of tiles containing tissue within the whole slide.

Methodology

The complete algorithm for classifying a whole slide image can be described as follows:

```
1. Create 3 empty lists for each lesion class
(Low grade, High grade and Invasive carcinoma)
2. Loop over tiles of the slide at zoom x2.5
3. For each tile:
    a. Convert it to a grayscale image.
    b. If 99% of pixels are greater than a white threshold or less than   a black threshold
       (i.e.   tile image is almost totally white or dark, so it does not contain tissue), then
       ignore it.
    c. Otherwise, subdivide this tile in several sub-tiles at zoox5.
    d. For each sub-tile:
        i. Detect if it contains tissue (same algorithm at 3.b). If not, ignore it.
        ii. Otherwise, infer the trained model on the original sub-tile. If one of the 3 lesion types is
            detected with a probability greater than 70%, then add the sub-tile coordinates to the
            associated list.
4. Loop over the 3 lesion lists:
    a. If all lists are empty, then assign the class 0 (normal, benign) to the image
    b. Otherwise, if Invasive cancer list contains at least one element, then assign the class 3
       (Invasive cancer) to the slide
    c. If Invasive carcinoma list is empty and High grade list contains at least one element, then
       assign the class 2 (High grade) to the slide
    d. Otherwise, if only Low grade lesion list contains elements, then assign the class 1 (Low
       grade) to the slide
```

Algoscope was granted access to the HPC resources of IDRIS (www.idris.fr): training was done using an Nvidia Tesla V100 installed on the Jean Zay super calculator (provided by GENCI - https://www.genci.fr to

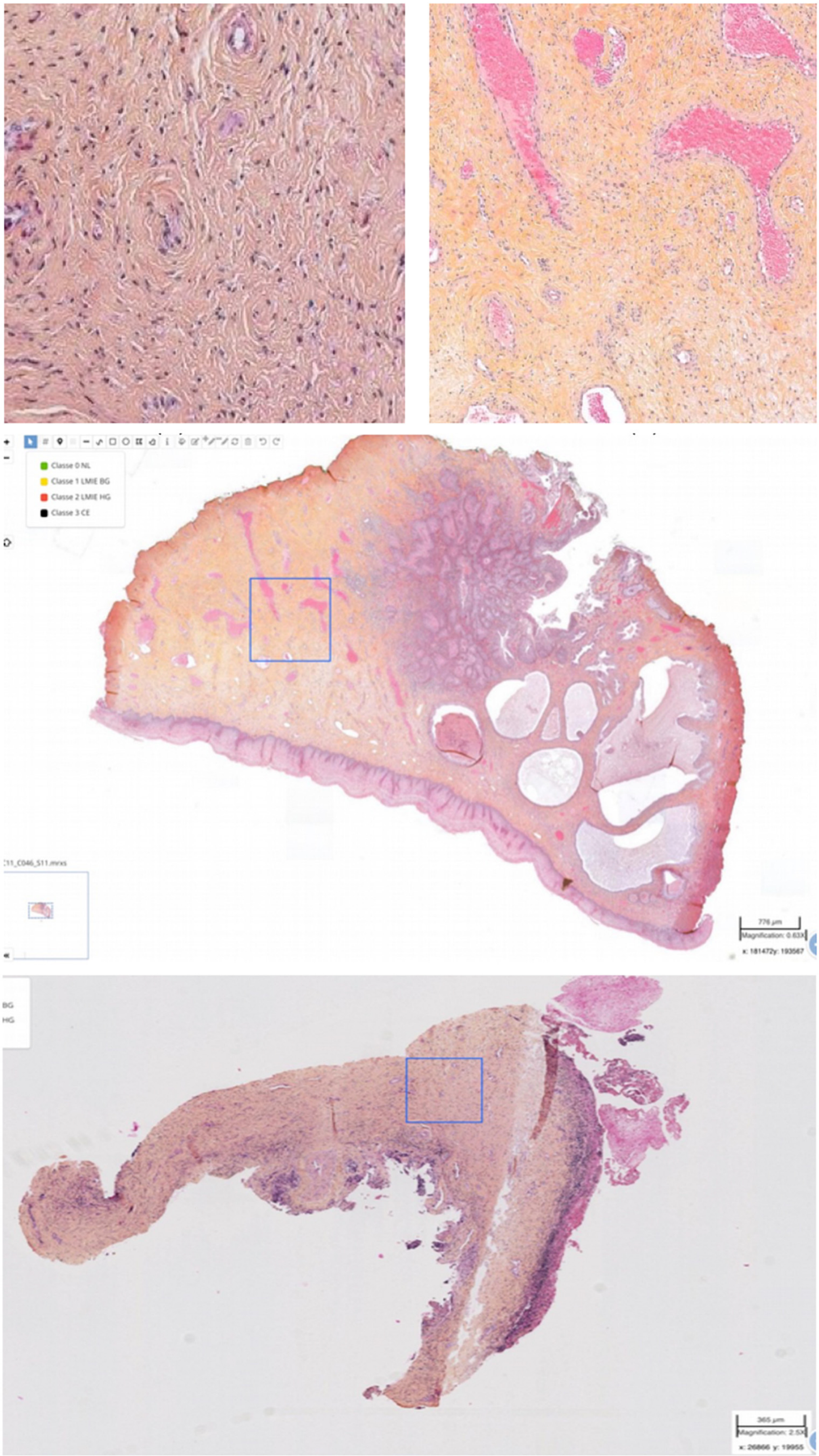


Fig. 8. From top to bottom: 2 patches from the original slide C11_C046_S11 (middle) with the location of the top right patch and the original slide C08_B023_S08 (bottom) with the location of the top left patch in the slide (see Discussion in Data challenge results section).

IDRIS). About 15 h were necessary to train the model. The following Python libraries were used: TensorFlow 2.2 (for DNNs) / PyVips (for reading TIFF images) / PIL (for image processing).

On a laptop equipped with an Nvidia GTX 1060 and Intel(R) Core(TM) i7-8750H CPU @ 2.20 GHz, the mean average duration for classifying a whole slide image with this algorithm is about 20 s.

The team believes that the efficiency of the algorithm could be greatly improved by adding images during the training phase. Since they manually labeled the dataset and the pathologist was not completely dedicated to this task, they were running out of time, so they decided to train neural networks on only a few hundreds of tile images for each class. They also consider some augmentation data to improve robustness.

WSI analysis results

Their algorithm is hereby explained over a couple of slides according to the contestant's choice:

- C04_B052_S04
- C12_B438_S12
- C06_B040_S21

The goal was to understand why it failed to give the correct score on those specific cases, and how they could improve the solution.

WSI C04_B052_S04:

Experts scored this slide in class 0. The submitted algorithm detected 2 zones identified as class 3 (see Fig. 9). The review of this misclassified area by pathologists experts showed that it consisted mostly of inflammatory cells. Nevertheless, the probability (Index of confidence of the algorithm) for this classification was relatively low (≈ 76.5%) compared to zones identified within most of other slides (≈ 95%). Therefore, increasing this threshold (e.g., 85% instead of current 70%) could improve the efficiency of the solution and eliminate some false-positive class 3 zones.

WSI C12_B438_S12:

Experts scored this slide in class 0 (see Fig. 10 for a global view). The output of the algorithm consists of some tiles labeled as class 1 (Blue, Fig. 11), and some as class 3 (Red, Fig. 12). So the given final score was therefore class 3.

They also put here some tiles of C06_B040_S21 slide (score 0 according to experts) that were considered as class 3 by their algorithm (see Fig. 13).

