2022

Automatic detection of dysplasia on digitized head and neck pathology slides using convolutional neural networks

Rory Gilliland
Western University

Automatic detection of dysplasia on digitized head and neck pathology slides using convolutional neural networks

Rory Gilliland
Western University, Honours Specialization in Medical Biophysics
Baines Imaging Research Laboratory

Supervisors:

Dr. Aaron D. Ward
Western University
Baines Imaging Research Laboratory

Dr. Matthew Cecchini
Western University
London Health Sciences Centre

Salma Dammak
Western University
Baines Imaging Research Laboratory

Abstract

Introduction: Head and neck squamous cell carcinoma (HNSCC) is primarily treated with surgery. This surgery is guided by a pathologist, who intraoperatively scans removed tissue for cancer and dysplasia (precancerous epithelial tissue). Dysplasia is sometimes not removed because it can be difficult to detect. This may result in HNSCC recurrence, so there is great need to detect dysplasia more accurately. Machine learning (ML; the use of algorithms to train mathematical models) has been successfully applied to other medical detection problems, making it an attractive approach for this task. In this study, we aim to build and evaluate a convolutional neural network (CNN; a type of ML model) -based tool to detect dysplasia on HNSCC pathology slides.

Methods: Pathologist-contoured digitized frozen section slides from seventeen HNSCC surgeries were preprocessed and tiled in MATLAB and the Groovy programming language. In Python, the slides were used to train, validate, and optimize a VGG16 CNN in a transfer learning approach. Model testing was reserved for future work. The tool was evaluated with quantitative performance metrics and binary heatmaps integrated into the digital pathology tool, QuPath.

Results: The model's accuracy, sensitivity, specificity, and positive predictive value (PPV) in validation were 83%, 74%, 83%, and 1.3%, respectively. Validation area under the curve (AUC) was 0.84. Qualitative comparison of the validation heatmaps with corresponding pathologist annotations revealed correct detection of most dysplasia but abundant false positive detection of nondysplastic epithelial tissue.

Conclusions: Low PPV and frequent false positives on the heatmaps suggest that the current tool struggles to discriminate between dysplasia and normal tissue, making it inappropriate for clinical

use. The poor model performance may be explained by model limitations, small tile size, and substantial class imbalance. Encouragingly, much of the nondysplastic epithelium classified by the tool as dysplastic had some dysplasia-like characteristics, suggesting that the model identifies some pathologically meaningful features. Future work may seek to improve model performance by applying a precursor model to screen out non-epithelial cells, thereby rebalancing the classes. This work represents the first steps towards building a novel ML-based model to detect dysplasia on HNSCC surgery slides. If a model of this type can be improved, it could be used by pathologists to detect dysplasia more easily and accurately during HNSCC surgery, which would in turn increase the efficacy of this treatment.

Keywords

Head and neck cancer, digital pathology, machine learning, convolutional neural networks, dysplasia, detection, VGG16

Acknowledgements

The author would like to thank Salma Dammak, Dr. Aaron D. Ward, and Dr. Matthew Cecchini for their support on this project, and David DeVries, Carol Johnson, and Jen Coats for their coding assistance. The author and his supervisors have no sources of funding to acknowledge for this project.

Table of Contents

List of Figures

Figure 8. A sample HNSCC FS slide from the validation dataset. Note that this slide contains no pathologist-contoured dysplasia.

Figure 9. A collection of validation tiles representing tissue types on which the final model tended to produce true positives (row A), false positives (row B), false negatives (row C), and true negatives (row D). Tiles B-I and B-II contain nondysplastic epithelial tissue, tile B-III contains neutrophil-infiltrated stroma, and tile B-IV contains stroma. Tiles D-I and D-II contain stroma, tile D-III contains muscle tissue, and tile D-IV contains epithelial tissue. Tiles in rows A and C contain dysplasia.

Figure 10. Slide 871765 C1FS 1, one of two slides on which the model produced a PPV greater than 0.75 (PPV=0.85 for this slide). Observe that nearly all epithelial tissue present is dysplastic.

Figure 11. Slide 835092 A1FS 1, one of two slides on which the model produced a PPV greater than 0.75 (PPV=0.76 for this slide). Observe that nearly all epithelial tissue present is dysplastic.

Figure 12. The ROC curve and AUC of the model on the validation dataset from the epithelium-only exploratory experiment. The dashed black line corresponds to the ROC curve of a random guess.

Figure 13. A sample validation HNSCC FS slide with the model predictions. The model was trained and validated on (A) epithelium-containing tiles only or (B) all tiles. Observe that the epithelium-only model correctly identified most of the top-left portion of the tissue slice (black arrow on A) as nondysplastic, while the original model incorrectly classified most of it as dysplastic. Also note the tissue folding artifact (blue arrow on A), on which the original model produced several false positive predictions.

List of Abbreviations

HNSCC	head and neck squamous cell carcinoma
ML	machine learning
CNN	convolutional neural network
PPV	positive predictive value
AUC	area under the ROC curve
FS	frozen section
ROC	receiver operating characteristic

1 INTRODUCTION

1.1 Head and Neck Cancer

Head and neck squamous cell carcinoma (HNSCC) has been identified as the sixth most common type of cancer in the world[1]. In 2020, approximately 55,000 new diagnoses and 11,000 deaths were reported in the United States alone[2]. This form of cancer originates in the epithelial lining of the mouth, throat, or upper respiratory tract[1,3]. In this disease, normal epithelial cells progress through dysplasia (increased cell count and abnormal cell presence) and then invasive carcinoma (Fig. 1)[1]. Because of its anatomical localization, HNSCC often presents with particularly intrusive symptoms such as non-healing mouth sores, ear pain, and difficulty chewing and swallowing[1].

During treatment of HNSCC, it is particularly important to balance curative effect with function preservation due to the delicate nature of structures in the head and neck. Radiation therapy, chemotherapy, surgical resection, and combinations thereof have all been used successfully to strike this balance[4]. However, developments in minimally invasive resection techniques are broadening the potential for using surgery as the primary treatment[1].

Surgical resection of HNSCC needs to be extremely precise. All cancerous tissue must be removed to prevent re-establishment of the disease, while minimal healthy tissue should be removed to preserve function. To achieve this precision, a pathologist intraoperatively examines the outside edge – or the margin (Fig. 2) – of resected tissue for the presence of cancerous cells via frozen section (FS; a method of quickly cooling tissue for microscopic analysis). If cancerous cells are identified in the margins of removed tissue, it is likely that additional cancerous cells

remain in the surgical site. In this case, surgery and intraoperative FS will continue until the margins of resected tissue are cancer-free[5–7]. This procedure is designed to maximize the resected cancerous tissue while minimizing the removal of healthy tissue.

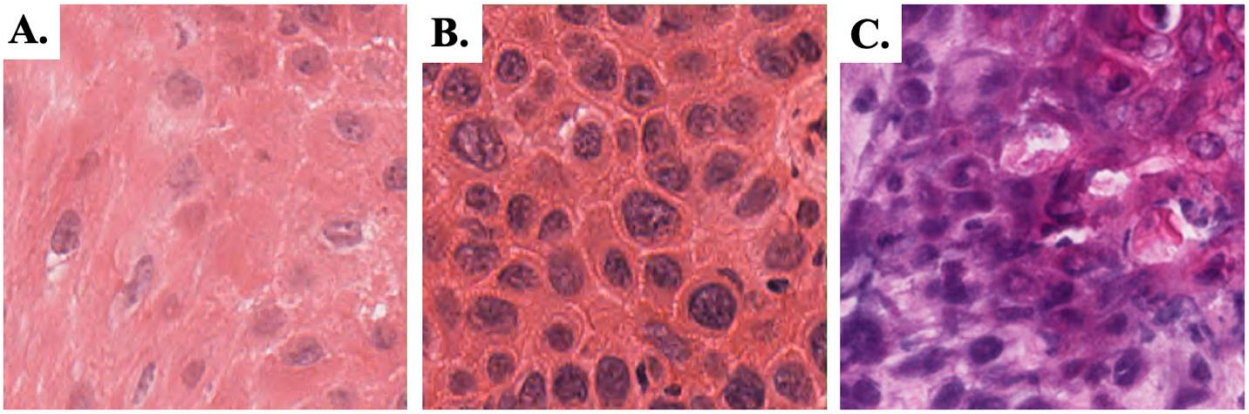

Figure 1. Examples of (A) normal epithelium, (B) dysplasia, and (C) invasive HNSCC.

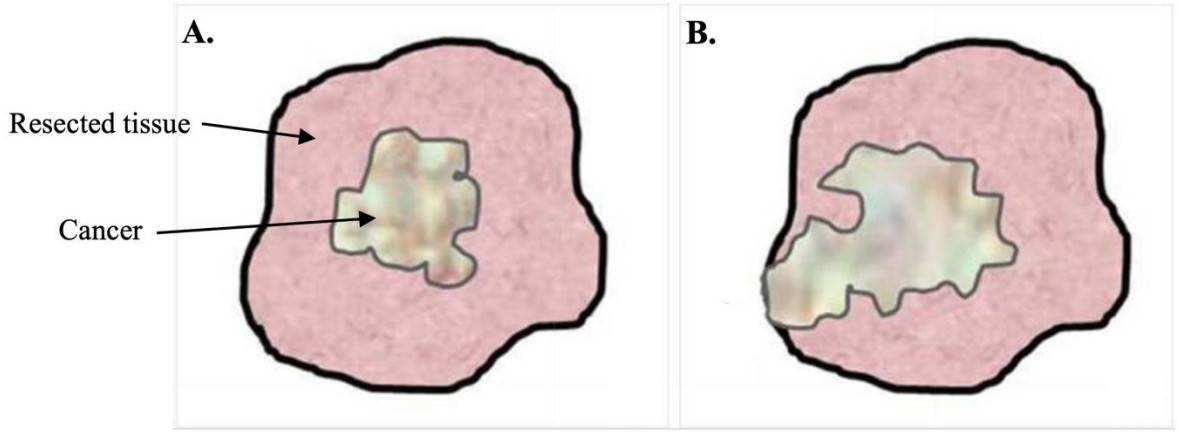

Figure 2. Illustrations of (A) a negative surgical margin and (B) a positive surgical margin. Note that in (B), the margin is positive because there are cancerous cells present at the outer edge of the resected tissue. A pathologist examines HNSCC surgical margins to determine if all cancer has been removed. A positive margin indicates that more cancer remains in the patient, and in this case, the surgeon must continue to remove tissue until a negative margin is achieved. Illustration adapted from Lee et al. (2012)[8] under the article's Creative Commons License (https://creativecommons.org/licenses/by-nc/3.0/).

1.2 Challenges in the HNSCC Resection Procedure

The pathologist's role in the HNSCC resection procedure is extremely important. Failing to correctly identify disease in resected tissue margins is strongly associated with recurrence and mortality[7,9,10]. However, two factors complicate the pathologist's job.

First, the assessment of resected tissue margins must be performed quickly to minimize the time the patient spends in the operating room. It has been shown that longer surgical procedures tend to result in a higher frequency of complications such as infection[11]. This time constraint

combined with the demand for accurate classifications can place considerable stress on the pathologist[6].

