Machine learning for grading and prognosis of esophageal dysplasia using mass spectrometry and histological imaging

Computers in Biology and Medicine 138 (2021) 104918

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Machine learning for grading and prognosis of esophageal dysplasia using mass spectrometry and histological imaging

Manon Beuque [a,1,*], Marta Martin-Lorenzo [b,c,**,1], Benjamin Balluff [b], Henry C. Woodruff [a,d], Marit Lucas [e], Daniel M. de Bruin [e], Janita E. van Timmeren [a,f], Onno J.de Boer [g], Ron MA. Heeren [b], Sybren L. Meijer [g,2], Philippe Lambin [a,d,2]

[a] Department of Precision Medicine, GROW – School for Oncology and Developmental Biology, Maastricht University, 6229 ER Maastricht, The Netherlands
[b] Maastricht MultiModal Molecular Imaging Institute (M4I), Universiteitssingel 50, 6229 ER, Maastricht, Maastricht University, the Netherlands
[c] Department of Immunology, IIS-Fundación Jiménez Díaz, UAM, Avda. Reyes Católicos, 28040, Madrid, Spain
[d] Department of Radiology and Nuclear Medicine, GROW – School for Oncology and Developmental Biology, Maastricht University Medical Centre+, 6202 AZ, Maastricht, The Netherlands
[e] Department of Biomedical Engineering & Physics, Amsterdam UMC, Meibergdreef 9, 1105 AZ, Amsterdam, University of Amsterdam, Amsterdam, the Netherlands
[f] Department of Radiation Oncology, University Hospital Zurich and University of Zurich, Rämistrasse 100, 8006, Zürich, Switzerland
[g] Department of Pathology, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, 1105 AZ, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Deep learning
Machine learning
Mass spectrometry imaging
H&E staining
Barrett's esophagus

ABSTRACT

Background: Barrett's esophagus (BE) is a precursor lesion of esophageal adenocarcinoma and may progress from non-dysplastic through low-grade dysplasia (LGD) to high-grade dysplasia (HGD) and cancer. Grading BE is of crucial prognostic value and is currently based on the subjective evaluation of biopsies. This study aims to investigate the potential of machine learning (ML) using spatially resolved molecular data from mass spectrometry imaging (MSI) and histological data from microscopic hematoxylin and eosin (H&E)-stained imaging for computer-aided diagnosis and prognosis of BE.

Methods: Biopsies from 57 patients were considered, divided into non-dysplastic (n = 15), LGD non-progressive (n = 14), LGD progressive (n = 14), and HGD (n = 14). MSI experiments were conducted at 50 × 50 μm spatial resolution per pixel corresponding to a tile size of 96x96 pixels in the co-registered H&E images, making a total of 144,823 tiles for the whole dataset.

Results: ML models were trained to distinguish epithelial tissue from stroma with area-under-the-curve (AUC) values of 0.89 (MSI) and 0.95 (H&E)) and dysplastic grade (AUC of 0.97 (MSI) and 0.85 (H&E)) on a tile level, and low-grade progressors from non-progressors on a patient level (accuracies of 0.72 (MSI) and 0.48 (H&E)).

Conclusions: In summary, while the H&E-based classifier was best at distinguishing tissue types, the MSI-based model was more accurate at distinguishing dysplastic grades and patients at progression risk, which demonstrates the complementarity of both approaches. Data are available via ProteomeXchange with identifier PXD028949.

1. Introduction

Esophageal adenocarcinoma (EAC) remains one of the deadliest cancers with a 5-year survival rate of less than 20% [1] and Barrett's esophagus (BE) is the only known precursor lesion. BE is a condition of the distal esophagus where the stratified squamous epithelium is replaced by columnar epithelium with goblet cells due to gastroesophageal reflux disease [2]. BE may progress from non-dysplastic metaplasia (NDBE) through low-grade dysplasia (LGD), to high-grade dysplasia (HGD) and esophageal adenocarcinoma (EAC). A histopathology diagnosis of LGD is an important independent risk factor to

* Corresponding author. The D-Lab, Department of Precision Medicine, GROW – School for Oncology and Developmental Biology, Maastricht University, P.O. Box 616, 6200 MD Maastricht, Universiteitssingel 40, room 4.549, 4th floor, 6229 ER, Maastricht, the Netherlands.
** Corresponding author. Immunoallergy and Proteomics Lab, IIS-Fundacion Jimenez Diaz, Avda. Reyes Catolicos 2, 28040-Madrid, Spain.
E-mail addresses: m.beuque@maastrichtuniversity.nl (M. Beuque), marta.martin@fjd.es (M. Martin-Lorenzo).
[1] These authors contributed equally.
[2] Share senior authorship.

https://doi.org/10.1016/j.compbiomed.2021.104918
Received 23 July 2021; Received in revised form 29 September 2021; Accepted 29 September 2021
Available online 4 October 2021

develop EAC [3]. This diagnosis however is hampered by inter- and intra-observer variability and international guidelines therefore mandate a second opinion. The individual rate of progression from BE patients with LGD to HGD/EAC is difficult to evaluate on hematoxylin and eosin (H&E) slides using light microscopy and ranges between 0.6 and 13.4% per patient per year [4]. Due to the lack of reliable indicators of progression, current clinical treatment guidelines for LGD patients are not well defined and range from immediate local treatments to further endoscopic surveillance [5]. Currently no objective biomarkers exist to identify BE patients with LGD that quickly progress to HGD and EAC from LGD lesions that remain stable for years.

List of abbreviations

AUC	area under the curve
BE	Barrett's esophagus
CBAM	Convolutional Block Attention Module
CI	confidence interval
DL	deep learning
EAC	Esophageal adenocarcinoma
H&E	hematoxylin and eosin stained tissue scans
HGD	high grade dysplasia
LOPOCV	leave-one patient-out cross validation
MSI	mass spectrometry imaging
MLP	multi-layer perceptron
ML	Machine learning
NDBE	non-dysplastic metaplasia
RF	random forest
TIC	total-ion-count

Computer-aided diagnostics of histological images and new molecular imaging modalities are therefore needed to assist the pathologist in grading BE lesions and give a reliable prediction of the disease progression.