The competitor compared those tiles with some expert annotations from public dataset, and it seems that at this magnification (zoom x5) it is difficult to make a correct classification for some specific cases. Any doubt quickly vanished when zooming to a higher magnification (zoom x10). The reason they did not exploit tiles at this high magnification was to

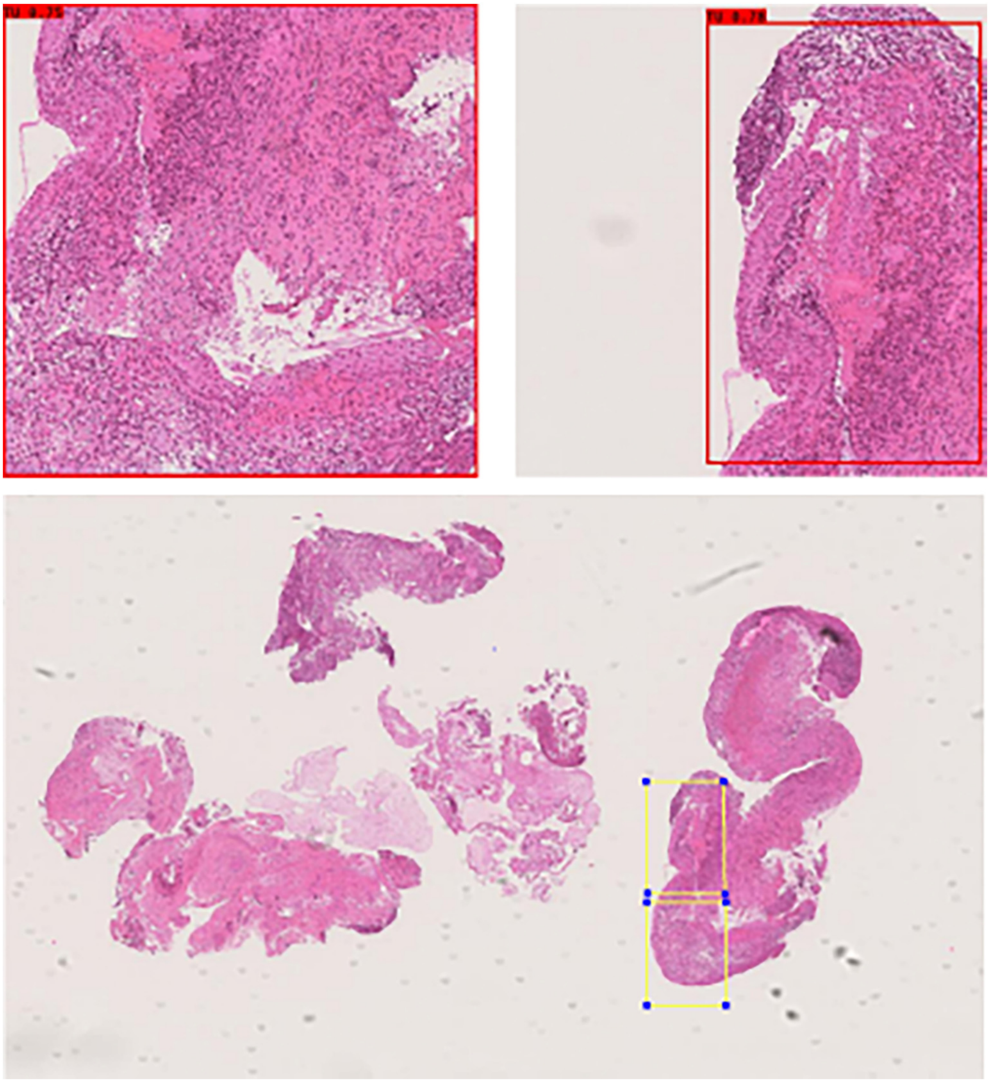


Fig. 9. Class 0 slide misclassified as class 3 on WSI C04_B052_S04 with details above of class 3 areas (See Supplementary Material to dig into the images at better resolution in Algoscope folder).

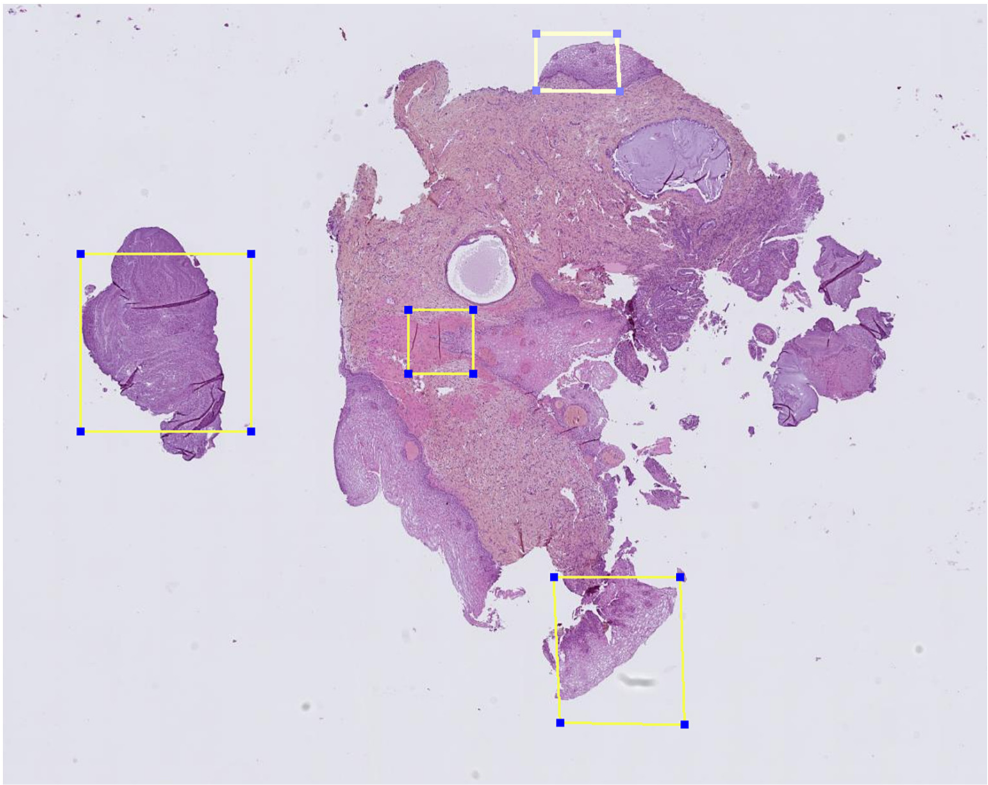


Fig. 10. Misclassified as class 3 instead of class 0 on WSI C12_B438_S12 (See Supplementary Material to dig into the images at better resolution in Algoscope folder and Figs. 11 and 12 hereinafter for details).

save processing time, since the proposed solution had to process many slides in a very short time. An easy way to overcome this problem would be to re-process the only tiles tagged with abnormalities (class 1, 2, or 3) at higher magnification.

To sum up, the competitor concluded that its algorithm's errors are "humanly understandable" by pathologists: spongiosis mistaken for a low grade lesion (class 1); tangential cut artifact or benign squamous metaplasia colonizing the endocervical glands mistaken for an invasive carcinoma (class 3). All these challenging areas of the slide were confirmed by the panel of expert pathologists. These errors could be made by a junior resident, and it shows that somehow their algorithm achieved the goal of adopting the most faithful as possible approach to the real practice of pathologists.

Tribun health processing pipeline and WSI analysis results

Methodology

Tribun health end-to-end processing pipeline consists of 4 steps, as shown in Fig. 14.

Tissue detection. They color-normalized the input WSI and segment the foreground tissue mask at image level 6 (16 μm/pix), which offers a compromise between the execution time and the detection quality. Color normalization ensures the WSI background to be white (no absorption). They corrected this by selecting a frame close to the outer edge of the scanned area - where no tissue is expected - and take the median of each channel. After normalization, they used Otsu's method to detect tissue areas,

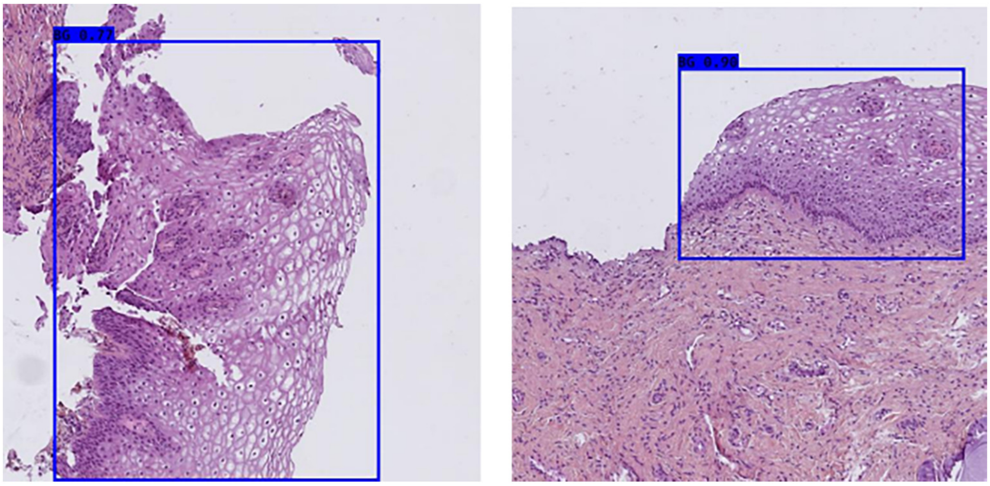


Fig. 11. Details of class 0 classification on WSI C12_B438_S12, normal or spongiotic epithelium.