The second complication for the pathologist is identifying dysplastic tissue. Dysplasia is an intermediate state between normal epithelium and invasive carcinoma[1,12]. It is a known risk factor for the development of invasive carcinoma, and its presence in the margin of a resection may be a predictor of local recurrence[1,12–15]. Thus, HNSCC resection procedures must aim to remove all cancerous and dysplastic tissue. Unfortunately, dysplastic cells can be difficult to identify. The differences between dysplastic and normal cells – especially normal cells which are inflamed or physically stressed – can be slight. This can make recognition challenging, especially for pathologists who are not specialized in HNSCC[16]. On top of this, the grading of dysplasia severity (which informs whether the tissue needs to be removed) is subject to substantial intra- and inter-observer variability[17].

Time constraints and dysplastic tissue identification are significant difficulties for pathologists assisting in HNSCC resection procedures. These challenges, as well as the lack of solutions to them in the literature, are the motivation for this study. There is a need for developing methods to help pathologists detect dysplastic cells on FS slides quickly and accurately. This would (1) minimize the amount of healthy tissue removed, (2) maximize the amount of cancerous and dysplastic tissue removed, and (3) reduce the duration of the operation. We aim to leverage recent advancements in digital pathology and machine learning to assist in this detection task.

1.3 Digital Pathology

Digital pathology uses imaging technology to view and analyze pathology slides on computers. Glass pathology slides are digitized by scanning them with a camera onto a computer in small tiles, which are then stitched together into whole slide images. Its popularization in the last two decades has fundamentally modernized clinical pathology. Digitized slides can be shared across multiple clinics for consultation, vast amounts of pathological data can be stored efficiently, and everyday pathology can be performed remotely[18]. Perhaps the most interesting opportunity afforded by digital pathology is the application of artificial intelligence. Digitization allows for quantitative analysis, and data can be extracted and used to build machine learning (ML) models[18–20].

1.4 Machine Learning

ML is a subfield of artificial intelligence that uses algorithms to train models to perform tasks based on input data. In the supervised approach to ML, many samples of features (input data) and their corresponding labels (desired output data) are presented to the model. During the training process, the model is tuned to recognize patterns in the features that are associated with the labels. When the trained model is subsequently presented with unlabeled features of new samples, it attempts to predict the label on its own based on the previously learned patterns[21]. The applications of ML are nearly limitless. Its use in the field of medicine has facilitated early prediction of diabetes mellitus, automatic detection of intracranial hemorrhages, and prediction of recurrence in non-small cell lung cancer[22–24].

The building of ML models can generally be split into three steps. The first step is data preparation, which involves the collection of features and labels for many samples of data. Features are chosen and input based on knowledge of factors that may have some predictive value for the label[21,25]. For example, to build a classical ML model to predict the diabetes status of a patient (the label), it might be valuable to use blood sugar concentration as a feature. Feature selection is then used to systematically limit the final number of features given to the model based on their predictive value. This is usually performed to simplify the model or avoid overfitting[21,26].

The second step in building ML models is training the algorithm. Training is performed on a distinct subset of the full dataset, called the training data. This subset is kept separate from the testing data[21]. During training, the algorithm is fit to the features and labels of the training data. A third subset of data, called validation data, is sometimes used after training to choose from a set of possible models or hyperparameter (pre-set values that change the behaviour of the model) combinations being considered for the task[21].

The final step in building a ML model is testing the model using the testing data. This step is designed to evaluate the model's performance on unseen data, as would be expected in a real-life application[21]. The model is given only the features of the testing data and must predict the labels itself. The predicted labels can then be compared to the actual labels of the testing data to calculate performance metrics. For a classification task, several performance metrics exist, such as accuracy, sensitivity, specificity, positive predictive value (PPV), and area under the receiver operating characteristic curve (AUC). An in-depth description of these metrics is given by Fawcett[27].

The choice of algorithm used in the model is worth some discussion. Many kinds of algorithms exist, and they vary drastically in both complexity and the types of problems they are

best suited for. Classical algorithms such as linear regression, support vector machines, and decision trees tend to be simpler[21,28,29]. They are therefore computationally cheap and better suited for smaller datasets. Deep learning algorithms, on the other hand, tend to be much more complex and computationally expensive. However, these algorithms are capable of handling complex problems with large datasets, and indeed they have consistently performed well on other digital pathology problems[30]. This strong performance makes deep learning algorithms particularly attractive as candidates to help pathologists detect dysplasia on digitized HNSCC FS slides[21].

1.5 Deep Learning

Deep learning is a subset of ML that has gained attention in the last two decades. Neural networks, the model type used in deep learning, feeds input data through networks of interconnected artificial neurons to make predictions. Each artificial neuron operates on input data (x_i) using certain weight values (w_i), a bias value (b), and an activation function (σ) to produce an output (y):

$$y = \sigma\left(\sum_{i=0}^{n-1} x_i w_i + b\right)$$

The output of each artificial neuron then becomes an input for other artificial neurons in the network until the last neuron(s) produce the final output of the model. Training neural networks involves iteratively adjusting the weights and biases of the network's neurons to emphasize or mute certain patterns in the input data that are associated with the labels[21]. The operation and training of these algorithms is dictated by several hyperparameters, such as learning rate, batch size, and epoch number[31–33].

A convolutional neural network (CNNs) is a type of neural network that can be applied directly to images. Each pixel of an image is an input into the algorithm, and the network of neurons works to discern the spatial relationships between pixels that are associated with the labels. CNNs use special convolution and pooling layers to extract this spatial information[21]. A common logistical practice in building CNN models is tiling, wherein large images are divided into small tiles a few hundred pixels across[34]. This allows for the adaptation of CNN models to problems where the input image size would otherwise be too large to load in random access memory, as is often the case in digital pathology.

Full-scale CNNs can take a great deal of computational power and time to train. One successful CNN model designed by the Visual Geometry Group at the University of Oxford, VGG16, took more than two weeks to train[35]. A common approach to reduce computation time is to use transfer learning, wherein existing successful CNNs are applied to new problems. In this approach, only the last few layers of pre-trained algorithms like VGG16 need to be retrained on the new data. Transfer learning leverages the success of well-trained models without the computational expense of training them from scratch[32,36].

Deep learning models are largely built in the same way as classical ML models: the algorithm is trained and validated to fine-tune the model before testing it and measuring performance. One key difference between deep learning and classical ML is the handling of features. Where classical ML requires predefined features, deep learning algorithms use raw data as input. The deep learning algorithm is free to decide what raw input data to use and how to use it[21]. This allows CNNs to identify visual patterns that may not be obvious to the human eye. This can be an important advantage of using CNNs over classical ML algorithms for visual tasks.

CNNs such as VGG16 have been successfully applied to numerous medical detection tasks[37,38]. This history of success and the key theoretical advantages of CNNs over classical ML algorithms are what motivate the use of CNNs in the present study. We aim to build a CNN-based tool to assist pathologists in HNSCC resection procedures. We hypothesize that a CNN-based tool can be developed to automatically detect dysplastic cells on digitized HNSCC FS slides. Such a tool could assist pathologists in differentiating dysplastic and normal tissue during HNSCC resection procedures. This could lead to faster and more accurate decisions by the pathologist, and ultimately more successful HNSCC resection operations. CNNs have previously been used to detect HNSCC on digital pathology slides[39], but to our knowledge, this research has not been extended to the more difficult task of detecting dysplasia on digitized HNSCC FS slides.

Our primary objective is to build this tool and evaluate if CNNs are appropriate for this task. We plan to gather pathologist annotated digitized HNSCC FS slides, perform tiling on these slides, and split the tiles into training, testing, and validation datasets. Using MATLAB R2020B (The MathWorks, Natick, MA) and Python 3.7 (Centrum voor Wiskunde en Informatica, Amsterdam), we will train and validate a VGG16 CNN classifier model using transfer learning. Finally, we aim to assemble the CNN's predictions into whole-slide binary heatmaps based on the model's predictions of dysplasia presence. These heatmaps will be callable from the popular open-source digital pathology tool, QuPath, for pathologist visualization[40]. Our secondary objective is to collaborate with a pathologist to identify opportunities to improve the tool.

2 METHODS

2.1 Data Preparation

FS slides from 17 anonymized HNSCC resection patients between 2018 and 2021 were collected from the pathology archives at the London Health Sciences Centre. Examination for dysplasia was reported by Dr. Christopher Tran, who was a surgical pathology fellow at the time. The accompanying pathology reports were used to identify the slide HNSCC content. Slides contained either (1) dysplasia, (2) invasive HNSCC, (3) both dysplasia and invasive HNSCC, or (4) neither (the negative case). A total of 258 slides were collected for these patients. Slides were digitized using an Aperio AT slide scanner at a pixel-level resolution of 0.4961μm. Those containing dysplasia were contoured by Dr. Tran in QuPath and confirmed by a board-certified pathologist, Dr. Matthew Cecchini. A sample of the pathologist-contoured digital slides is available in Fig. 3.

Slides were then divided into 224 x 224 pixel (111 x 111 μm) nonoverlapping tiles using the Groovy programming language[41]. Tiles were screened in MATLAB for tissue content using thresholding: tiles which had (1) over 85% of their pixels exceeding an intensity value of 232 and (2) a pixel intensity variance exceeding 148 were kept and used as data for building the ML model. These criteria were determined by inspection of 20 representative tiles from the training dataset. A total of 673,344 tiles met these criteria. A sample of these tiles is available in Fig. 4. Tiles were then labeled as dysplastic if the tile contained any amount of dysplasia based on the pathologist contours. All other tiles were labeled as nondysplastic.

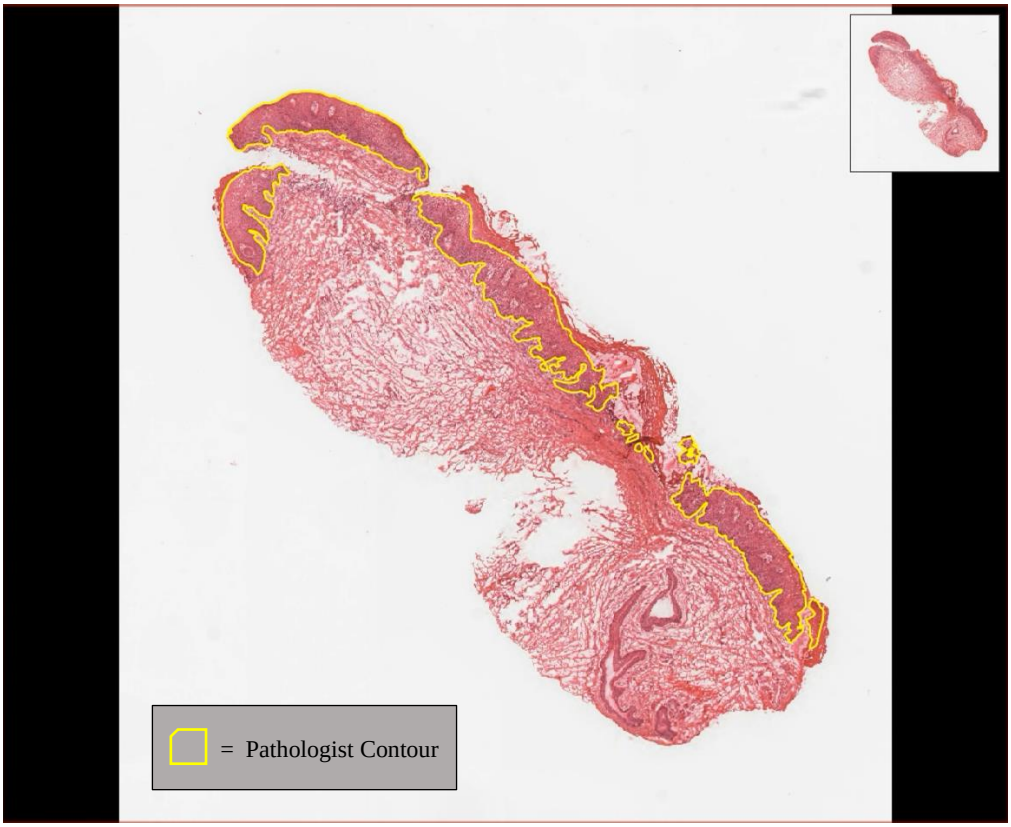

Figure 3. A sample digitized HNSCC surgery FS slide with pathologist-contoured dysplasia shown in yellow.