For the molecular analysis of histological tissue section, mass spectrometry imaging (MSI) is a young technique in expansion. MSI enables the acquisition of spatially resolved molecular profiles from tissue sections without any labelling. MSI has demonstrated during the past decade to be a powerful tool to extract clinically relevant information beyond histology from the molecular setup of different cancer types [6]. In the context of EAC, the group of Walch and coworkers has already used MSI to find several proteins to be indicative of poor survival, metastasis, and chemosensitivity [7].

MSI and histology can be used in combination and we hypothesize that the complementarity of both can potentially reinforce the accurate grading of BE and prognosis. From a technical point of view, both imaging modalities (optical microscopy and MSI) provide copious amounts of data: histological images are usually high resolution whereas MSI data is high-dimensional in its feature space, making them both suited for machine learning (ML) approaches [8].

Our general objective is to investigate ML solutions applied to MSI and H&E data and analyse its ability to discriminate epithelial from stromal tissue and to classify BE samples according to the grade of dysplasia. We propose here a workflow based on ML, which can classify the tissue between epithelial tissue and stroma and display where the classifier identifies dysplastic areas of interest in the epithelial region. This way the experts can focus on the specific region of the H&E stained slides. This would provide a cheap and fast auxiliary observation and would help the experts to give a faster and more accurate diagnosis. The second aim of the study is to use ML solutions to distinguish LGD-lesions at risk to progress from those that display stable disease.

2. Methods

2.1. Patient material

Formalin-fixed paraffin-embedded (FFPE) esophagus tissue biopsies were retrieved from the archives of the Department of Pathology of the Amsterdam UMC, location Meibergdreef. A total of 57 biopsy samples from 57 patients were collected and covered the complete spectrum of BE, ranging from NDBE (n = 15) to LGD (n = 28) and HGD (n = 14). Based on the patients' follow-up LGD samples were sub-classified into LGD non-progressors (n = 14, no progression to HGD or EAC within a period of two years) and LGD progressors (n = 14, developed HGD or EAC within 2 years). All samples were anonymized for further use and did not require approval from the relevant Institutional Ethics Committee under applicable local regulatory law ('Code of conduct', FEDERA).

2.2. Mass spectrometry imaging experiments

For this unique dataset, FFPE esophagus tissue samples were cut at 5 μm thickness and randomly distributed on a total of 19 indium tin oxide-coated conductive glass slides (Delta Technologies). For MSI peptide measurements, samples were prepared as previously described by Vos et al. in [9]. Briefly, samples were deparaffinised with xylene, exposed to antigen-retrieval and on-tissue tryptic digested using the Antigen Retriever 2100 (Aptum Biologics, UK) and a SunCollect pneumatic sprayer (SunChrom GmbH, Germany), respectively. After a 17h long incubation, alpha-cyano-4-hydroxycinnamic acid was applied using the same SunCollect sprayer. Before MSI, optical images of the glass slides were taken with a high-quality film scanner (Nikon LS-5000) with a true optical resolution of 4000 dpi (i.e. one pixel is 6.35 μm) in order to define the measurement region. This image therefore acts as anchor image and is later also used to co-register high-resolution H&E images to the MSI data. MSI experiments were performed at 50 μm lateral pixel size (40 × 40 μm laser beam scan range) on a rapifleX MALDI-ToF mass spectrometer (Bruker Daltonics) in reflectron and positive-ion mode within an m/z range of 800–3000. The instrument was calibrated beforehand using Red Phosphorus. Line scan sequence was non-random (i.e. spectra are acquired sequentially from upper left to lower right). Random walk within one pixel was deactivated and 700 spectra were averaged per pixel with a MALDI laser repetition rate of 10 kHz. All individual spectra underwent on-the-fly smoothing (Savitzky-Golay 5%) and baseline subtraction (TopHat). Digitization rate was 1.25 GS/s resulting in 55,000 data points per spectrum (i.e. per MSI pixel), which was reduced to 80% of its original size in FlexImaging (Bruker Daltonics). MSI datasets contained on average 4500 MSI pixels (min = 1370; max = 11,647) and were exported separately as imzML files from FlexImaging.

2.3. Hematoxylin and eosin staining

The same tissue sections analysed by MSI were concurrently stained with H&E to minimize possible staining differences. For this, the matrix was first washed-off from the slides using 70% ethanol for 2–3 min, followed by a 3 min wash with Milli Q water. Slides were stained with hematoxylin (3 min), washed for 3 min with tap water to remove excess hematoxylin, then stained with eosin (30 s), washed again with tap water for 3 min to remove excess eosin, followed by a 1 min ethanol wash and a 30 s xylene wash before attaching coverslips to the slides using Entellan as a mounting medium. The stained slides were scanned with a digital slide scanner at 20x magnification (Mirax Desk, Carl Zeiss MicroImaging, Göttingen, Germany). Tissue scans were exported using Pannoramic Viewer (3DHistech, Hungary) in JPG file format (at 90% quality compression and at original resolution), resulting in pixel sizes of 0.52x0.52 μm^2. The images were superposed to the MALDI-MSI data in FlexImaging using a 3-control-point co-registration of previously

applied fiducial markers. Stained tissue sections were annotated by an expert pathologist according to the tissue type (epithelial/stroma) and BE grade. In order to evaluate the co-registration quality, all datasets were individually inspected visually. In all cases, the eye-estimated average error (<10 μm) was significantly smaller than the laser spot size (50 μm) (Supplement Fig.1). An example of the aligned information comprising histological images, MSI, and annotations is shown in Fig. 1.