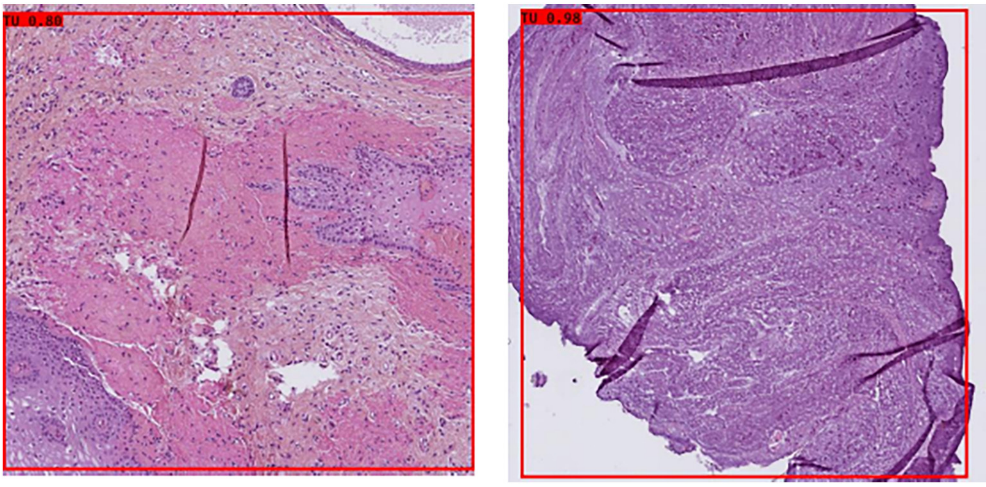


Fig. 12. Details of area misclassified as class 3 on WSI C12_B438_S12, squamous metaplasia of endocervical gland on the top panel, mucus on the bottom panel.

followed by some morphological filtering opening/closing to clean small defects in the mask.

Multi-resolution, ensemble-based patchwise classification. They used a multi-resolution ensemble CNN to predict healthy and CIN1-3 class probabilities per patch.

Patch resolution. Accurate CIN classification requires both sufficiently high resolution and enough context on the full thickness of the epithelium. Therefore, their ensemble CNN uses a range of patch sizes and resolutions: 256 × 256 at level 3 (2 μm/pix, high resolution, less context), 256 × 256 at level 4 (4 μm/pix, lower resolution, more context) as well as 384 × 384 at 2 μm/pix.

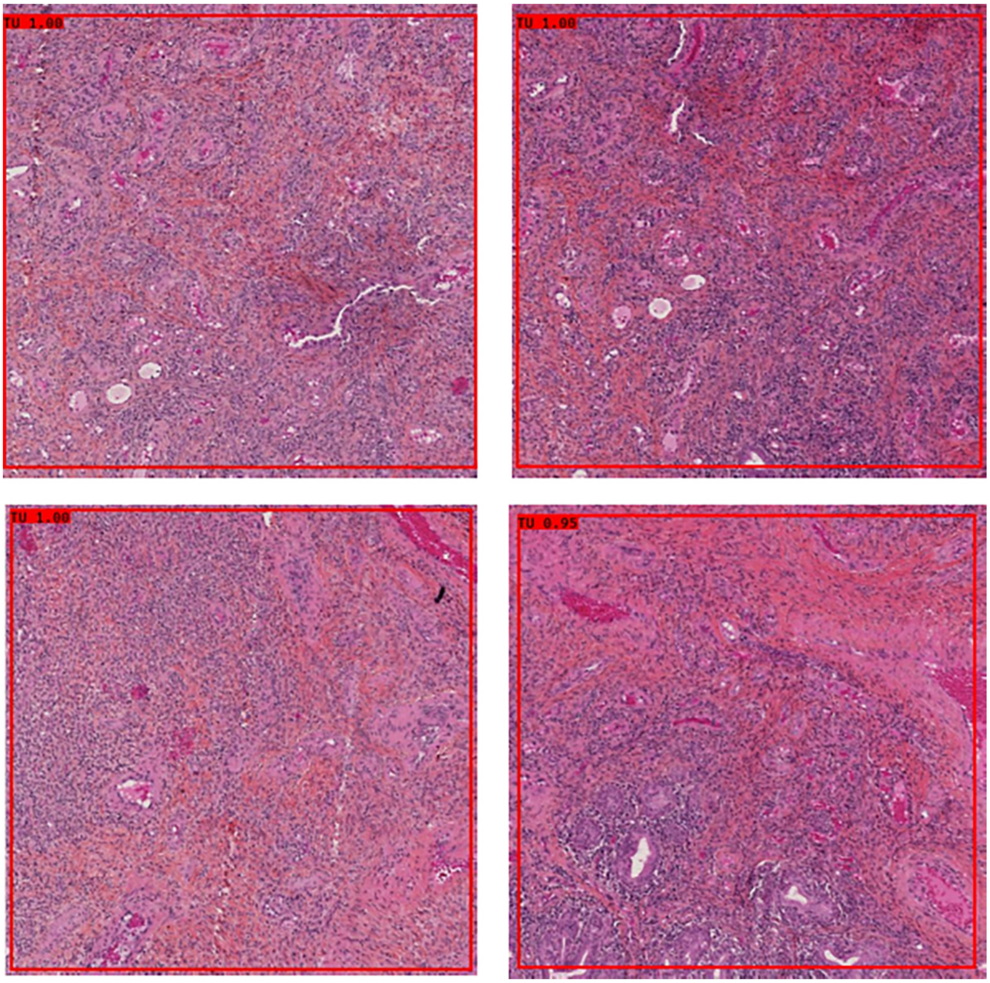


Fig. 13. Algoscope: Patches classified as class 3 while the slide C06_B040_S21 was diagnosed class 0 by the experts. The misclassified area contains vessels and inflammatory infiltrates that can be confusing (See Supplementary Material to dig into the images at better resolution in Algoscope folder).

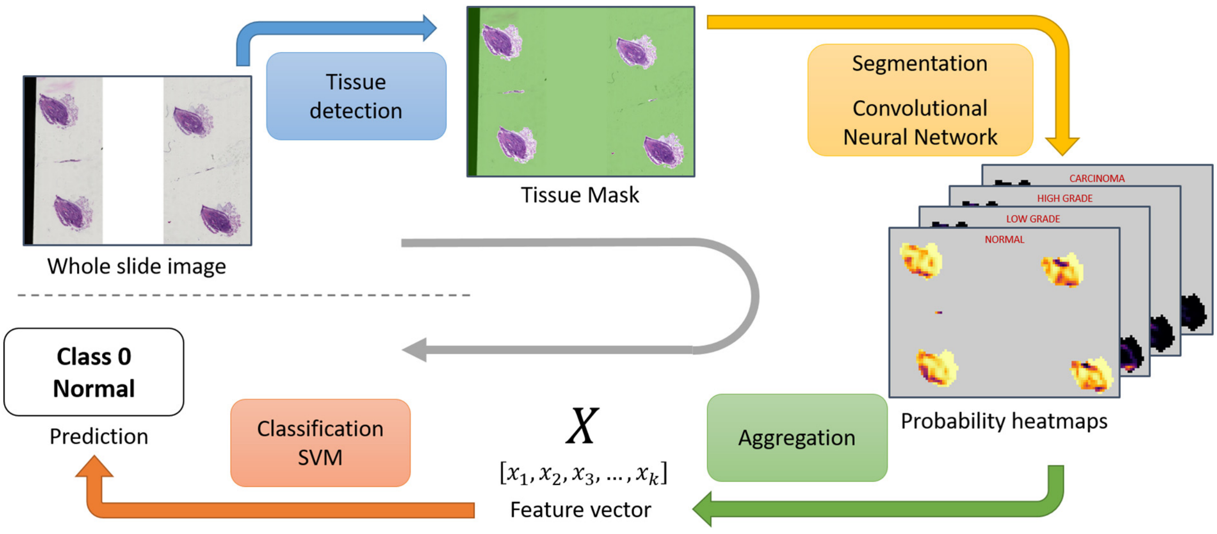


Fig. 14. Tribun Health's end-to-end cervical cancer classification pipeline.

CNN ensemble. For their CNN ensemble, they used DenseNets,[1] as they found it outperforms other (ResNet) architectures. They used DenseNet169 on level 4 patches and DenseNet121 on Level 3 patches, both pre-trained on ImageNet. To improve performance, they trained a linear combination of the 5 models (2 on level 4, 3 on level 3), followed by a softmax layer to output the final 4 CIN class probabilities per patch.