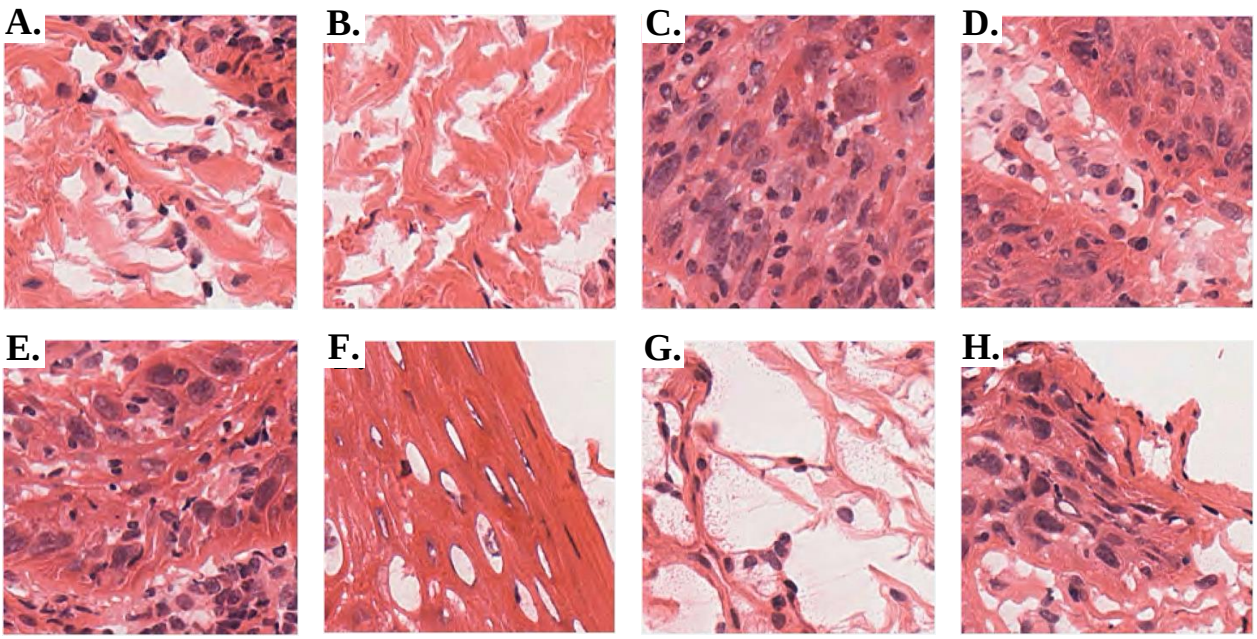

Figure 4. A sample of 224 x 224 pixel tiles taken from digitized HNSCC surgery FS slides. Tiles C, D, E, and H contain dysplasia; tiles A, B, and G contain subepithelial connective tissue; and tile F contains superficial nondysplastic epithelium.

The tiles were manually split by patient into training, testing, and validation datasets, targeting approximately a 50/25/25 split and similar proportions of dysplastic tiles. A breakdown of tiles from each patient is included in Table 1, and the resulting composition of each dataset is summarized in Table 2. One patient's tiles included over 75% of the full dataset's dysplastic tiles; this patient was assigned to the training dataset.

Table 1. A breakdown of the post-screening total slide count, total tile count, and positive tile count for each patient. Note that slides from patient 3591737 contain most of the full dataset's positive tiles.

Patient Anonymized ID	Total Slides	Total Tiles	Positive Tiles
3496325	15	89,208	973
3504266	18	67,753	0
3507259	11	20,458	0
3587371	18	71,153	0
3591737	15	95,871	6,375
8335486	15	131,629	640
835092	12	19,143	183
835537	7	21,337	0
839362	9	7,812	13
8707630	11	10,966	42
8717228	8	43,961	0
871765	11	21,125	209
8719032	2	7,126	42
87314	3	11,812	0
8786897	1	1,793	0
8787119	13	20,276	0
8788542	14	31,921	0

Table 2. A summary of the training, validation, and testing datasets following manual splitting of the tiles by patient. Split refers to the percentage of all tiles apportioned to a dataset, and percentage positive is the fraction of tiles in a dataset that contain dysplasia. Observe that the training dataset has a substantially higher percentage positive than validation or testing because it includes a patient that over 75% of the positive tiles came from.

Dataset	Total Slides	Total Tiles	Positive Tiles	Split (%)	Percentage Positive (%)
Training	93	376,182	7,348	55.87	1.95
Validation	63	145,909	476	21.67	0.33
Testing	27	151,253	653	22.46	0.43

2.2 Training and Validation

The VGG16 CNN model was trained for this problem with a transfer learning approach. Only the last layer of neurons in this pre-trained CNN were allowed to be trained. The training and validation datasets were used to train and validate the algorithm using in-house scripts in MATLAB and Python (Appendix A). The algorithm's output predictions on validation data (as confidence levels of dysplasia presence) were then compared to the actual validation labels to quantitatively assess validation performance. A receiver operating characteristic (ROC) curve was plotted for the validation dataset, and AUC was calculated. Accuracy, sensitivity, specificity, and PPV of the model on this dataset was calculated at the threshold producing the point on the ROC curve closest to the top left corner. Accuracy, sensitivity, specificity, and PPV were also calculated on a per-slide basis using the same threshold.

2.3 Model Optimization and Testing

The algorithm was alternately trained and validated several more times to search for the optimal training parameters. Validation performance metrics were calculated as described above after each validation. Performance on each validation was used to systematically adjust the model for subsequent runs, with the goal of optimizing validation AUC and accuracy. Adjustments were made to algorithm hyperparameters, which include learning rate, epochs, and batch size. In general, these hyperparameters were optimized one at a time, with batch size optimized first, epochs second, and learning rate last. The hyperparameter values attempted were 50, 100, 200, 300, and 400 for batch size; 1, 2, and 3 for epochs; and 0.001, 0.002, 0.01, and 0.02 for learning rate.

Once the model with the best performing combination of hyperparameters was identified, bootstrapping was used to resample the model confidences on the validation tiles 1,000 times. The AUC was calculated for each resampled validation dataset, and a Wilcoxon signed rank test was used to compare these AUCs to 0.50 (the AUC of a random guess). A *p*-value of less than 0.05 was considered statistically significant. Additionally, the performance of the final model its performance on the validation dataset was visualized by comparing binary heatmaps (see section 2.4) for each validation slide to pathologist contours of dysplasia.

Testing of the model on the unseen testing dataset was reserved for future work that may achieve sufficiently strong validation performance.

2.4 Output Visualization in QuPath

A binary heatmap of the model's dysplasia detection for a slide was produced by mapping the model's tile predictions to their corresponding locations on the slide. Tiles on each slide predicted to be dysplastic by the model were represented by small square annotations on top of the slide display in QuPath. The Groovy programming language was used to create the heatmaps. Binary heatmaps were produced for the final model's predictions on the validation dataset as well as the testing dataset. Model performance on the validation and testing datasets was assessed qualitatively by comparing the binary heatmaps to the pathologist's dysplasia contours on each slide in QuPath.

2.5 Epithelium-Only Exploratory Experiment

Trends in the types of tissues classified correctly or incorrectly by the model were identified in collaboration with Dr. Cecchini. Based on observed patterns, the general study design described up to this point was re-applied to only the epithelial tissue on a small subset of the collected FS HNSCC resection slides.

For this exploratory experiment , a set of 11 slides was randomly selected from the training and validation datasets. Epithelium was manually contoured on these slides by the author, and 224 x 224 pixel (111 x 111 μm) tiles containing any of the epithelium contour were extracted as described in Section 2.1. Screening of the tiles for tissue content was not performed. Resulting tiles that contained any pathologist-contoured dysplasia were labeled as dysplastic, while all other tiles were labeled nondysplastic. Tiles were split by patient into training and validation datasets using the approach described in Section 2.1. The breakdown of the epithelium-containing tiles

from each patient included in this exploratory experiment is detailed in Table 3, and the resulting composition of each dataset is presented in Table 4.

Training, validation, optimization, and performance assessment of the VGG16 CNN was then performed for this exploratory experiment as described in Sections 2.2, 2.3, and 2.4. Testing was not performed on the final model from this experiment.

Table 3. A breakdown of the slide count, total tile count, and positive tile count for each patient from the epithelium-only exploratory experiment.

Patient Anonymized ID	Total Slides	Total Tiles	Positive Tiles
3496325	2	443	171
3504266	1	821	0
3591737	3	1,653	308
835537	1	885	0
8707630	1	134	46
871765	1	191	0
8719032	1	315	43
8788542	1	386	0

Table 4. A summary of the training and validation datasets following manual splitting of the tiles by patient for the epithelium-only exploratory experiment. Split refers to the percentage of all tiles apportioned to a dataset, and percentage positive is the fraction of tiles in a dataset that contain dysplasia. Note the larger percentages positive in this exploratory experiment compared to those in Table 2.

Dataset	Total Slides	Total Tiles	Positive Tiles	Split (%)	Percentage Positive (%)
Training	6	3,493	354	72.35	10.14
Validation	5	1,335	214	27.65	16.03

3 RESULTS

3.1 Quantitative Validation Performance

The batch size, epochs, and learning rate values that resulted in the best model validation performance are 300, 1, and 0.002, respectively. The accuracy, sensitivity, and specificity of this model on the whole validation dataset are 83%, 74%, and 83%, respectively. Notably, the model's PPV on this dataset is relatively low at 1.4%, meaning that only around 1 in 100 tiles predicted to be positive by the model truly contained dysplasia. Fig. 5 presents the model's ROC curve for the validation dataset. The model's AUC for this dataset is 0.84. The mean AUC for the 1,000 bootstrapped validation datasets is 0.84. This value is significantly different from the theoretical

AUC of 0.50 that represents random guesses by a Wilcoxon signed rank test with alpha=0.05 (p<0.0001).

Histograms of the model's validation accuracy, sensitivity, specificity, and PPV by slide are reported in Fig. 6. In general, the performance metrics on the slide level are similar in trend to those on the whole dataset: PPV is low, while accuracy, sensitivity, and specificity are relatively high. Two exceptions are the validation slides 835092 A1FS 1 and 871765 C1FS 1, on both of which the model has PPV values above 0.75 (note the outliers in Fig. 6d).