2.4. MSI data pre-processing

A recalibration was performed in Flex Analysis v3.4 using the lock masses *m/z* 842.510 and 1045.564 with a peak assignment tolerance of 500 ppm and *m/z* 1303.615, 1508.750, 1833.954, 1835.957, 2104.190, 2105.190, 2106.190 with a tolerance of 250 ppm. All recalibrated MSI data, coregistered H&E images, and annotations were imported to SCiLS Lab (Bruker Daltonik) where each spectrum was normalized to its total-ion-count (TIC). From SCiLS Lab, the overall spectra from on- and off-tissue regions were exported to mMass (http://www.mmass.org/) for peak picking using the following parameters: (1) Baseline correction precision = 40; (2) Peak-picking: S/N 5.0; Picking height = 90; (3) Deisotoping: maximum charge = 1, isotope mass tolerance *m/z* = 0.15, isotope intensity tolerance = 70%, isotope mass shift = 0.0. The peak picking lists from on- and off-tissue were subsequently compared with a tolerance of 0.2 Da and common peaks were removed from the on-tissue peak list after visual inspection and confirmation (Supplement Table 1). This final peak list (Supplement Table 2) was then imported, together with the imzML files and the histological images of every patient, into Python 3.7. All of these data are available via ProteomeXchange (https://www.ebi.ac.uk/pride/) using the identifier PXD028949. In Python, the mass spectrometry pixels were normalized to their total-ion-count before extracting the maximum intensity for every peak in its ± 0.5 *m/z* interval across all MSI spectra.

2.5. MSI and histology data extraction

Data extraction was performed using Python 3.7 with ImzMLParser and OpenCV libraries. Affine geometric transformations were performed in order to spatially link MSI and H&E based on the co-registrations previously done in FlexImaging, which were accessible via the respective .mis XML files. These files also contained the annotations as sets of

Table 1
Number of tiles distributed over the different grades and tissue types Abbreviations used: NDBE, non-dysplastic Barrett Esophagus; LGD, low-grade dysplasia; HGD: high-grade dysplasia.

Data	Portion of the total tiles [%]	Portion of epithelial tissue [%]	Total number of tiles
Stroma	50%	–	72516
NDBE	36%	73%	52704
LGD	8%	16%	11615
HGD	6%	11%	7988
Total	100%	100%	144823

Table 2
Tile-based classifier performance for predicting tissue type and grade on the test sets. Abbreviations used: H&E, hematoxylin and eosin-stained tissue scans; MSI, mass spectrometry imaging.

Prediction of …	Data type	Labels	Precision	Recall	f1-score	Support (number of tiles)
Tissue type	MSI	Epithelial tissue	0.80	0.81	0.81	10602
		Stroma	0.81	0.82	0.80	11121
	H&E	Epithelial tissue	0.89	0.87	0.88	10602
		Stroma	0.88	0.90	0.89	11121
Grade	MSI	Non dysplastic Barrett's Esophagus	0.97	0.84	0.90	7836
		Low-grade dysplasia	0.68	0.90	0.77	1787
		High-grade dysplasia	0.70	0.98	0.82	1224
	H&E	Non dysplastic Barrett's Esophagus	0.93	0.70	0.80	7836
		Low-grade dysplasia	0.37	0.75	0.50	1787
		High-grade dysplasia	0.44	0.47	0.45	1224

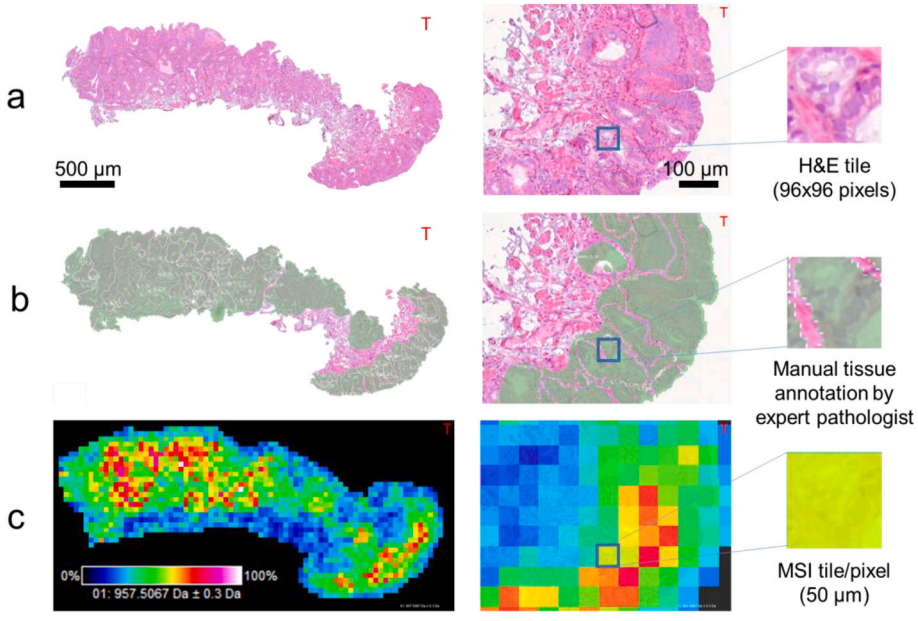

Fig. 1. Illustration of the multimodal imaging data used in this study. Three increasing magnification levels (left to right) of the three spatially co-registered layers of information and their size-matched representation on a tile level are shown in one of the Barrett's esophagus tissue biopsies: (a) The H&E-stained microscopic image (b) the manual pathological annotations made by an expert pathologist on the same H&E image in this case indicating the glandular areas (grey colour) and (c) the MSI data, here represented by the visualization of a particular mass channel (*m/z* 957.5) which co-localizes with the glandular areas shown in (b). Abbreviations used: H&E, hematoxylin and eosin-stained tissue scans; MSI, mass spectrometry imaging; TIC, total-ion-count.

polygonal coordinates (Supplement Fig. 2). For each of the MSI pixels (see Fig. 1 c), the corresponding histological patch was extracted with a size of 96x96 pixels (see Fig. 1 a). Henceforth, the word "tile" will be used to describe both the MSI pixel and the matching histological patch. The histology tiles were labelled according to the annotation of the centre pixel.