Data augmentation and training parameters. They augmented the data using digital pathology specific HED color augmentation,[2] random linear transformation, Cutout[3] and CutMix.[4] To train each CNN, they used 200 epochs with a learning rate of 5×10^{-2} and use Xavier weight initialization. For classification, they used the categorical cross-entropy loss and add a l_2 weight regularization with $\lambda \in [0.01 - 0.5]$ depending on the dataset.

Hard Negative Mining (HNM). Healthy tissue annotations were given primarily in the epithelium, giving the network a bias towards non-epithelium healthy tissue. They used HNM to enrich the training data. After a training on annotated patches, they did a full inference on all normal tissue slides and added misclassified CIN1, CIN2, and CIN3 patches to the training data. They did the same in CIN1 slides for CIN2 and CIN3 predictions, and CIN2 slides for CIN3 predictions.

Slide-wise SVM classification. Depending on the amount of tissue in the slides, the number of tiles actually analysed varies from less than 5 up to 7000. The aggregation step consists in calculating, for each class separately, statistics on the previously calculated patch predictions, and concatenating them into a vector of fixed size. Specifically, they calculated a histograms features $h_{m,i}$, with m being the m^{th} histogram bin value and i being the slide class, i = 0, 1, 2, 3, as well as percentiles $p_{90,i}$, $p_{95,i}$, $p_{99,i}$ such that we obtain feature vector

$$\ddot{X} = [h_{0,0}, \ldots, h_{m-1,0}, h_{0,1}, \ldots, h_{m-1,1}, \ldots h_{m-1,3}, p_{90,0}, \ldots, p_{99,3}].$$

Slide-wise SVM classification. Given feature vector $\ddot{X}$, they used SVM to predict the probability distribution $p = [p_0, \ldots, p_3]$ that the slide belongs to class 0–3. They did this prediction using 10-fold cross-validation on the patch prediction validation data. They also used the contest reward matrix $R \in R^{4 \times 4}$ to maximize the expected score. The final prediction is chosen from gain $r = pR$ as $i^* = argmax_i\, r_i$.

WSI analysis results

Some of the errors made by Tribun Health algorithm are explained over a couple of slides according to the contestant's choice:

- Slide C11_S046_S11 is classified as class 0 by the algorithm while it is class 3 according to the pathologists' consensus.

- C11_S002_S11 is classified as class 3 by the algorithm while it is class 0 according to the pathologists' consensus.

To understand the algorithm decision, the heatmaps showing the probability of belonging to class i, i = 0, 1, 2, 3 at a patch level are displayed in Figs. 15 and 16 for slides C11_S046_S11 and C11_S002_S11 respectively.

The heatmaps of slide C11_S046_S11 show that the CNI 3 region is well detected by the patchwise classifier but not considered by the whole slide classifier.

This could be explained by 2 reasons:

1) the extracted vector features are not enough informative for representing the heatmaps. One should consider additional ones like the class size or average intensity for example.
2) The features are well representing the heatmaps but the slide classifier is biased toward class 0. Hard negative mining at slide level is a good option to improve these cases.

However, the heatmaps of slide C11_S002_S11 show clear false-positive detections in class 3 misleading the model decision toward class 3. To avoid these errors, one can add another round of hard negative mining of the patchwise model.

They proposed heatmaps exhibiting the different regions responding at different class test from 0 to 3 as illustrated in Fig. 15 for slide C11_C046_S11 and Fig. 16 in slide C11_C002_S11.

Pathologists' discussion. The slide C11_C046_S11 in Fig. 15 contains an area of deep invasive carcinoma not connected to the surface of the sample, probably explaining why none of the 3 algorithms could identify it as a class

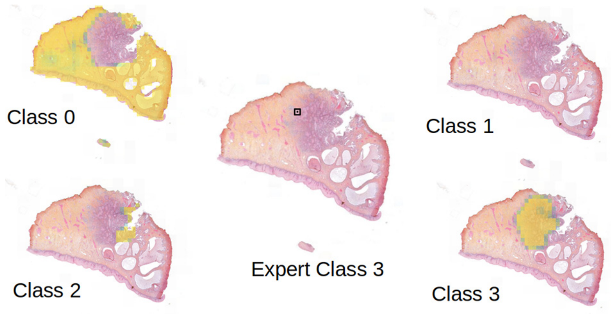


Fig. 15. Slide C11_C046_S11 classified as class 0 and the various heatmaps for every class (most of the interesting WSI and heatmaps are available as Supplementary Material at a good resolution in order to probe them in further details: Folder Tribun_HeatMaps/C11_C046).

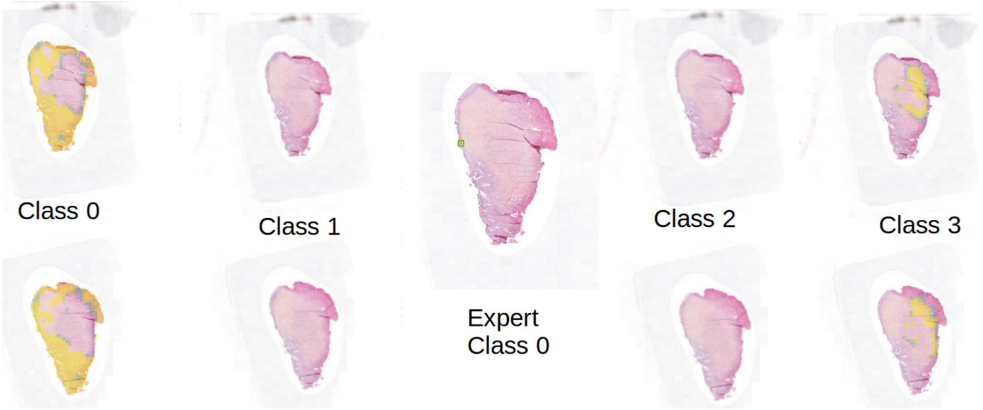


Fig. 16. Slide C11_C002_S11 aclassified as class 3 and the various heatmaps for every class (Folder Tribun_HeatMaps/C11_C002 in Supplementary Material: Folder Tribun_HeatMaps/C11_C002).

3 area. Moreover, the sample is a big piece of tissue, and the training set did not contain many of those big samples, making it difficult for algorithms that learned mostly on small biopsies. Again, the slide C11_C002_S11 in Fig. 16 showed a bigger tissue sample than most of the others, and 2 of the 3 competitors falsely identified class 3 areas in this slide while the experts classified it as class 0. The misclassified areas contained some vessels and technical artifacts such as tissue folding. This pathologist discussion will be further extended in the Pathologists' discussion section 3.3 about the *post hoc* analysis of the results by a board of pathologists.

Data challenge results and detailed medical post-analysis

Challenge results

Global performance of each algorithm was evaluated according to a custom metric devised by a panel of expert pathologists. The score for each prediction equals 1 minus the error, where the error weighting for misclassification has been set by an expert consensus within the scientific council as defined in Table 1 below. The total error is the average error across all predictions. Note that the metric is symmetric, e.g., predicting class 3 when it is actually class 0 produces the same error as predicting class 0 when it is actually class 3.

Table 1
Error table of misclassification.

	Class 0 (pred)	Class 1 (pred)	Class 2 (pred)	Class 3 (pred)
Class 0 (actual)	0.0	0.1	0.7	1.0
Class 1 (actual)	0.1	0.0	0.3	0.7
Class 2 (actual)	0.7	0.3	0.0	0.3
Class 3 (actual)	1.0	0.7	0.3	0.0

More formally, for the development of their algorithm the competitors had access to a training set of roughly 1000 whole slide images $\mathrm{WSI}_i^{\mathrm{train}}{}_{\mathrm{i\ in\ 1..N}}$ for which a global label $y_i^{\mathrm{train}}{}_{\mathrm{i\ in\ 1..N}}$ (with values in {0,1,2,3}) was available for each WSI (in addition a leaderboard of roughly 500 WSI was provided over the course of the challenge training in order for the competitors to compare each other performance).