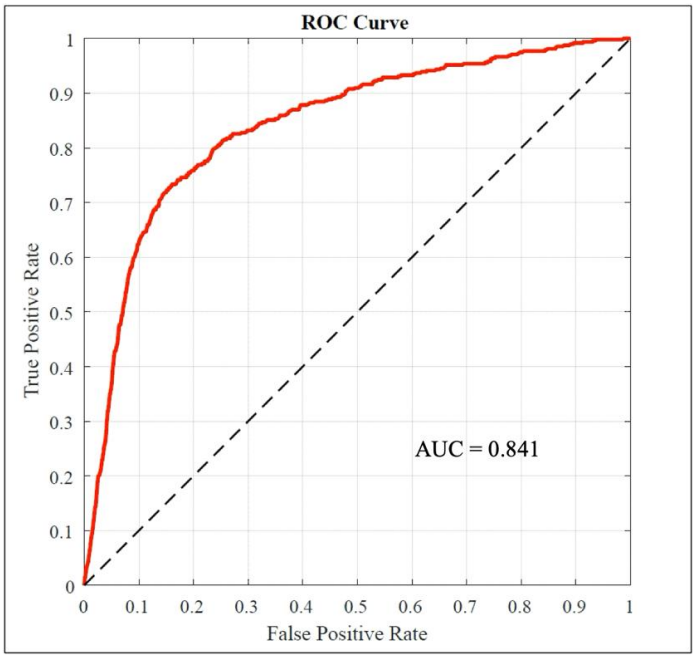

Figure 5. The ROC curve and AUC of the final model on the entire validation dataset. The dashed black line corresponds to the ROC curve of a random guess.

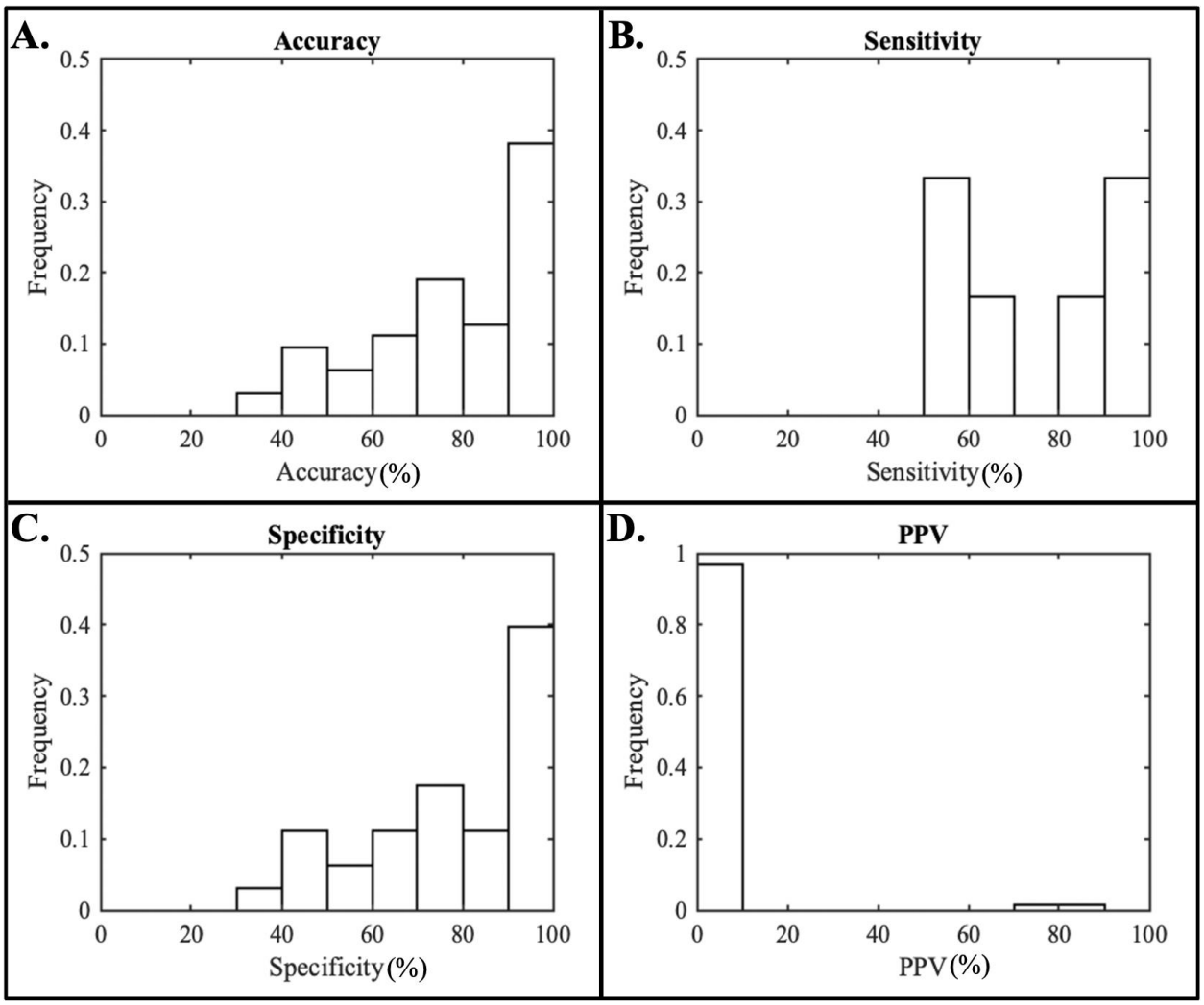

Figure 6. Frequency histograms of the final model's accuracy (A), sensitivity (B), specificity (C), and PPV (D) on each validation slide. Only six slides are represented in B, as only this number of validation slides contained dysplasia.

3.2 Qualitative Validation Performance

Figs. 7 and 8 present the heatmaps and pathologist contours for two slides representative of the model's validation performance. In general, the validation slide heatmaps reveal that most regions of dysplasia are detected by the model. However, the model also produces substantial false

positive detection, which varies in amount from a few dozen tiles to nearly entire slices of tissue. Considerable false positive detection tends to occur whether dysplasia is present on the slide or not. However, on slides where dysplasia is present, the area of the slide covered by the model's positive detections is often much larger than the dysplasia contour.

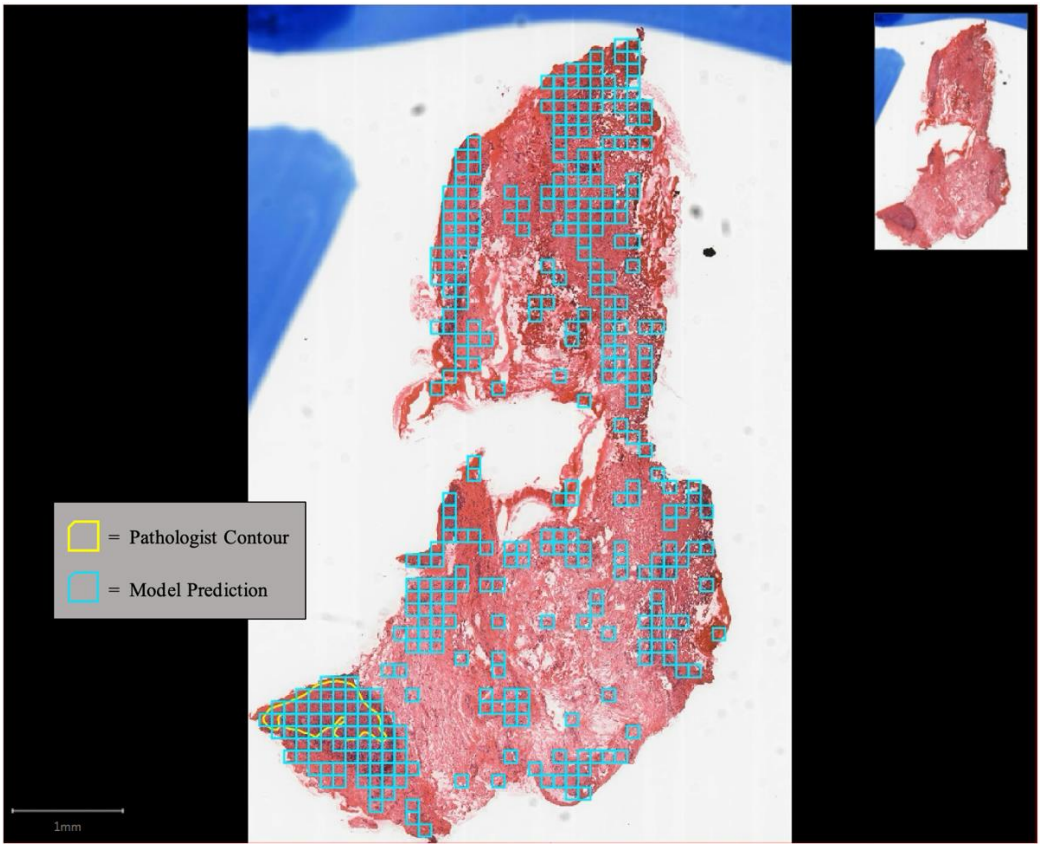

Figure 7. A sample HNSCC FS slide from the validation dataset.

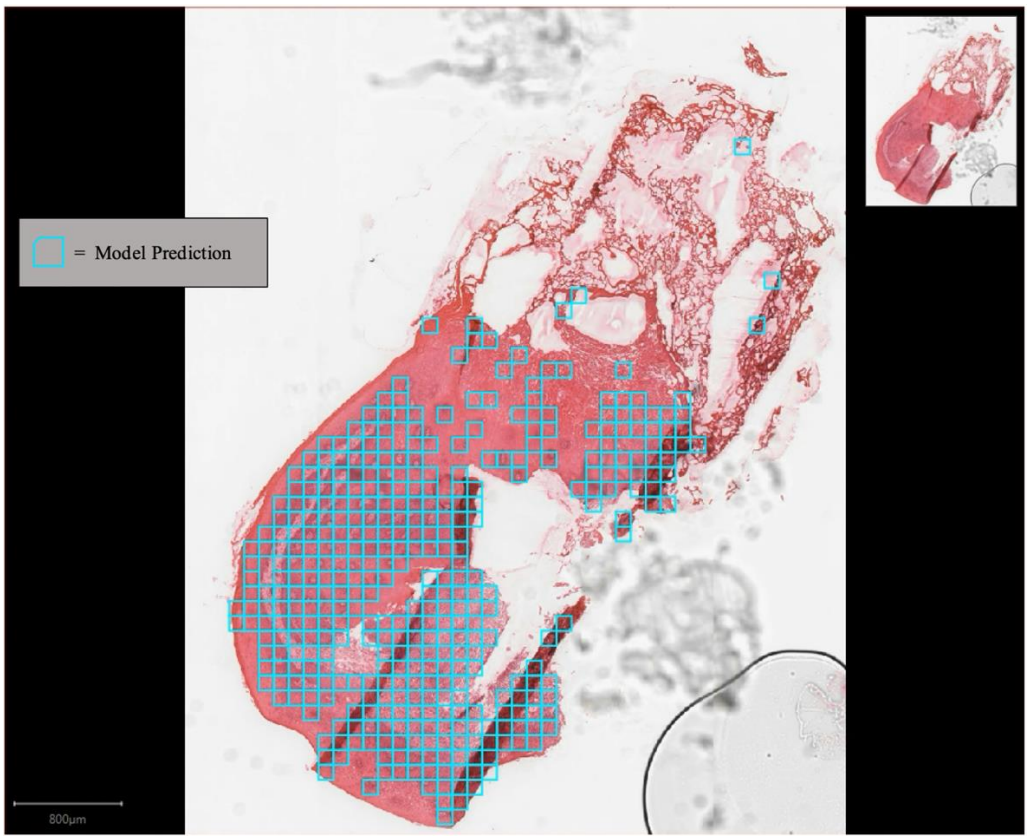

Figure 8. A sample HNSCC FS slide from the validation dataset. Note that this slide contains no pathologist-contoured dysplasia.