2.6. Machine learning models

For each modality (MSI and H&E), three tile-based ML classifiers were trained: the first classified the tiles between epithelial tissue and stroma (tissue type prediction), the second predicted the dysplastic grade of a tile, and the third the progression of dysplasia on a patient level. The data was randomly split into training, validation, and test datasets in a ratio of 0.70/0.15/0.15, respectively. For the patient-level classifiers, we performed a leave-one-patient-out cross validation (LOPOCV) by excluding the data of one patient from the training/validation sets and splitting the training dataset and validation dataset in a ratio of 0.90/0.10 and repeat the process for all the patients. The model was computed using Pytorch 1.2.0 on Python 3.7, on a GPU-clusters of 10 GPUs (NVIDIA GeForce RTX 2080 Ti). The implementation of the machine learning models can be found at: https://github.com/precision-medicine-um/ML-and-MS-in-esophageal-cancer.git.

2.7. Tissue type prediction

The signals of each MS-tile were rescaled by multiplying all values by a factor 10^5 for a better compliance in the software. Then, the features were mapped with a Gaussian distribution per patient with Box-Cox transformation in order to obtain a similar range of signal intensities in all datasets and to remove possible patient/acquisition biases. After pre-processing, the parameters of three models were optimized on the training dataset with three independent grid searches: The impurity measure, the number of estimators, and the maximum depth for random forest (RF), the weight decay, the batch size, the hidden layer sizes, the maximum iteration, and the optimizer for multi-layer perceptron (MLP), and the learning rate, the number of estimators, the maximum depth, the minimum child weight, and subsample for XGBoost. The rest of the parameters were the default parameters from the python library scikit-learn 0.24.2. The three models were then merged with an ensemble modelling method, a voting classifier, which used argmax function to obtain a final probability class prediction. The pipeline can be found as flowchart in Supplement Fig. 3.

As the H&E and MSI datasets were co-registered, we could use the equivalent split of H&E data to form the training dataset and validation dataset and all the tiles were resized to 224x224 pixels to match the required input size of the DL model. A data augmentation step using the library albumentations within Python was applied to the training dataset where transformations were applied on the images with a probability of 0.5 for each augmentation: rotation by 90°, transposition, flipping around the horizontal/vertical or both axes, random intensity filtering, and random affine transformations. All the tiles were normalized on the three channels individually (red, green, and blue). We used a Convolutional Block Attention Module (CBAM) [10] with Resnet50 as the backbone [11] as proposed by the work of Tomita et al. [12] with minor modifications. The CBAM was added between two convolutional blocks, aggregating max pooling and average pooling into a channel attention module and a spatial attention module to focus on representation. The parameters chosen were binary cross-entropy and Adam optimizer with a learning rate of $4x10^{-4}$. The model was trained on batches of 600 tiles. The training stopped once the loss on the validation dataset stopped decreasing after one epoch. Then, the model was evaluated on the test dataset using test time augmentation. Ten different augmentation functions were used on the test dataset, such as a 90° rotation, transposition, horizontal and vertical flip. The model gave a probability per class for each transformation following which the tissue type predictions were averaged, and the class predicted with the highest probability was chosen. The workflows for the MSI and H&E data are presented in Fig. 2.

2.8. Grade of dysplasia prediction

The workflow for grade prediction on a tile-level follows the same workflow as described for tissue type prediction (see in Fig. 2) for MSI and H&E with the difference that the analysis focuses only on tiles from the epithelial regions since BE grading is based on morphological changes in the epithelial structures [13]. As the dataset was highly unbalanced, undersampling was performed in the proportion of the high-grade tiles (lowest number of tiles), choosing randomly the tiles among the 3 classes to not overfit during the training phase.

2.9. Multi-modal classifier

For the two different tasks, the same approach was pursued to establish a multi-modal classifier: we extracted the last layer of features from the trained DL model (2048 features) and combined them to the mass spectrometry features. Then we used grid-search in combination with a MLP to obtain an optimized classifier.

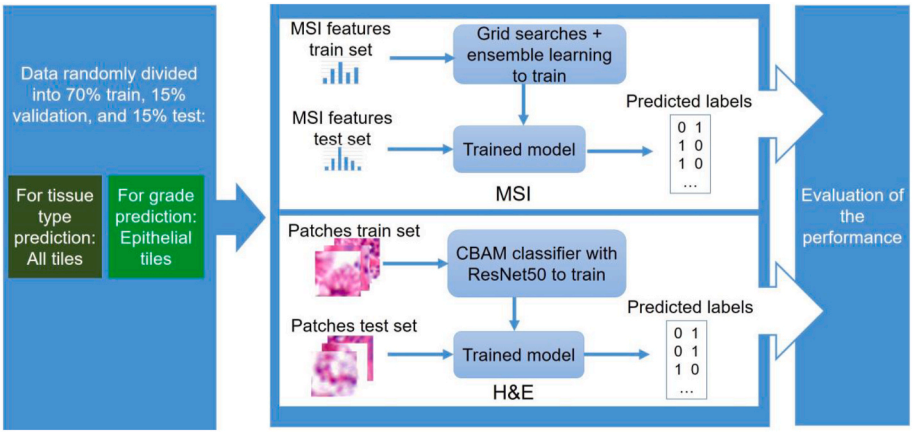

Fig. 2. Workflow for the prediction of tissue type (using all tiles) and grading (using tiles belonging to epithelial tissue only) on a tile-level using the MSI data (top row) and H&E data (bottom row). Abbreviations used: CBAM, Convolutional Block Attention Module; H&E, hematoxylin and eosin-stained tissue scans; MSI, mass spectrometry imaging.