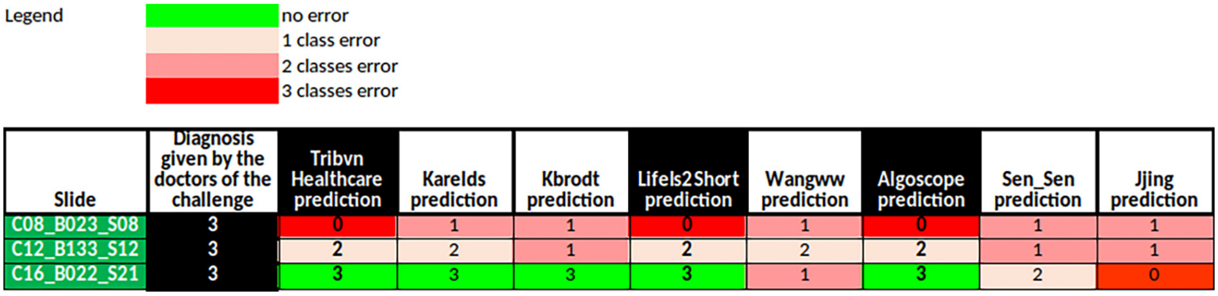
The final score is calculated over a testing set of 1024 other WSIs $\mathrm{WSI}_i^{\mathrm{test}}{}_{\mathrm{i\ in\ 1..N'}}$ for which the labels $y_i^{\mathrm{test}}{}_{\mathrm{i\ in\ 1..N'}}$ (with values in {0,1,2,3}) were not be provided to the competitors. Given a testing whole slide image $\mathrm{WSI}_i^{\mathrm{test}}$, let y_i^{test} in {0,1,2,3} be the predicted label by one of the competitors' algorithm. Then the algorithm score is computed as follows: $Score = 1 - Error$

where $Error = 1/N\, \Sigma\, f(y_i, \hat{y}_i)$ having the function $f(y_i, \hat{y}_i)$ defined by the values defined in Table 1 for which the possible values of actual classes y (resp. predicted classes ŷ) are indicated in the rows (resp. columns). The Error varies from 0.0 (no error) to 1.0 (worst case: only very bad errors with a penalty of 1.0 over each of the N' test WSIs. This worst case will never happen because of the label 1 and label 2 WSI in the data set. A statistical study has proven that a minimum score of 0.776 (max error of 0.224) corresponds to a potential naive algorithm giving the label 1 to all slides.

In the following, we analyze the results in more details and specifically we focus on the most severe misclassification by at least one of the top competitors (class 0 ↔ class 3) (see Fig. 17).

Detailed analysis over three representative WSI

Out of the 1024 final test WSI, 12 WSI were found of particular interest: these slides were problematic for at least one of the top competitors by making a error level of at least 2 grade levels. Among these 12 slides, the board of doctors selected 3 specific ones that were able to illustrate not only the



Legend
🟩 no error
▨ 1 class error
▨ 2 classes error
🟥 3 classes error

Slide	Diagnosis given by the doctors of the challenge	Tribvn Healthcare prediction	Karelds prediction	Kbrodt prediction	Lifels2Short prediction	Wangww prediction	Algoscope prediction	Sen_Sen prediction	Jjing prediction
C08_B023_S08	3	0	1	1	0	1	0	1	1
C12_B133_S12	3	2	2	1	2	2	2	1	1
C16_B022_S21	3	3	3	3	3	1	3	2	0

Fig. 17. Error analysis for a chosen subset of slides and competitors. For the sake of clarity, we selected three competitors among the top ones to analyze the results (see more details on the challenge website[14])

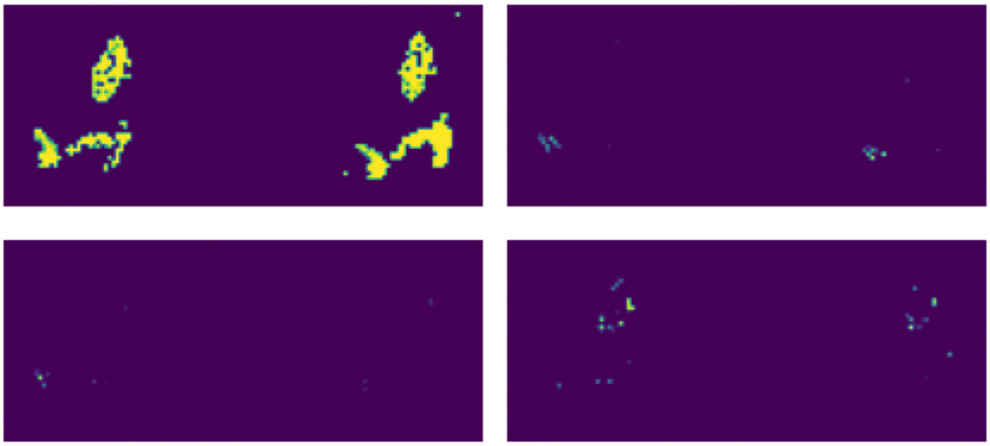


Fig. 18. Lifeis2Short heatmaps for the problematic slide C08_B023_S08 (one slide of the original slide in Fig. 19). From left to right and top to bottom Class 0, 1, 2, 3 heatmaps.

predictive power of AI algorithms for clinical usage but also the room for improvements (see Fig. 17). It is worth noting that the errors were very often explainable by the pathologists as ambiguous cases for instance and will be discussed later on in this report. In complement, 2 interesting additional slides (C11_C046_S11 and C11_C002_S11) were analyzed in the Pathologist discussion paragraph concerning Tribun Health results.

The medical question addressed by this data challenge was the ability for AI algorithms to assist the diagnosis of pathologists in order to classify

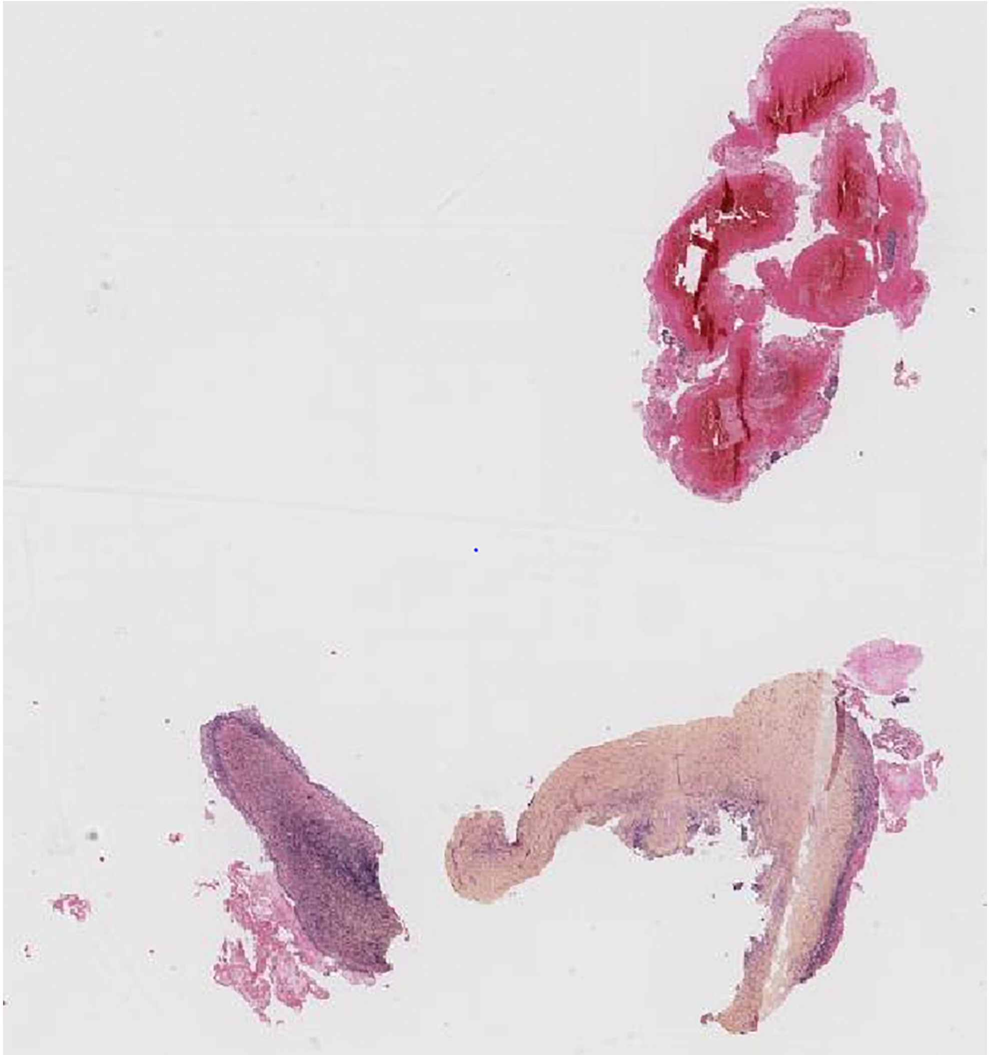


Fig. 19. Algoscope: no detection of any abnormal regions in the problematic slide C08_B023_S08 graded 3 by the expert (one slice tissue over the WSI as presented in Fig. 20).