A collection of validation tiles exemplifying tissue on which the model produced true positives, true negatives, false positives, and false negatives is given in Fig. 9. Most false positives produced by the model are on nondysplastic epithelial tissue (Fig. 9b). Epithelial tissue on which the model produced false positive detection tends to be darker in colour, less organized in cell arrangement, and more densely packed with cells. Other false positive detections by the model sometimes involves neutrophil-infiltration (Fig. 9b). Dysplastic tiles that are not detected by the model (false negative tiles) tend to contain white regions (Fig. 9c). Nondysplastic tissue types that are usually classified correctly by the model include stroma and muscle (Fig. 9d).

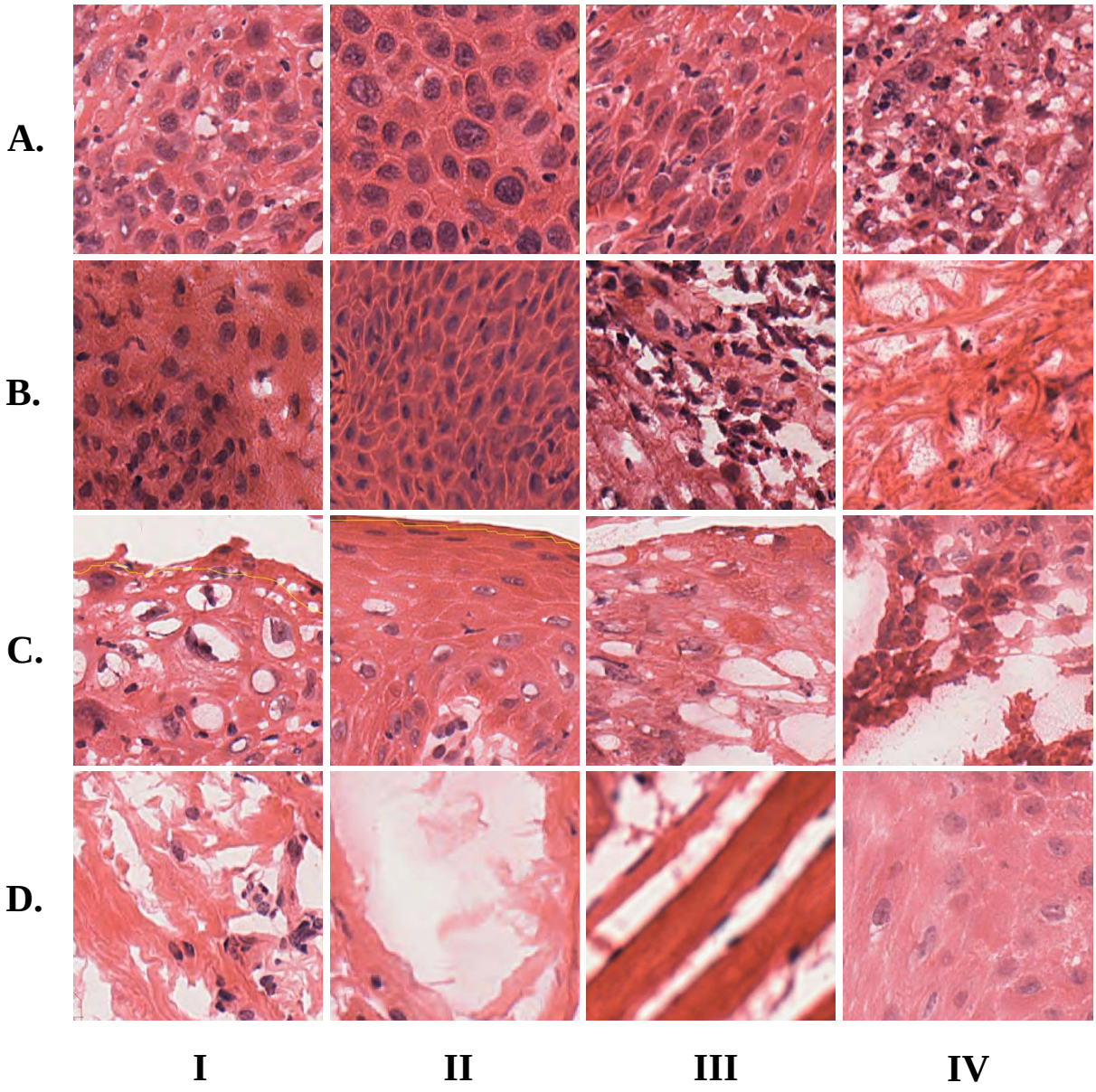

Figure 9. A collection of validation tiles representing tissue types on which the final model tended to produce true positives (row A), false positives (row B), false negatives (row C), and true negatives (row D). Tiles B-I and B-II contain nondysplastic epithelial tissue, tile B-III contains neutrophil-infiltrated stroma, and tile B-IV contains stroma. Tiles D-I and D-II contain stroma, tile D-III contains muscle tissue, and tile D-IV contains epithelial tissue. Tiles in rows A and C contain dysplasia.

The heatmaps and pathologist contours for the two slides on which the model achieved notably high PPV values (835092 A1FS 1 and 871765 C1FS 1) are presented in Figs. 10 and 11. The model identifies nearly all the epithelial tissue in these slides as dysplastic, and most of the non-epithelial tissue as nondysplastic. Unlike most other slides, though, most of the epithelial cells on these slides are dysplastic. Non-epithelial tissue on these slides tends to be tissue that is visually distinct from epithelium, such as stroma.

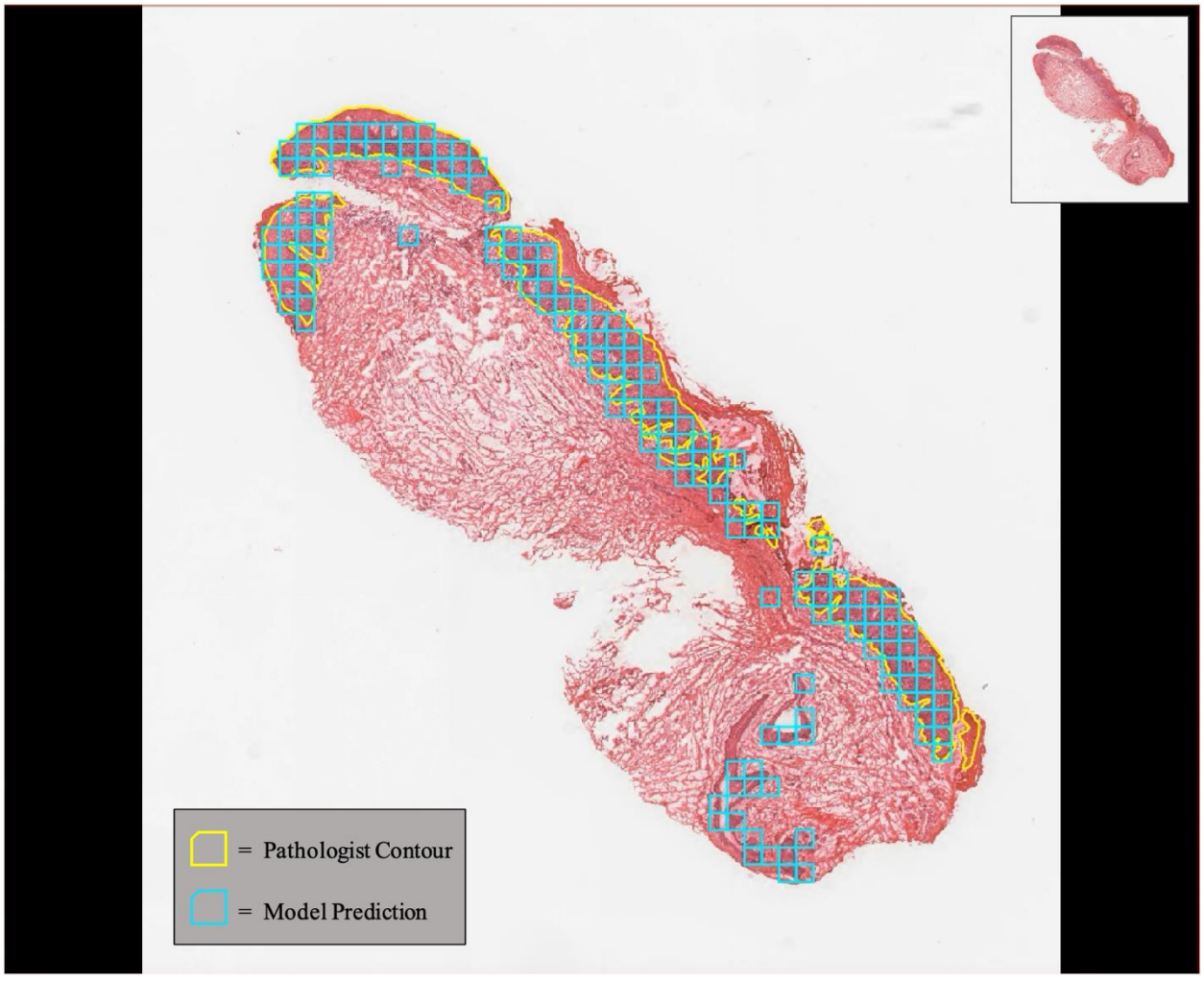

Figure 10. Slide 871765 C1FS 1, one of two slides on which the model produced a PPV greater than 0.75 (PPV=0.85 for this slide). Observe that nearly all epithelial tissue present is dysplastic.