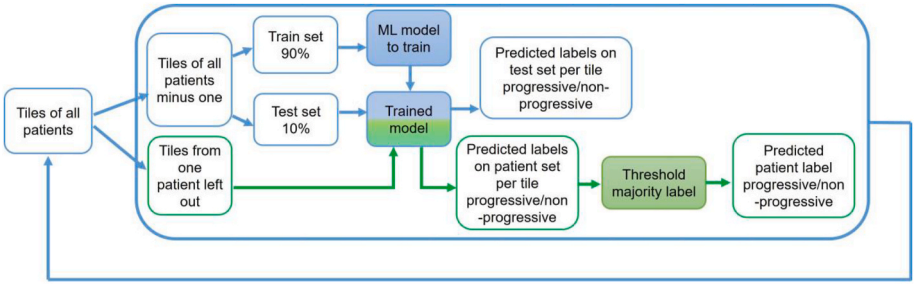

Fig. 3. Workflow for the prediction of low-grade dysplasia progression on a patient-level which is used for both the MSI as well as the H&E data. Abbreviations used: MSI, mass spectrometry imaging; H&E, hematoxylin and eosin-stained tissue scans.

2.10. Identify low grade dysplasia progression

To predict the progression of BE dysplasia, only the tiles belonging to epithelial tissue from 25 patients annotated with LGD regions were considered. We used a LOPOCV where, at each iteration, a new model was trained on 90% of the remaining data. We used 10% for the validation dataset to make sure that the classifiers didn't overfit on the training dataset.

The signals of each tile in the MSI dataset were mapped with a Gaussian distribution per patient with the Box-Cox method and all the *m/z* features were used. The classifier used was the same MLP classifier with the same hyperparameters computed during the training of grade classification.

To build a classifier using the H&E data, a CBAM architecture was used with ResNet50 as the backbone, trained on 2 epochs, re-trained, and validated using LOPOCV as described above. The workflow is presented in Fig. 3. In both MSI and H&E, a patient was classified as progressive or stable according to the majority vote of the predicted tiles.

2.11. Evaluation of the generated models

We reported the confusion matrices, the precision, the recall, the f1-score, and the number of samples per category for both validation and test datasets as calculated with the libraries sklearn and matplotlib within Python 3.7. For the classification of the tiles, the receiving operating characteristic (ROC) and the area under the ROC curves (AUC) were calculated. The confidence intervals of the AUC at 95% were computed with the DeLong algorithm, using the pROC library in R 3.6.3. The significance of every feature was calculated with the *feature_importances* function of sklearn based on the mean Gini decrease. The dice coefficient per tile was calculated for the test dataset to evaluate the grading performance: when the grade was correctly predicted, the dice was calculated with the delineations made by the pathologist and the full shape of the tile. In any other case, the dice coefficient was considered zero.

3. Results

Processing of the MSI data led to the detection of 321 on-tissue peptide signals, which were all used for training the models. Matching the MSI pixel size of 50 × 50 μm to the tile size in the histological images (96x96 pixels) made up a total of 144,823 tiles for the whole dataset. Table 1 inventories all the extracted tiles from all the images with the corresponding annotated tissue type and grade.

The epithelial tissue and stroma tiles dataset were equally balanced. However, the grades were highly unbalanced with the NDBE at 73%, LGD class 16%, and HGD class 11% of the epithelial dataset. These results were expected since NDBE and stroma classes can be found in all the samples, but dysplastic regions can only be found in specific areas.

3.1. Tissue type classification

The first task of our study was to use all 144,823 annotated pixels and classify the tiles as either epithelial or stroma regions.

For this, three models were trained on the MSI dataset and optimized with a grid search: the best parameters for MLP were a weight decay of 10^{-5}, with a batch size of 32, hidden layer sizes of (20, 10), maximum iteration of 1100, using Adam as optimizer. The best parameters for RF were the entropy as criterion with a maximum depth of 16, number of estimators at 200. For XGBoost we found that a learning rate of 0.1, number of estimators at 140, maximum depth of 9, a minimum child weight of 1 with a subsample at 0.8 worked best. Then the results were combined using a voting classifier with argmax function. This model gave an AUC of 0.89 (95% confidence interval (CI): 0.89–0.90) on the test dataset. The list of features importance computed on random forest, xgboost and their average is provided in supplement (Supplement Table 3). The model using H&E data as an input was trained for 2 epochs. The model achieved an AUC of 0.95 (95% CI: 0.94–0.95) on the test dataset.

The performances of both models were assessed with normalized confusion matrices on both validation (Supplement Fig. 4) and test dataset (see Fig. 4 a), and with the display of ROC curves (both Supplement Fig. 5). The classification reports (Table 2) show the balance of precision and recall in both models.

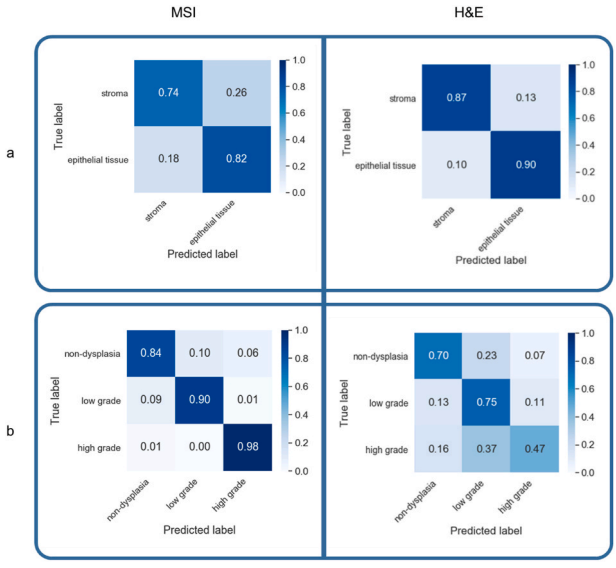

Fig. 4. Normalized confusion matrices of the test datasets for a prediction of the tissue type (a) and grade (b) using the mass spectrometry imaging (MSI) data (left) and the H&E data (right). Abbreviations used: MSI, mass spectrometry imaging; H&E, hematoxylin and eosin-stained tissue scans.