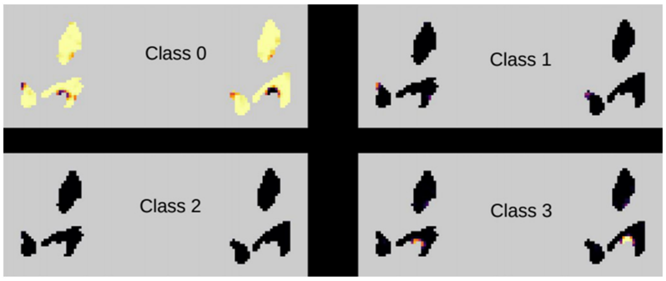


Fig. 20. Tribun Health: misclassification 0 for 3 by Tribun Health despite a small region of class 3 detected in the slide (see next Fig. 21).

cervical biopsy slides into 1 of the 4 WHO diagnostic categories (WHO 2020), from "normal" to invasive carcinoma, through low- or high grade lesions. The algorithms were asked to identify only the most severe diagnostic class for each slide. A few months after the final event of the challenge, the expert pathologists set up a board to assess the results and come up with a medical analysis on where and why the best algorithms failed.

For the sake of comparison, we present hereby the analysis of 3 competitors. The one by Tribun Health and the one by Lifeis2Short that globally make the same errors are similar pipelines, the one by Tribun Health outperforming most of the competitors over the whole test data set. Algoscope team outperforms the other competitors in the identification of class 3 (and 2) slides which can be a big asset for automatic screening of patients (i.e., not missing critical cases). The other teams did not allocate time resources to explain their methods afterwards: interestingly, most of these competitors are called serial data challengers that can proceed any kind of data without any kind of specific expertise on the topic. In particular, one of the leading competitors' processing pipeline did not make use of any patch annotation but worked at the slide level annotation (see [14] and [15]). Nevertheless, in statistical learning, scoring well does not imply being usable at the clinical level. This is what we explore in the following with the help of a post-challenge committee teaming up senior and junior pathologists as well as data scientists.

To sum up the methodology, we decided to compare 3 very representative WSI classification results: 1 problematic case C08_B023_S08, 1 perfect slide C16_B022_S21, and 1 in-between case C12_B133_S12, given by the 2 teams that mixed pathologists and data scientists for the challenge, known as Tribun Health and Algoscope. These slides were chosen by the medical subgroup of the scientific council of the challenge and were provided to the contestants for their feedback about the failure cases in particular. We also discuss results of the third team we elected Lifeis2Short.

All in all, from now on, we give a thorough analysis of results by 3 representative leading teams in the challenge over 3 representative slides (highlighted with a black background in Fig. 17): 1 case very problematic (not classified as 3 but 0), 1 perfect case(classified as 3 by the 3 chosen competitors), and 1 in-between case (classified as 2 instead of 3 but explainable).

The problematic case: WSI C08_B023_S08

This slide C08_B023_S08 was diagnosed class 3 by the experts and both Tribun Health and Algoscope rated class 0 the slide. As stated in the methodology section, Lifeis2Short team did not classified it in class 3 but in class 0 as illustrated in Fig. 18 just as the 2 other competitor algorithms.

Algoscope's algorithm did not detect any abnormality on this slide (neither class 1, class 2, nor class 3 zone) as illustrated in Fig. 19, so the output was class 0. After reviewing the whole slide with their pathologists team, it seems that within the limit of these histological levels, without further information (deeper sections or concomitant biopsies on another slide) the rare areas with cellular atypia are too equivocal to clearly state whether it is a high grade lesion with eroded epithelium (class 2) or an infiltrative carcinoma (class 3).

In Fig. 20, illustrations for Tribun Health results on this problematic case is presented with detailed views in Fig. 21. The post-challenge annotation by the experts confirms that a small area of class 3 has been well detected by the patch level classifier but discarded from the final prediction by the slide level classifier. This error is probably due to the histogram-based features that does not sufficiently well model the tissue class distribution when one of the classes is of small size. Improving the feature extraction part is important for avoiding such an error.

Neither Algoscope nor Lifeis2Short did detect any abnormality on this slide (neither class 1, class 2, nor class 3 zone), so the output was class 0. After reviewing the whole slide with the expert pathologists, the lesion is indeed difficult since it is very small and at the border of the tissue section. Careful reviewing of this case showed that some experts suggested that this small island might be rather class 2 than class 3, while other experts maintained class 3 diagnosis. For Tribun Health, as mentioned earlier, the small focus was however labeled class 3 on the heatmap but the final score was class 0. The determination of the threshold used for taking into account a small area is therefore an important feature of an algorithm, with a direct impact on sensitivity and specificity.

One perfect case:WSI C16_B022_S21

This slide was misclassified by one of the best competitors (see Fig. 17) but the algorithm of the 3 competitors who participated in this paper all agreed with the expert pathologists panel and scored it as a class 3 slide.

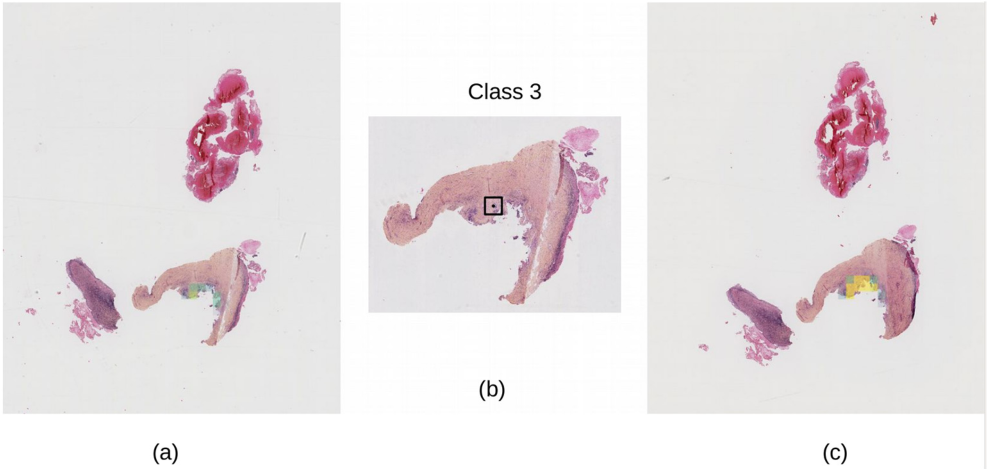


Fig. 21. Tribun Health: (a) A mitigated detection of the class 3 region on tone of the 2 tissue slices in the WSI. (c) The yellow regions corresponds to the class 3 regions detected by the algorithm. (b) The black square corresponds to a post-challenge annotation made by the pathologists to explain their WSI level annotation of class 3 slide (Folder: Tribun_HeatMaps/C08_B023 in Supplementary Material).