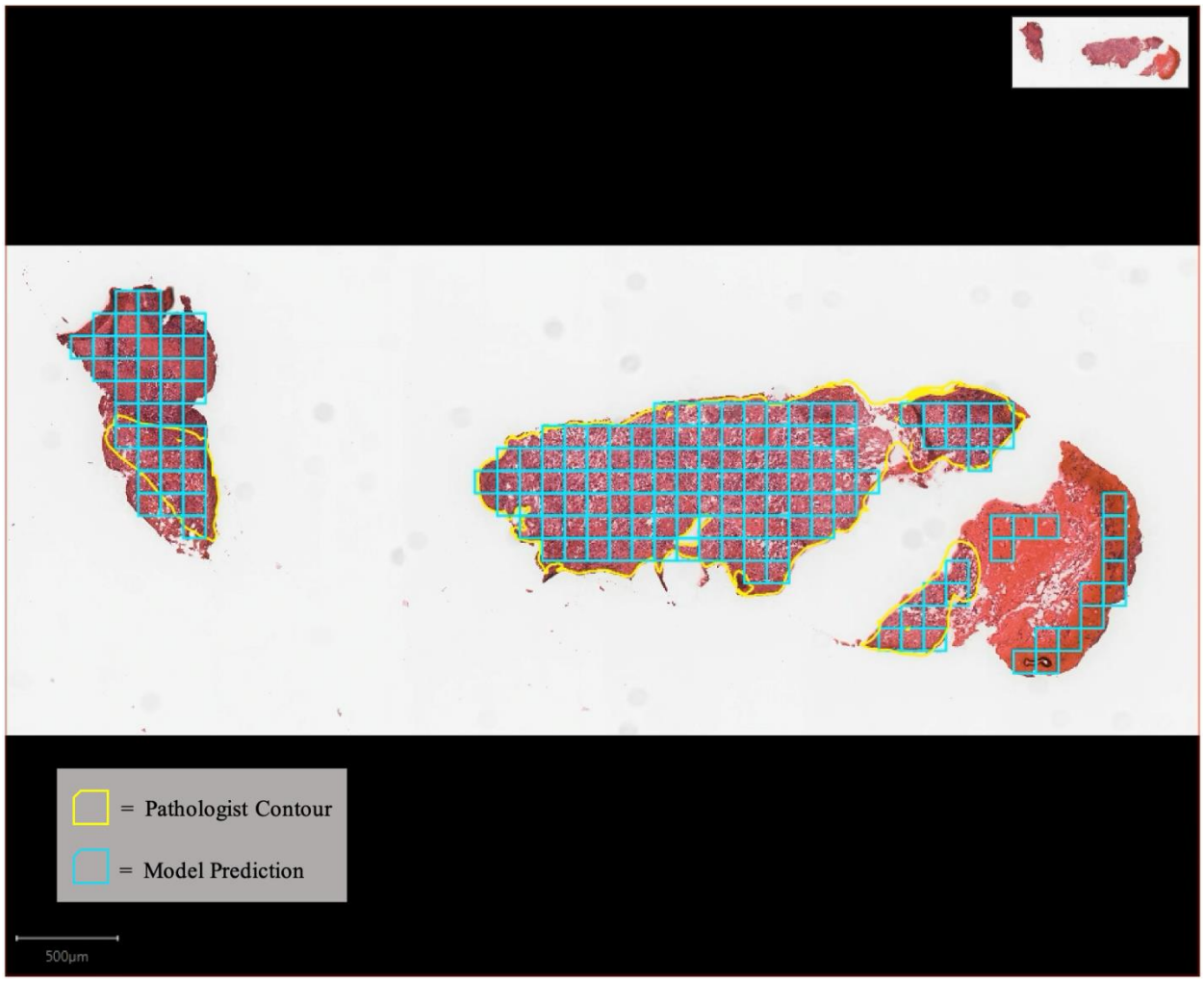

Figure 11. Slide 835092 A1FS 1, one of two slides on which the model produced a PPV greater than 0.75 (PPV=0.76 for this slide). Observe that nearly all epithelial tissue present is dysplastic.

3.3 Epithelium-Only Exploratory Experiment

Batch size, epochs, and learning rate values of 300, 2, and 0.002 (respectively) resulted in the best model validation performance in detecting dysplasia on the epithelium-containing tiles. Accuracy, sensitivity, and specificity of this model on the validation dataset are 84%, 86%, and 84%, respectively. The PPV of the model is 48%. Fig. 12 presents the model's ROC curve for the

validation dataset of this exploratory experiment. The model's AUC for this dataset is 0.92. The model's heatmap, the pathologist-contoured dysplasia, and the epithelium contour for a sample validation slide is presented in Fig. 13a.

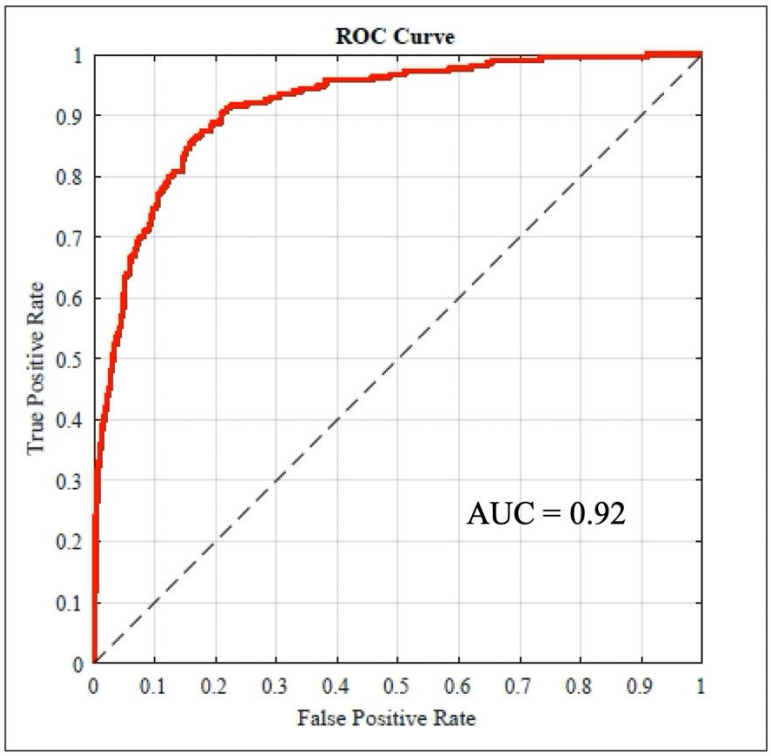

Figure 12. The ROC curve and AUC of the model on the validation dataset from the epithelium-only exploratory experiment. The dashed black line corresponds to the ROC curve of a random guess.

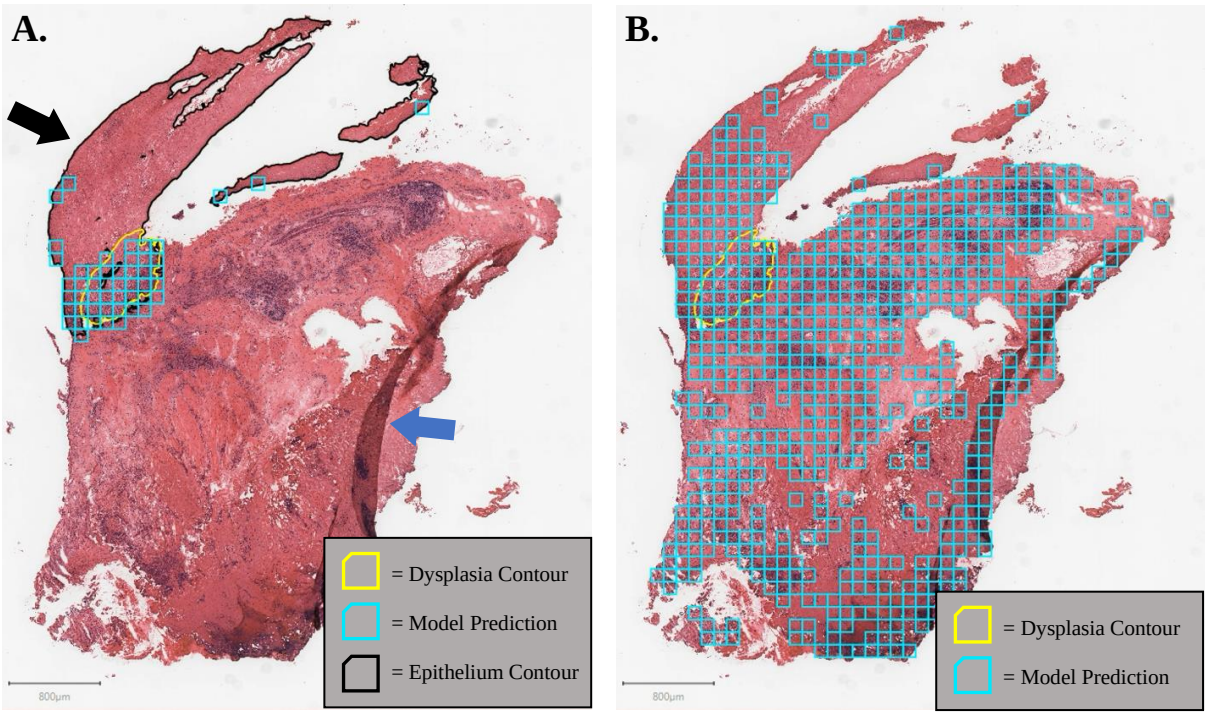

Figure 13. A sample validation HNSCC FS slide with the model predictions. The model was trained and validated on (A) epithelium-containing tiles only or (B) all tiles. Observe that the epithelium-only model correctly identified most of the top-left portion of the tissue slice (black arrow on A) as nondysplastic, while the original model incorrectly classified most of it as dysplastic. Also note the tissue folding artifact (blue arrow on A), on which the original model produced several false positive predictions.

4 DISCUSSION

4.1 Overall Performance

In this study, we investigated the use of a CNN-based tool to detect dysplasia on HNSCC resection FS slides. We found that such a tool had relatively high validation AUC, accuracy, sensitivity, and specificity. High values in these metrics are usually an indication that an ML model is performing well. However, when one notes the very poor validation PPV, it becomes clear that this is not necessarily the case. PPV is a measurement of how likely a model's positive prediction is to be truly positive. Our model's low validation PPV suggests that the tiles identified by the model as dysplastic are only rarely dysplastic in reality. Visualization of the model's binary heatmaps confirms this finding: on most validation slides, false positives produced by the model cover large areas of the tissue, effectively 'drowning out' any true positive detections.

This critical flaw means that the tool, in its current state, is not appropriate for use by a pathologist assisting in HNSCC resections. Ideally, this tool would detect potential regions of dysplasia for a pathologist to manually review. In this way, it would 'narrow down' the slide so that, instead of scanning the entire slide manually, the pathologist could analyze high-probability areas more carefully. This could allow the pathologist to make quicker and more accurate decisions about the slide's dysplasia content. In this context, the tool does not need to distinguish dysplasia from nondysplasia with perfect accuracy. However, for the tool to make a tangible difference in how quickly and accurately a pathologist can analyze an HNSCC resection FS slide, it must have two general characteristics. First, it must detect dysplasia with high sensitivity such that, if there is dysplasia present on the slide, the tool will highlight most of it for the pathologist to review.

Second, it should take less time to analyze a slide by reviewing the model's detections than it would to scan the whole slide manually. This means that the model must not produce excessive false positives. Our tool appears to satisfy the first characteristic reasonably well, evidenced by a validation sensitivity of 74%. In contrast, our tool fails in the second characteristic. Only around 1 in every 100 of the model's positive predictions are truly dysplastic, meaning it would be easier and faster for a pathologist to ignore the model and scan the whole slide manually. Therefore, our current tool is not appropriate for use by pathologists assisting in HNSCC resections.

It is crucial to note that the results we present here are on validation data, and not testing data. Thus, the model's observed validation performance is subject to some data leakage. As a result, performance metrics such as AUC and sensitivity may be overestimated, so it is important to keep this caveat in mind when interpreting our results. An entirely separate testing dataset was taken from our original data (Table 2), but it was not used in the present study. Ideally, a testing dataset should be used only once so that a realistic and leakage-free model performance assessment may be reported. In this light (and given that large volumes of HNSCC resection FS slides are relatively rare), we decided to reserve our testing dataset for future work that may achieve sufficiently strong validation performance with this model.

4.2 Possible Explanations

Halicek and his colleagues recently built a CNN-based tool (using the Inception-v4 CNN[42,43]) to automatically detect HNSCC on digitized permanent pathology slides[39]. In testing, they achieved an accuracy, sensitivity, specificity, and tile-level AUC of 85%, 85%, 85%, and 0.92, respectively. Though PPV is not reported, heatmaps show that the model tends to detect HNSCC sensitively and without excessive false positives (Fig. 14). In general, the performance of

this tool is substantially stronger than the model we propose here. Two factors may be responsible for this difference in performance. The first is that detecting HNSCC is fundamentally a simpler visual task than detecting dysplasia[16]. Second, the use of permanent, paraffin-embedded pathology slides rather than FS slides means that the dataset used by Halicek and his colleagues was free of FS-related tissue artifacts. These artifacts may compromise analysis by human pathologists[44], and they sometimes cause false positive detections by our model as well (see the blue arrow in Fig. 13a). These two methodological differences may explain our model's poor performance in comparison with Halicek and his colleagues'. Nevertheless, it is still worth exploring other contributing factors to identify opportunities to improve our model.