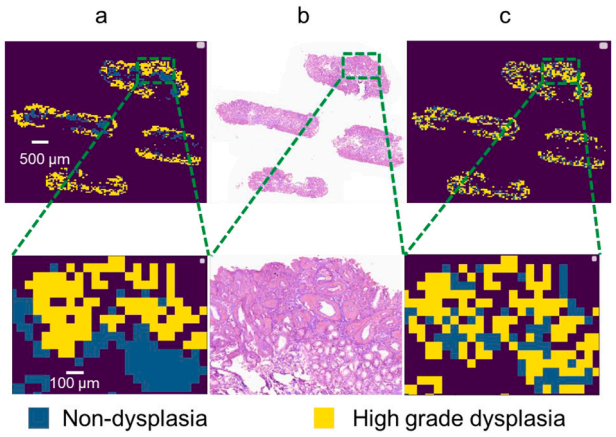

Fig. 5. Example of tile-based classification of the grade of a BE tissue: (a) ground truth label per tile of the full slide, (b) H&E of the full slide, and (c) prediction of tile labels on the full slide based on the classification made with MSI data Magnifications are shown in the lower row. Abbreviations used: H&E, hematoxylin and eosin-stained tissue scans; MSI, mass spectrometry imaging.

3.2. Tile-based dysplastic grade prediction

The second aim consisted in determining the grade of the tiles belonging to the epithelial regions (n = 16,920). Also, here 321 MSI features were used to train the model for grade prediction. The optimal weight decay was 10^{-6} with a batch size of 32, hidden layer sizes of (20, 20), a maximum iteration of 1100, and Adam as optimizer for the MLP model. The best parameters found for RF were entropy as criterion, a maximum depth of 16, number of estimators at 200 and for XGBoost we found that a learning rate at 0.1, a number of estimators at 140, a maximum depth at 9, a minimum child weight at 5, and using a subsample at 0.8 worked best. The results were combined using a voting classifier with argmax function.

The model performance as per micro-average AUC was 0.97 (95% CI: 0.96–0.97) on the test dataset. The comparison of performance on both the test and validation datasets allows us to observe that the model does not overfit. Because the test dataset was unbalanced but the training dataset was balanced, we can observe that the f1-scores are not consistent, achieving a worse performance on the low-grade tiles and the high-grade tiles (Table 2). The ROC curves of the micro-average and the macro-average ROC curves were also computed to give a better understanding of the overall prediction performance (Supplement Fig. 6). The list of features importance computed on random forest, xgboost and their average is provided in Supplement Table 3.

For the H&E-based classifier, the weights from the 17th epoch gave the best average accuracy on the validation dataset and were selected to evaluate the model. The micro-average AUC was 0.85 (95% CI: 0.85–0.86) on the test dataset. The results are visualized by confusion matrices of both the test (Fig. 4 b) and validation dataset (Supplement Fig. 4b). The ROC curves of the different grade predictions are provided in Supplement Fig. 6. The average dice coefficients calculated on the test dataset are given in Supplement Table 4. An example of a tile-wise full slide prediction for a patient diagnosed with LGD using the H&E classifier is given in Fig. 5.

3.3. Multi-modal prediction

We trained an MLP model by combining the features extracted from the DL model trained to distinguish the epithelial tissue from stroma with the MSI features. The model was trained with an L2-regularization coefficient of 0.1, a batch size of 32, hidden layer sizes of (10,10), a maximum of 1100 iterations and Adam as optimizer. Using this configuration, we obtained an AUC of 0.95 (95% CI: 0.95–0.95) on the test dataset. We repeated the same logic for the grade prediction and we obtained an optimal MLP with the following parameters: The optimal weight decay was 10^{-6}, with a batch size of 64, hidden layer sizes of (10, 10), maximum 1100 iterations, and Adam optimizer. The micro-average gave an AUC of 0.96 (95% CI: 0.96–0.96) on the test dataset. The corresponding confusion matrices are provided in Supplement Fig. 7.

3.4. Prediction of disease progression

Finally, models were trained for both modalities with the aim of forecasting the progression of low-grade dysplasia to a higher grade. The predictions of the models were thereby assumed to make statements on progression on a patient-level. The accuracies of the MSI-model and H&E-model were 0.72 (95% CI: 0.54–0.90) and 0.48 (95% CI: 0.28–0.68), respectively (Supplement Fig. 8).

4. Discussion

As a use-case, we chose the task of classifying the grade of dysplasia in BE and identify LGD lesions at high risk of progression. While deep learning has recently been used to detect neoplasia in BE using endoscopy [14], the application to histopathology is novel and ML-based classification of dysplasia grade or risk of progression has not been done with the combination of the two modalities as far as we know. When automatically classifying the tissue into epithelial and stromal structures, we could observe that the results are comparable between the validation (Supplement Table 5) and the test datasets (Table 2), indicating that the model was not overfitting the validation dataset very strongly for both models. Analysing the precision and recall scores, we observe similar results, which indicate well-balanced models. The model based on H&E data (AUC: 0.95 (95% CI: 0.94–0.95)) obtained a better performance at classifying epithelial tissue versus stroma than the model based on MSI data (AUC: 0.89 (CI: 0.89–0.90)). Given the clear visual differences between these tissue types, the superior performance of the H&E data is not surprising.

When predicting the grade of the dysplasia on a pixel-level, we observed similar prediction scores between the validation and the test datasets. The different grading implementation exhibited similar predictions with the model based on MSI features but the model based on H&E images obtained poor results for the classification of high-grade tiles. This allowed us to conclude that the models did not overfit on the validation dataset and the model based on MSI did not over-predict one class rather than another, but the model based on the H&E images did over-predict low grade tiles. With the classification reports (Supplement Table 5) we observed that the precision/recall was unbalanced on the validation dataset and on the test dataset (Table 2). This was caused by unbalanced data. In contrast with the previous task, we found that the model based on MSI data (AUC: 0.97 (CI: 0.96–0.97)) outperformed the model based on H&E data (AUC: 0.85 (CI: 0.85–0.86)). Moreover, the average dice coefficients obtained on the grades (Supplement Table 4) were lower but close to the true positive values obtained on the test dataset; thus confirming the capability of our model to reliably identify dysplastic regions. Indications for the potential of MSI data in similar scenarios can be found in previous MSI literature, where Elsner et al. [7] were able to distinguish metaplasia from carcinoma with an accuracy of 91% using a pattern of 31 proteins, albeit on a sample-level.