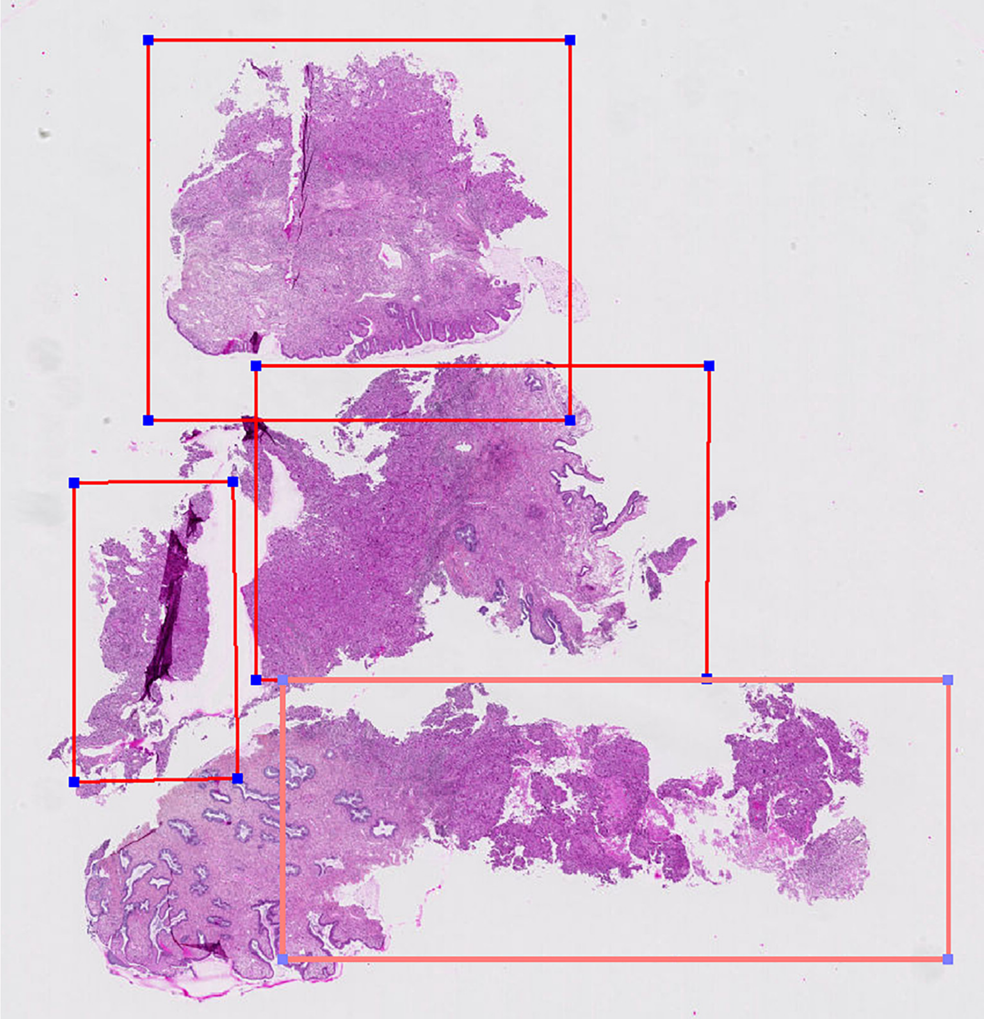


Fig. 22. Algoscope: detection of class 3 regions on the slide correctly classified as class 3 for slide C16_B022_S21 (See Supplementary Material to dig into the images at better resolution in Algoscope folder).

The Algoscope algorithm correctly classified this slide as a class 3 WSI. It detected infiltrating carcinoma all over the WSI as illustrated in Fig. 22. The Tribun Health algorithm correctly assigned the slides to class 3 as well (see Fig. 23).

The in-between case: WSI C12_B133_S12

This slide has been diagnosed as class 3 by the experts but all 3 algorithms diagnoses it as class 2.

The Algoscope algorithm detected a single area of invasive carcinoma (the red square annotation in Fig. 24 as class 3), several areas of high grade lesion (the green squares annotation in Fig. 25 as class 2) and one area of low grade (the blue square annotation in Fig. 24 as class 1). The set diagnosis was: high grade lesion (class 2), because much more areas of class 2 were detected by the algorithm and likely to fit with the misclassification risk score for winning the competition. However, it is interesting to note that several areas of infiltrating carcinoma (on the upper left biopsy fragment) were not detected (probably due to the fact that the detection threshold is slightly too high). All in all, the algorithm detected an area of invasive carcinoma (class 3) but still made the diagnosis of a high grade lesion because it had found many more regions of class 2.

The Tribun Health algorithm classified as well the slide as class 2 instead of class 3 but did not detect class 3 areas (see Fig. 25). Again, these findings point out the importance of choosing the right threshold for taking into account, or not, a small area compared to much bigger areas.

The pathologists' discussion

This data challenge focused on epithelial lesions of the uterine cervix. These lesions account for an important amount of cases in the daily practice of pathologists. Tools that would pre-identify these lesions and prioritize most severe cases would be therefore very useful for pathologists in order to save diagnostic time.

Among over 500 competitors, we chose the best algorithms according to their final scores on the evaluation set. Their capacity in distinguishing the four diagnostic classes of cervical epithelial lesions (normal, low grade, high grade, invasive tumor) was excellent since, out of 1024 slides of the final validation set, there were only 12 slides showing at least one “two classes error compared to ground-truth” for at least one of the best competitors.

Beside these excellent results, we decided in this work to focus on these discordant situations, looking for potential explanations to these discrepancies, leading to a better understanding of the “A.I. black box” by pathologists and health practitioners.

The results presented in this paper showed that the detection sensitivity of some algorithms was extremely high even with very small lesions made of only a few cells, as shown on one competitor heatmap (see Fig. 20).

However, the final output of the algorithms still contained a few “2 or 3 classes or grade error discrepancy” that the medical panel analyzed one by one. In a few situations, the algorithm indeed found a challenging area on the slide but misinterpreted it, mostly due to a lack of training of unusual

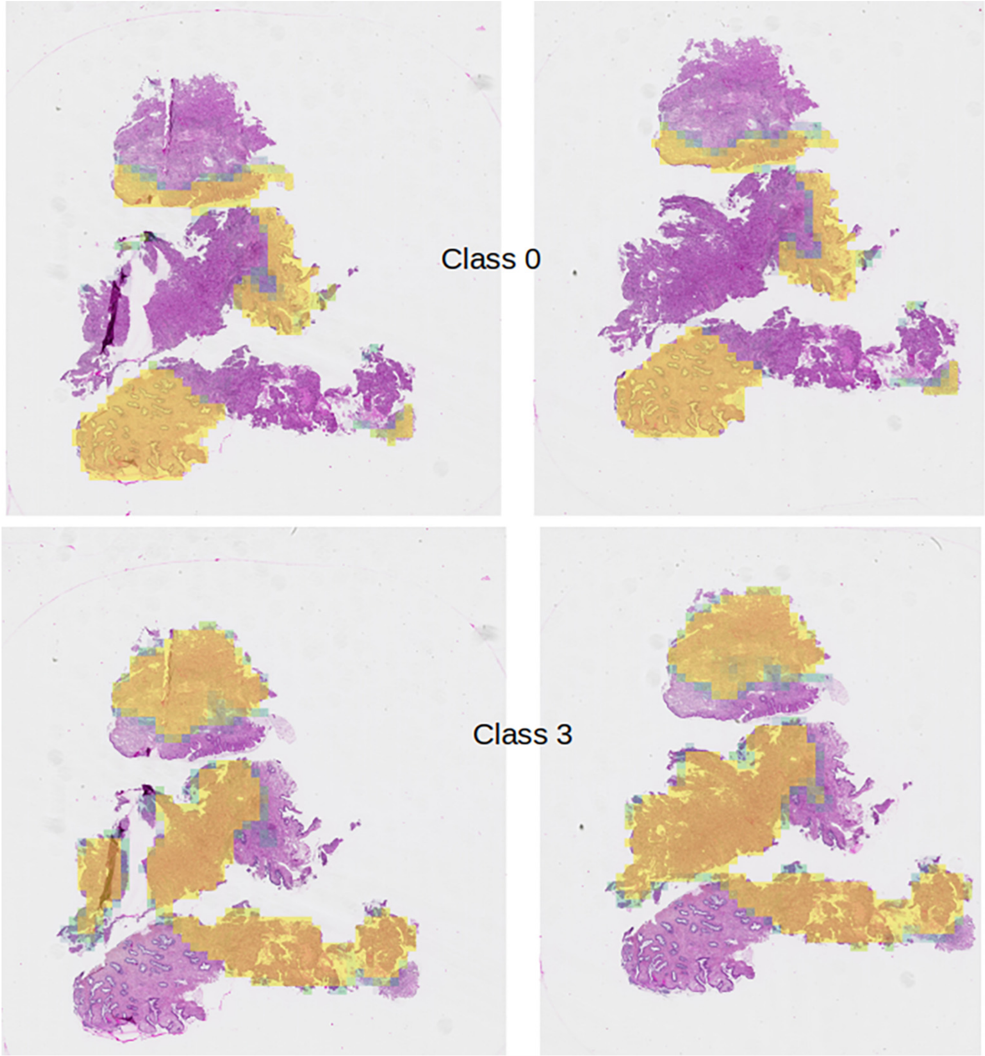


Fig. 23. Tribun Health: class 0 and 3 regions detected for a slide correctly classified as 3 for slide C16_B022_S21 (Folder Tribun_HeatMaps/C16_B022 in Supplementary Material).

images such as inflammatory infiltrates, vessels, or technical artifacts. Increasing the amount of training slides will improve the classifications. We specifically noticed that most significant errors of the algorithms were made on large pieces of tissue. Usually, cervical biopsies are 1–3 mm large but in some cases, they can be larger and the problems in large samples can be different, with other histological structures appearing in the deep part of the sample, that are not present in more superficial biopsies. This lack of training data issue emphasizes the need for large amounts of training sets of all kind of samples and tissues. This is a challenge in itself regarding the regulations and constraints about obtaining human tissues in many countries.