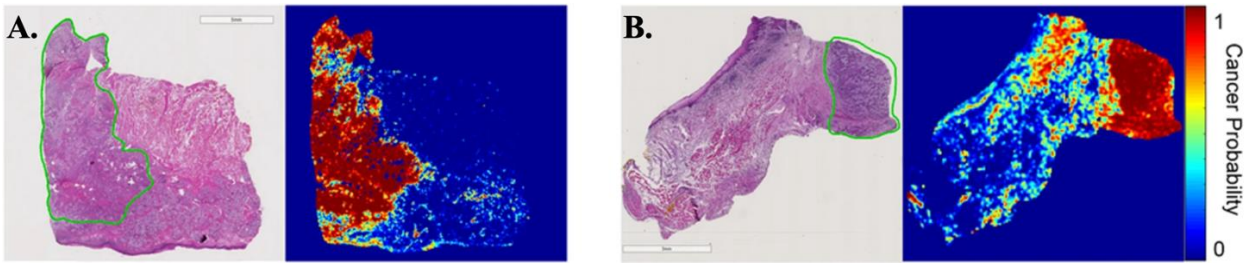

Figure 14. Sample tissue slides with HNSCC contoured in green and heatmaps corresponding to predicted probability of cancer presence by Halicek et al.'s model[39]. Observe that high predicted probability corresponds well to regions of true HNSCC. Adapted from Halicek et al. (2019)[39] under the article's Creative Commons License (http://creativecommons.org/licenses/by/4.0/).

Diagnosis of where our model is going wrong may be performed directly using activation analysis, but this technique can be resource-intensive[45]. Instead, it may be useful to take a more indirect approach by observing trends in tissue types that our model tends to perform well or poorly on.

One of the strongest trends in the model output is substantial false positive detection of normal epithelium. This is unsurprising given the known visual similarities between normal epithelium and dysplasia[16]. Interestingly though, many of the false positive, epithelium-containing tiles share some general visual hallmarks of dysplastic tissue, such as relatively dark colour (Figs. 9b-i and 9b-ii), densely packed cells (Fig. 9b-ii), and disorganized cell distribution (Fig. 9b-ii). This pattern is encouraging because it may indicate that these basic features of dysplasia are being used by the model to predict dysplasia content.

The model produced several false positives on tiles containing neutrophil infiltration (note the small dark spots in Fig. 9b-iii) of epithelium and connective tissue. This may be an indication that the model mistakes neutrophils for epithelial cell nuclei, and subsequently identifies these tiles as dysplastic because of the neutrophils' disorganized nature.

Many of the tiles on which the model produces false negatives include substantial white space, which is the background of the slide (Figs. 9c-i and 9c-iv). Ideally, the model would be able to detect dysplasia even when it contains white space. However, associating white space with nondysplasia might be another sign that the model uses colour or tone as a predictive feature.

A nondysplastic tissue type that the model almost always classifies correctly is stroma (Figs. 9d-i and 9d-ii). Stroma tends to be lighter in colour and more sparsely populated with cells, so it is encouraging that the model tends not to classify this tissue as dysplastic.

In general, it appears that the model successfully identifies some general characteristics of dysplasia, such as dark colour, dense cell population, and cell disorganization. This allows the model to distinguish dysplasia from other tissue types that are quite different from dysplasia in these regards, such as stroma. However, it seems that the model does not pick up more subtle signals that would allow it to better distinguish dysplasia from normal epithelium, for example. One such signal that is present in normal epithelium, but not dysplasia, is the gradient of enlargement and flattening of cells towards the outer edge of the tissue (partially visible in Fig. 4f).

Some methodological choices and limitations in this study may be responsible for these trends, and therefore for the poor overall performance. One potential factor is the use of a transfer learning approach. The VGG16 CNN was designed for a very general visual task[35]. Retraining only the last layer of this 16-layer CNN may have been insufficient to adapt this model for our problem. Another limitation that may have led to the poor performance of the model was the small size of the tiles we used (224 x 224 pixels). These tiles may have been too small to depict high-level spatial features – such as gradients of cell flattening – thus preventing the model from using them to detect dysplasia. Finally, the severe class imbalance in our dataset may explain the poor performance of our model. Only 1.3% of the tiles (post-screening) used were positive. This limitation might have meant that there simply were not enough positive cases provided to the model for it to adequately learn the physical features that make a tissue dysplastic. The imbalance may have also meant that too many negative cases and irrelevant tissue types (recall that dysplasia occurs primarily in the epithelium) were provided to the model, leading to substantial noise that interfered with model training.

A puzzling contradiction in our results is the model's high validation accuracy, sensitivity, and specificity despite its low PPV. It seems paradoxical that these performance metrics could simultaneously indicate strong and poor performance, but this discrepancy may be explained by the dataset imbalance. CNNs are trained by adjusting the parameters of its constituent artificial neurons to minimize a loss function. The loss function quantifies how far the predicted outcomes are from the true outcomes. Normally, this means that when a model produces too many false positives, it is penalized by the loss function, and consequently the model parameters are adjusted to make the model less sensitive. However, when there are many times more negative samples than positive ones – as is the case in our imbalanced dataset – poor performance on positive cases confers only a small penalty in the loss function. Therefore, as long as the model classifies negative cases reasonably well, it can minimize the loss function even if it produces too many false positives. In other words, the model can 'get away with' having only a vague idea of what the positive class looks like because positive cases are so rare. This effect precipitates the low PPV. Specificity and accuracy remain high in this situation because the number of false positives, even when high enough to be practically unusable, is still relatively small compared to the number of correctly classified negative cases.

4.3 Case Studies

Slides 871765 C1FS 1 (Fig. 10) and 835092 A1FS 1 (Fig. 11) are worth some discussion. While the model produces PPVs of less than 10% on every other validation slide, its PPV exceeds 75% on these slides (Fig. 6). These two exceptional performances may be explained by the composition of the slides. Most of the epithelial tissue on these slides is dysplastic, and most of the nondysplastic tissue on these slides are tissues that are very distinct from dysplasia, such as

stroma. As established above, the model tends to classify epithelial tissue (dysplastic or not) as dysplastic and stromal tissue as nondysplastic (see Section 4.2). Therefore, the model may achieve such high PPVs on slides 871765 C1FS 1 and 835092 A1FS 1 simply because (1) most epithelium is coincidentally dysplastic and (2) most nondysplastic tissue is coincidentally a tissue type that the model handles well.

4.4 Opportunities for Improvement

As described above, three factors that may be limiting the tool's performance are the transfer learning approach, the tile size, and the class imbalance. Future work may wish to address these three potentially limiting factors to improve the performance of the tool.

First, several alternatives to our transfer learning approach may be explored to improve our model in the future. By adjusting the transfer learning scheme to allow more layers of the VGG16 CNN to be trained, it is possible that the model could become more specialized to our problem, thus improving performance. The main advantage of transfer learning, however, is that it reduces computational expenses. Training more layers of our CNN would compromise some of this benefit. It may also be worth applying other pre-trained CNNs, such as AlexNet[46] or ResNet[47], to our existing transfer learning scheme for comparison. These other CNNs may be able to learn the characteristics of dysplasia better than VGG16 can. Alternatively, an entirely new CNN could be built from scratch. This approach would allow maximal customizability of the model to our problem, but it would entail substantially more expensive computations.

Future work on this tool could also experiment with the use of different tile sizes and techniques. Larger tiles sizes may improve the tool by providing a larger area of tissue to the model

for prediction. A larger area may allow the model to identify higher level spatial features, such as the gradient of cell flattening that distinguishes dysplasia from normal epithelium. A disadvantage of larger tiles is that it there are more pixels, and thus inputs, into the model. More inputs may increase the processing time of a tile by a CNN, unless the larger tile is first downsampled. A post-processing technique that may improve the PPV of the model is applying some criteria about adjacent tile predictions that must be satisfied for a positive tile prediction to be displayed. For example, one could ignore all positively predicted tiles that are not part of a contiguous, 2 x 2 square of positively predicted tiles. On slides such as the one shown in Fig. 7, this technique would filter out some of the false positive predictions, thereby improving PPV.

A third approach to improving our tool in the future is addressing the class imbalance. A possible solution might be to apply a precursor model that detects epithelial tissue and gives only tiles containing this tissue to our existing tool. This technique would screen out irrelevant tissue types (recall that dysplasia occurs primarily in epithelial tissue), thereby decreasing the class imbalance. It would also mean that our model would only need to distinguish dysplasia from normal epithelium. When alone, this might be a simpler task to handle. To explore this approach, we applied our model to the epithelium-containing tiles of a small number of slides (see Section 2.5). The model produced high validation accuracy, sensitivity, specificity, and AUC values on these tiles, and notably achieved a validation PPV of 48%. Heatmap visualization corroborates these encouraging results (Fig. 13). Though this exploratory experiment was small in scale, the strong performance of the model here suggests that applying a precursor model to detect epithelium before passing the data over to our existing model may be a viable option for improving our tool's performance. Of course, the success of this approach depends on the performance of this

hypothetical precursor model. Indeed, it is possible that detecting epithelium is just as difficult a task as detecting dysplasia.

In this work, we made some first steps towards building a tool for detecting dysplasia on HNSCC resection FS slides. To our knowledge, this tool would be the first of its kind. Currently, the tool lacks the performance needed for a clinical translation, as its PPV is simply too low to be practically usable. However, if the performance can be improved in the future with one or more of the solutions detailed above (and rigorously tested on unseen testing data), it may become suitable for use by pathologists. If successfully translated to a clinical context, the tool could be used to make faster and more accurate intraoperative decisions about whether to continue HNSCC surgery. This would improve the overall efficacy of surgical resection in achieving remission of HNSCC and reduce the time that the patient spends in the operating room.

5 CONCLUSIONS

Dysplasia must be removed during HNSCC resections to prevent disease recurrence. However, dysplasia can be difficult to detect manually by the pathologists guiding these procedures. This work aimed to assist these pathologists by building a tool to automatically detect dysplasia on HNSCC resection FS slides. To build this tool, we implemented the VGG16 CNN in a transfer learning approach using MATLAB and Python. The tool was evaluated with quantitative performance metrics and binary heatmaps integrated into the popular digital pathology tool, QuPath. A tool for this problem was successfully built, but its current performance is too poor to

be clinically useful. The poor performance may be due to limitations in dataset size or a substantial class imbalance.

Despite the tool's poor current performance, encouraging trends are present in the tissue types that the tool classifies well and poorly. These patterns suggest that our CNN-based tool extracts some predictive value from HNSCC resection FS slides, which motivates future work aiming to enhance the tool's detective power. Such future work may wish to apply different CNN models, larger tiles, or an epithelium-detecting precursor model to the tool. If the tool could be sufficiently improved, it may be used by pathologists assisting in HNSCC resection procedures to make faster and more accurate intraoperative decisions to guide the surgery. This would, in turn, improve the efficacy of surgical resection as a treatment for HNSCC.