The use of multi-modal classifiers didn't improve the results obtained by using the modalities separately. H&E is better than MSI to distinguish the tissue type and MSI is better at predicting the grade of the tiles.

Despite endoscopic surveillance of patients with non-dysplastic BE or BE with low-grade dysplasia, up to 25% of EACs and HGDs are diagnosed within one year after last screening [15]. Our study shows the potential of MSI coupled to DL to identify patients that are higher at risk to progress to HGD with 72% accuracy. The performance of the model using MSI was similar to other studies such as the study of Kate and

co-workers who used clinical features in combination with p53 immunohistochemistry and histology criteria to obtain an AUC of 0.77 [16]. Our method was independent of clinical features and still obtained similar results. Our classifier could be therefore a useful addition to the existing surveillance strategies.

At the moment, the size of the cohort (57 in total) limits any strong clinical conclusions. A larger external validation sample dataset is therefore required to evaluate and confirm the predictions made by both approaches. In such a follow-up study, the predictive values of already known biomarkers in BE or EAC could be evaluated and compared to the MSI/H&E based approach. As mentioned, p53 is a biomarker for progression observed in 75% of the patients with multifocal aggregates of positive cells [13]. Another indicator for progression could be alpha-methyl-CoA racemase, an enzyme with high specificity and low sensitivity for the progression of indefinite for dysplasia towards dysplasia [17].

When comparing the classificatory power of H&E and MSI, an explanation of the improved classification capability of the model based on MSI data might be the fact that the dataset was annotated based on the H&E staining making it compile information from both approaches. Furthermore, the H&E dataset was not exploited to its full potential. We restricted the optimal parameter space of the H&E classification models by fixing the size of the patches (96x96 pixels) to the size of the MSI pixel (50 μm lateral pixel size), although similar tile sizes are being used in this field at 20x magnification [18]. In this context, it would be interesting to use multiple magnification levels for the classification of the data as done by Han et al. [19]. In contrast, we believe that using histomics to extract features at a cellular level combined with a ML model which classifies tissues instead of DL would help the pathologists to understand better what characteristics of the H&E are important for the classification [20].

Nevertheless, this study reveals the strength of each modality and their complementarity to address diagnostic and prognostic challenges in pathology using advanced ML. In our study, H&E provided higher accuracies for diagnostic purposes where the information is visually located in the histological phenotype. MSI, conversely, seems better suited for purposes where clinically relevant molecular alterations are present but still not morphologically visible at the microscopic level [21]. One can, therefore, envision a cascade-like application of the presented ML classifiers, where the optimum data and model are used for different tasks. In our example, the sequence would start with the H&E-based classifier for the detection of epithelial tissue regions. The MSI-based grade classifier would be applied to determine the grade of these epithelial structures, followed by the second MSI-based classifiers to predict if a patient's lesion is at risk of progressing.

In summary, the intention of our work was to investigate the complementarity and suitability of histological and molecular images using ML approaches for different clinical tasks in BE (tissue annotation, pathological grading, and patient prognosis). We have found that MSI can add valuable prognostic information beyond the histological level, whereas histology remains strong at the tissue annotation level. Based on these results we conclude that both, histological imaging and MSI, can complement each other for different clinical questions, which could ultimately help pathologists in diagnosing BE patient biopsies.

Declaration of competing interest

Dr. Lambin is co-inventor of two non-issues, non-licensed patents on Deep Learning-Radiomics-Histomics (N2024482, N2024889). Dr. Woodruff and Dr. Lambin have (minority) shares in the company Oncoradiomics. The rest of the authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was made possible through the support of Marie Skłodowska-Curie grant (PREDICT - ITN - No. 766276). BB and MML acknowledge the financial support of the European Union and the Dutch Cancer Society (ERA-NET TRANSCAN 2; Grant No. 643638). MML acknowledges the financial support of CAM (2018-T2/BMD-11561). Part of this work was conducted with financial support of the Province of Limburg through the LINK program. Authors furthermore acknowledge financial support from ERC advanced grant (ERC-ADG-2015 no. 694812 - Hypoximmuno, ERC-2018-PoC: 813200-CL-IO, ERC-2020-PoC: 957565-AUTO.DISTINCT). Authors also acknowledge financial support from SME Phase 2 (RAIL no.673780), EUROSTARS (DART, DECIDE, COMPACT-12053), the European Union's Horizon 2020 research and innovation programme under grant agreement: ImmunoSABR no. 733008, FETOPEN- SCANnTREAT no. 899549, CHAIMELEON no. 952172, EuCanImage no. 952103, TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY no. UM 2017–8295) and Interreg V-A Euregio Meuse-Rhine (EURADIOMICS no. EMR4). This work was supported by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018–2.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104918.

Author contribution statement

Manon Beuque performed all the ML analysis, analysed the results of the classifiers and wrote the manuscript. Marta Martin-Lorenzo performed the MSI experiments and H&E staining. Benjamin Balluff designed the MSI experiments, supervised the progression of the project, helped to pre-process the mass spectrometry data and made the high-resolution data handling and machine ML compatible. Henry Woodruff supervised the progression of the project and the writing of this article and guarantees the integrity of the analysis and results presented. Marit Lucas transferred the tissue an-notations from the H&E staining to the MSI data. Daniel M. de Bruin supervised the work of Marit Lucas. Janita van Timmeren helped with the analysis. Onno de Boer managed and organized the BE histopathology dataset (together with Sybren Meijer). Ron MA Heeren supervised the progression of the project. Sybren Meijer devised the project's aim, collected and provided the samples, and annotated the tissues. Philippe Lambin supervised the progression of the project. All authors have participated in writing the manuscript.