Also, optimizing all technical pre-analytical steps of the slides will also reduce technical artifacts and possible pitfalls for AI.

Sensitivity thresholds are also a crucial choice that the competitors have to deal with in a data challenge. It is indeed important in order to win the competition to optimize the choices made by the algorithm as compared to the metric established by the organizers. The algorithms have all been built in order to fulfill this competition goal. In the true diagnostic setting, detection thresholds will not be optimized to a metric, but rather will be adjusted in order to maximize the chance of detecting important lesions, especially high grade or malignant ones. In real life, pathologists will indeed certainly prefer one heatmap showing a tiny area with possible malignant cells, rather than an overall AI judgment that will choose not to mention this tiny area and will refer only to the majority of non-tumoral slide. In other words, it is likely that pathologists will prefer highly sensitive tools rather than highly specific ones, the final choice between benign and malignant lesions being done by the pathologist. The examples that we report in this paper where an area of “class 3 – tumor area” is finally not mentioned in the overall score should not be encountered in the real diagnostic setting (see Fig. 21).

Another limit for the training of algorithms are the unavoidable borderline situations, when even experts might be slightly divergent regarding some complex lesions, because of their rarity, because of the need of other staining on the sample, or because of non optimal tissue sampling or non-optimal tissue processing. During the preparation of the challenge, a few cases were already rejected from the challenge because of these kind of situations. However, when the pathologists looked again in details at the 12 slides selected for this paper, 1 slide was still somehow problematic, and the experts could not fully agree on a small lesion, whether it was a class 2 or a class 3 lesion. This situation actually accounts for true life and true difficult pathology cases, often combining intrinsic complexity and technical/sampling limits. Such cases without clear-cut ground truth are not surprisingly difficult as well for algorithms.

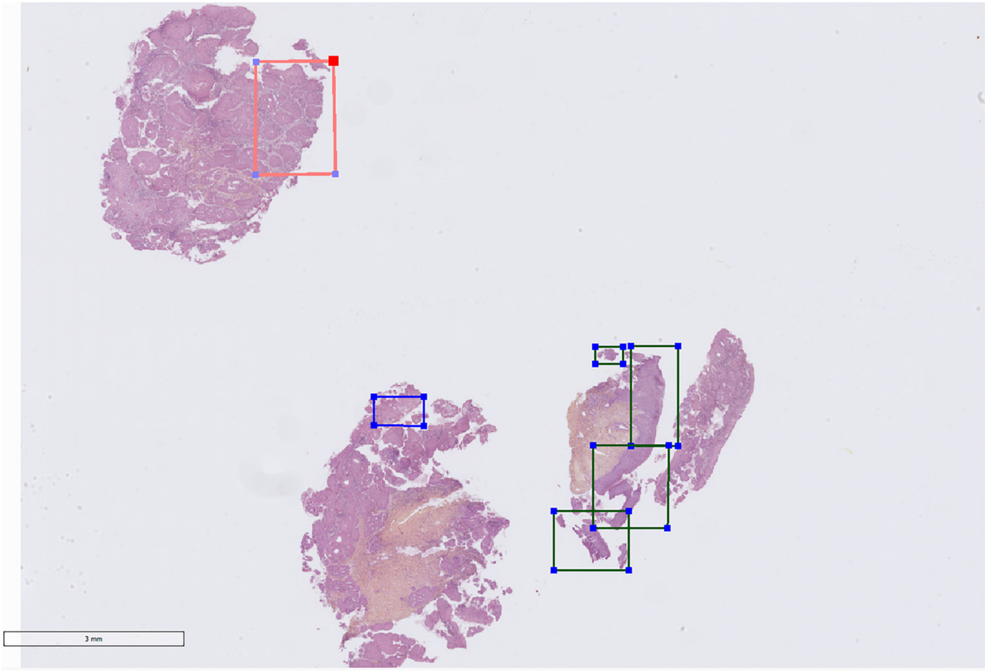


Fig. 24. Algoscope on the class 3 in-betweeen C12_B133_S12 case: detection of class 2 (green annotation) and class 3 (red annotations) regions on the slide finally labeled 2 by the algorithm instead of class 3. Regions detected as class 1 areas are also highlighted within blue squares (See Supplementary Material to dig into the images at better resolution in Algoscope folder).

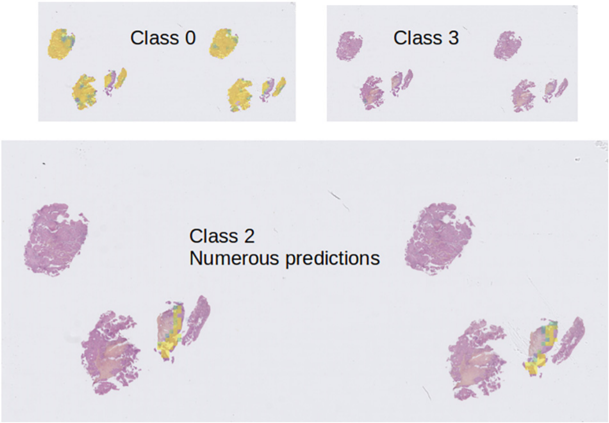


Fig. 25. Tribun Health on the class 3 in-betweeen C12_B133_S12 case: class 0, 2, and 3 detection region for a class 3 slide (Folder Tribun_HeatMaps/C12_B133 in Supplementary Material).

Conclusion

In this work, the pathologists realized that the AI box is actually not so black, and that, at least for some of the best algorithms that emerged in this competition, discrepancies between AI and ground truth can be explained most of the time. Our findings are of course very preliminary and focused on one type of pathology and one type of tissue sample, but they pave the way for improving the level of confidence of health practitioners who will one day work with the help of these algorithms.

Conflict of interest

None.

Acknowledgments

We thank the Health Data Hub (HDH) and the Banque Publique d'Investissement (Bpifrance) for their support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpi.2022.100149.

References

1. Kleppe A, Skrede O-J, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. Nat Rev Cancer 2021. https://doi.org/10.1038/s41568-020-00327-9.
2. Echle A, Rindtorff NT, Brinker TJ, Luedde T, Pearson AT, Kather JN. Deep learning in cancer pathology: a new generation of clinical biomarkers. Br J Cancer 2020. https://doi.org/10.1038/s41416-020-01122-x.
3. Kann BH, Hosny A, Aerts HJWL. Artificial intelligence for clinical oncology. Cancer Cell 2021. https://doi.org/10.1016/j.ccell.2021.04.002.
4. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Heal 2021;7500:1–9. https://doi.org/10.1016/s2589-7500(20)30292-2.
5. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577:89–94. https://doi.org/10.1038/s41586-019-1799-6.
6. 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;16:703–715. https://doi.org/10.1038/s41571-019-0252-y.
7. Tizhoosh HR, Diamandis P, Campbell CJV, et al. Searching images for consensus: can AI remove observer variability in pathology? Am J Pathol 2021. https://doi.org/10.1016/j.ajpath.2021.01.015.
8. Colling R, Pitman H, Oien K, et al. Artificial intelligence in digital pathology: a roadmap to routine use in clinical practice. J Pathol 2019;249:143–150. https://doi.org/10.1002/path.5310.
9. Lassau N, Bousaid I, Chouzenoux E, et al. Three artificial intelligence data challenges based on CT and MRI. Diagn Interv Imaging 2020;101. https://doi.org/10.1016/j.diii.2020.03.006.
10. Lassau N, Estienne T, de Vomecourt P, et al. Five simultaneous artificial intelligence data challenges on ultrasound, CT, and MRI. Diagn Interv Imaging 2019;100:199–209. https://doi.org/10.1016/j.diii.2019.02.001.
11. Litjens G, Bandi P, Bejnordi BE, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. Gigascience 2018;7. https://doi.org/10.1093/gigascience/giy065.
12. Hartman D, Laak JWM Van Der, Gurcan M, Pantanowitz L. Value of public challenges for the development of pathology deep learning algorithms. J Pathol Inform 2020;11:7. https://doi.org/10.4103/jpi.jpi_64_19.
13. Female Genital Tumours. *WHO Classification of Tumours5th ed.* . 2020.
14. https://www.drivendata.co/blog/tissuenet-cervical-biopsies-winners/. Last access 1st May 2022.
15. https://github.com/drivendataorg/tissuenet-cervical-biopsies. Last access 1st May 2022.