6 REFERENCES

1. Johnson, D. E. *et al.* Head and neck squamous cell carcinoma. *Nat Rev Dis Primers* **6**, 1–22 (2020).
2. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians* **70**, 7–30 (2020).
3. Head and Neck Cancers - National Cancer Institute. https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet (2021).
4. Cognetti, D. M., Weber, R. S. & Lai, S. Y. Head and Neck Cancer: An Evolving Treatment Paradigm. *Cancer* **113**, 1911–1932 (2008).
5. Layfield, E. M., Schmidt, R. L., Esebua, M. & Layfield, L. J. Frozen Section Evaluation of Margin Status in Primary Squamous Cell Carcinomas of the Head and Neck: A Correlation Study of Frozen Section and Final Diagnoses. *Head Neck Pathol* **12**, 175–180 (2017).
6. Agaimy, A., Stelzle, F., Zenk, J. & Iro, H. Intraoperative Gefrierschnittdiagnostik von Kopf-Hals-Tumoren. *Pathologe* **33**, 389–396 (2012).
7. Olson, S. M., Hussaini, M. & Lewis, J. S. Frozen section analysis of margins for head and neck tumor resections: reduction of sampling errors with a third histologic level. *Mod Pathol* **24**, 665–670 (2011).
8. Lee, J., Lee, S. & Bae, Y. Multiple Margin Positivity of Frozen Section Is an Independent Risk Factor for Local Recurrence in Breast-Conserving Surgery. *J Breast Cancer* **15**, 420–426 (2012).
9. Smits, R. W. H. *et al.* Resection margins in oral cancer surgery: Room for improvement. *Head & Neck* **38**, E2197–E2203 (2016).

10. Jesse, R. H. & Sugarbaker, E. V. Squamous cell carcinoma of the oropharynx: Why we fail. *The American Journal of Surgery* **132**, 435–438 (1976).
11. Cheng, H. *et al.* Prolonged operative duration is associated with complications: a systematic review and meta-analysis. *Journal of Surgical Research* **229**, 134–144 (2018).
12. Helliwell, T. R. & Giles, T. E. Pathological aspects of the assessment of head and neck cancers: United Kingdom National Multidisciplinary Guidelines. *J Laryngol Otol* **130**, S59–S65 (2016).
13. Chu, F. *et al.* Laryngeal dysplasia: Oncological outcomes in a large cohort of patients treated in a tertiary comprehensive cancer centre. *American Journal of Otolaryngology* **42**, 102861 (2021).
14. Jabarin, B., Pitaro, J., Marom, T. & Muallem-Kalmovich, L. Dysplastic Changes in Patients with Recurrent Laryngeal Leukoplakia: Importance of Long-Term Follow-Up. *Isr Med Assoc J* **20**, 623–626 (2018).
15. Luers, J. C., Sircar, K., Drebber, U. & Beutner, D. The impact of laryngeal dysplasia on the development of laryngeal squamous cell carcinoma. *Eur Arch Otorhinolaryngol* **271**, 539–545 (2014).
16. Pai, S. I. & Westra, W. H. Molecular Pathology of Head and Neck Cancer: Implications for Diagnosis, Prognosis, and Treatment. *Annu Rev Pathol* **4**, 49–70 (2009).
17. Fleskens, S. & Slootweg, P. Grading systems in head and neck dysplasia: their prognostic value, weaknesses and utility. *Head & Neck Oncology* **1**, 11 (2009).
18. Baxi, V., Edwards, R., Montalto, M. & Saha, S. Digital pathology and artificial intelligence in translational medicine and clinical practice. *Mod Pathol* **35**, 23–32 (2022).

19. Jahn, S. W., Plass, M. & Moinfar, F. Digital Pathology: Advantages, Limitations and Emerging Perspectives. *J Clin Med* **9**, 3697 (2020).
20. Pantanowitz, L. *et al.* Twenty Years of Digital Pathology: An Overview of the Road Travelled, What is on the Horizon, and the Emergence of Vendor-Neutral Archives. *J Pathol Inform* **9**, 40 (2018).
21. Nichols, J. A., Herbert Chan, H. W. & Baker, M. A. B. Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophys Rev* **11**, 111–118 (2018).
22. Sneha, N. & Gangil, T. Analysis of diabetes mellitus for early prediction using optimal features selection. *J Big Data* **6**, 13 (2019).
23. Christie, J. R., Abdelrazek, M., Lang, P. & Mattonen, S. A. A multi-modality radiomics-based model for predicting recurrence in non-small cell lung cancer. in *Medical Imaging 2021: Biomedical Applications in Molecular, Structural, and Functional Imaging* vol. 11600 116000L (International Society for Optics and Photonics, 2021).
24. Wang, X. *et al.* A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head CT scans. *NeuroImage: Clinical* **32**, 102785 (2021).
25. Awais, M., Chiari, L., Ihlen, E. A. F., Helbostad, J. L. & Palmerini, L. Classical Machine Learning Versus Deep Learning for the Older Adults Free-Living Activity Classification. *Sensors (Basel)* **21**, 4669 (2021).
26. Chen, R.-C., Dewi, C., Huang, S.-W. & Caraka, R. E. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data* **7**, 52 (2020).
27. Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874 (2006).

28. Sherafatian, M. & Arjmand, F. Decision tree-based classifiers for lung cancer diagnosis and subtyping using TCGA miRNA expression data. *Oncol Lett* **18**, 2125–2131 (2019).
29. Son, Y.-J., Kim, H.-G., Kim, E.-H., Choi, S. & Lee, S.-K. Application of Support Vector Machine for Prediction of Medication Adherence in Heart Failure Patients. *Healthc Inform Res* **16**, 253–259 (2010).
30. Srinidhi, C. L., Ciga, O. & Martel, A. L. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis* **67**, 101813 (2021).
31. Kandel, I. & Castelli, M. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express* **6**, 312–315 (2020).
32. Motta, D. *et al.* Optimization of convolutional neural network hyperparameters for automatic classification of adult mosquitoes. *PLOS ONE* **15**, e0234959 (2020).
33. Lacerda, P., Barros, B., Albuquerque, C. & Conci, A. Hyperparameter Optimization for COVID-19 Pneumonia Diagnosis Based on Chest CT. *Sensors (Basel)* **21**, 2174 (2021).
34. Lee, A. L. S., To, C. C. K., Lee, A. L. H., Li, J. J. X. & Chan, R. C. K. Model architecture and tile size selection for convolutional neural network training for non-small cell lung cancer detection on whole slide images. *Informatics in Medicine Unlocked* **28**, 100850 (2022).
35. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556v6 [cs.CV]* (2015).
36. Weiss, K., Khoshgoftaar, T. M. & Wang, D. A survey of transfer learning. *Journal of Big Data* **3**, 9 (2016).
37. Yang, D. *et al.* Detection and analysis of COVID-19 in medical images using deep learning techniques. *Sci Rep* **11**, 19638 (2021).

38. Dabeer, S., Khan, M. M. & Islam, S. Cancer diagnosis in histopathological image: CNN based approach. *Informatics in Medicine Unlocked* **16**, 100231 (2019).
39. Halicek, M. *et al.* Head and Neck Cancer Detection in Digitized Whole-Slide Histology Using Convolutional Neural Networks. *Sci Rep* **9**, 14043 (2019).
40. Humphries, M. P., Maxwell, P. & Salto-Tellez, M. QuPath: The global impact of an open source digital pathology system. *Comput Struct Biotechnol J* **19**, 852–859 (2021).
41. Groovy Language Documentation. https://docs.groovy-lang.org/docs/next/html/documentation/#_introduction.
42. Szegedy, C. *et al.* Going deeper with convolutions. in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1–9 (2015). doi:10.1109/CVPR.2015.7298594.
43. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv:1602.07261v2 [cs.CV]* (2016).
44. Thomson, A. M. & Wallace, W. A. Fixation artefact in an intra-operative frozen section: a potential cause of misinterpretation. *J Cardiothorac Surg* **2**, 45 (2007).
45. Stano, M., Benesova, W. & Martak, L. S. Explaining Predictions of Deep Neural Classifier via Activation Analysis. *arXiv:2012.02248v1 [cs.AI]* (2020).
46. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1* 1097–1105 (Curran Associates Inc., 2012).
47. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016). doi:10.1109/CVPR.2016.90.

7 APPENDIX A

```
%========================================================================%
Experiment.StartNewSection('Set up parameters')
%========================================================================%

% Set up parameters as strings. Python will parse to correct type
chEpochs = '2';
chLearningRate = '0.002';
chBatchSize = '350';
chExperimentFolderName = 'Test 1';


%========================================================================%
Experiment.EndCurrentSection()
Experiment.StartNewSection('In Python')
%========================================================================%

% Convert datapaths to the format python uses
sTrainDataCSVPath = strrep(Experiment.GetDataPath('Training'),'\','\\');
sValDataCSVPath = strrep(Experiment.GetDataPath('Validation'),'\','\\');
sResultsDir = [strrep(Experiment.GetResultsDirectory,'\','\\'),'\\'];

% Collect all input arguments to be passed to python script
c1chPythonScriptArguments = {sTrainDataCSVPath, sValDataCSVPath,
sResultsDir,...
                             chEpochs, chLearningRate, chBatchSize,...
                             chExperimentFolderName};
% Run the python code
PythonUtils.ExecutePythonScriptInAnacondaEnvironment(...
    'main.py', c1chPythonScriptArguments,'C:\Users\rgilliland\miniconda3',
'keras_env');

% Load the mat file python drops its MATLAB-compatible variables in
% This has: viTruth and vsiConfidences where si means single
load([Experiment.GetResultsDirectory(),'\Workspace_in_python.mat'])

disp("Num epochs completed: " + num2str(dNumEpochs))
%========================================================================%
Experiment.EndCurrentSection()
Experiment.StartNewSection('Error metrics')
%========================================================================%
iPositiveLabel = int32(1);
vdConfidences = double(vsiConfidences);
viTruth = viTruth';

dAUC = ErrorMetricsCalculator.CalculateAUC(viTruth, vdConfidences,
iPositiveLabel);
disp("MATLAB AUC: " + num2str(dAUC,'%.2f'))

dThreshold = ErrorMetricsCalculator.CalculateOptimalThreshold(...
    {"upperleft","MCR"}, viTruth, vdConfidences, iPositiveLabel);
```

```
dAccuracy = 1 - ErrorMetricsCalculator.CalculateMisclassificationRate(...
    viTruth, vdConfidences, iPositiveLabel,dThreshold);
disp("MATLAB accuracy is: " + num2str(round(100*(dAccuracy)))+ "%")

dTrueNegativeRate = ErrorMetricsCalculator.CalculateTrueNegativeRate(...
    viTruth, vdConfidences, iPositiveLabel,dThreshold);
disp("MATLAB TNR is: " + num2str(round(100*(dTrueNegativeRate)))+ "%")

dTruePositiveRate = ErrorMetricsCalculator.CalculateTruePositiveRate(...
    viTruth, vdConfidences, iPositiveLabel,dThreshold);
disp("MATLAB TPR is: " + num2str(round(100*(dTruePositiveRate))) + "%")

dPPVTotal =
sum((int32(vdConfidences>dThreshold))&(viTruth==1))/sum(vdConfidences>dThresh
old);
disp("MATLAB PPV is: " + num2str(round(100*(dPPVTotal))) + "%")
```