References

[1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics 70, CA Cancer J. Clin., 2020, 7–30, 1 2020.
[2] N.J. Shaheen, G.W. Falk, P.G. Iyer, L.B. Gerson, ACG clinical guideline: diagnosis and management of barrett's esophagus, Am. J. Gastroenterol. 111 (2016) 30–50.
[3] Y. Zhang, Epidemiology of esophageal cancer, World J. Gastroenterol. 19 (2013) 5598.
[4] L.C. Duits, K.N. Phoa, W.L. Curvers, F.J.W. Ten Kate, G.A. Meijer, C.A. Seldenrijk, G.J. Offerhaus, M. Visser, S.L. Meijer, K.K. Krishnadath, J.G.P. Tijssen, R. C. Mallant-Hent, J.J.G.H.M. Bergman, Barrett's oesophagus patients with low-grade dysplasia can be accurately risk-stratified after histological review by an expert pathology panel, Gut 64 (5 2015) 700–706.
[5] S.A. Gross, J. Kingsbery, J. Jang, M. Lee, A. Khan, Evaluation of dysplasia in Barrett esophagus, Gastroenterol. Hepatol. 14 (4) (2018) 233–239.
[6] P.-M. Vaysse, R.M.A. Heeren, T. Porta, B. Balluff, "Mass spectrometry imaging for clinical research–latest developments, applications, and current limitations, Analyst 142 (2017) 2690–2712.
[7] M. Elsner, S. Rauser, S. Maier, C. Schöne, B. Balluff, S. Meding, G. Jung, M. Nipp, H. Sarioglu, G. Maccarrone, M. Aichler, A. Feuchtinger, R. Langer, U. Jütting, M. Feith, B. Küster, M. Ueffing, H. Zitzelsberger, H. Höfler, A. Walch, MALDI imaging mass spectrometry reveals COX7A2, TAGLN2 and S100-A10 as novel prognostic markers in Barrett's adenocarcinoma, J. Proteomics 75 (8 2012) 4693–4704.

[8] R. Lazova, K. Smoot, H. Anderson, M.J. Powell, A.S. Rosenberg, F. Rongioletti, L. Pilloni, S. D'Hallewin, R. Gueorguieva, I. Tantcheva-Poór, O. Obadofin, C. Camacho, A. Hsi, H.H. Kluger, O. Fadare, E.H. Seeley, "Histopathology-guided mass spectrometry differentiates benign nevi from malignant melanoma, J. Cutan. Pathol. 47 (3 2020) 226–240.

[9] D.R.N. Vos, I. Jansen, M. Lucas, M.R.L. Paine, O.J. de Boer, S.L. Meijer, C.D. Savci-Heijink, H.A. Marquering, D.M. de Bruin, R.M.A. Heeren, S.R. Ellis, B. Balluff, Strategies for managing multi-patient 3D mass spectrometry imaging data, J. Proteomics 193 (2 2019) 184–191.

[10] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, CBAM: convolutional block Attention module, Computer Vision – ECCV (2018) 3–19, 2018.

[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 in *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[12] N. Tomita, B. Abdollahi, J. Wei, B. Ren, A. Suriawinata, S. Hassanpour, Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides, JAMA Netw Open 2 (2019) e1914645, 11.

[13] F. Yin, D. Hernandez Gonzalo, J. Lai, X. Liu, Histopathology of Barrett's esophagus and early-stage esophageal adenocarcinoma: an updated review, Gastrointestinal Disorders 1 (2019) 147–163.

[14] A.J. de Groof, M.R. Struyvenberg, J. van der Putten, F. van der Sommen, K. N. Fockens, W.L. Curvers, S. Zinger, R.E. Pouw, E. Coron, F. Baldaque-Silva, O. Pech, B. Weusten, A. Meining, H. Neuhaus, R. Bisschops, J. Dent, E.J. Schoon, P. H. de With, J.J. Bergman, Deep-learning system detects neoplasia in patients with barrett's esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking, Gastroenterology 158 (e4) (3 2020) 915–929.

[15] K. Visrodia, S. Singh, R. Krishnamoorthi, D.A. Ahlquist, K.K. Wang, P.G. Iyer, D. A. Katzka, "Magnitude of missed esophageal adenocarcinoma after Barrett's esophagus diagnosis: a systematic review and meta-analysis, Gastroenterology 150 (2016) 599–607, e7.

[16] F.J.C.T. Kate, F.J.C. ten Kate, D. Nieboer, F.J.W. ten Kate, M. Doukas, M.J. Bruno, M.C.W. Spaander, L.H.J. Looijenga, K. Biermann, Improved progression prediction in barrett's esophagus with low-grade dysplasia using specific histologic criteria, Am. J. Surg. Pathol. 42 (2018) 918–926.

[17] S.A. Sonwalkar, O. Rotimi, N. Scott, E. Verghese, M. Dixon, A.T.R.A. Axon, S. M. Everett, A study of indefinite for dysplasia in Barrett's oesophagus: reproducibility of diagnosis, clinical outcomes and predicting progression with AMACR (α-methylacyl-CoA-racemase): indefinite for dysplasia in Barrett's oesophagus, Histopathology 56 (5 2010) 900–907.

[18] N. Dimitriou, O. Arandjelović, P.D. Caie, Deep learning for whole slide image analysis: an overview, Front. Med. 6 (2019).

[19] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, S. Li, Breast cancer multi-classification from histopathological images with structured deep learning model, Sci. Rep. 7 (6 2017) 4172.

[20] M. Nalisnik, M. Amgad, S. Lee, S.H. Halani, J.E. Velazquez Vega, D.J. Brat, D. A. Gutman, L.A.D. Cooper, Interactive phenotyping of large-scale histology imaging data with HistomicsML, Sci. Rep. 7 (11) (2017) 14588.

[21] M. Aichler, M. Elsner, N. Ludyga, A. Feuchtinger, V. Zangen, S.K. Maier, B. Balluff, C. Schöne, L. Hierber, H. Braselmann, S. Meding, S. Rauser, H. Zischka, M. Aubele, M. Schmitt, M. Feith, S.M. Hauck, M. Ueffing, R. Langer, B. Kuster, H. Zitzelsberger, H. Höfler, A.K. Walch, Clinical response to chemotherapy in oesophageal adenocarcinoma patients is linked to defects in mitochondria, J. Pathol. 230 (8 2013) 410–419.