Deep Learning for Computer Assisted Differential Laryngoscopic Tissue Analysis

John Faucett, jwaterfaucett@gmail.com
Jörg Lohscheller, J.Lohscheller@hochschule-trier.de

***Abstract*—This paper studies tissue classification for laryngoscopic narrow-band images using deep learning techniques, specifically, convolutional neural networks (CNN). Solving the problem of tissue differentiation in an automated way can improve medical professionals decision-making capabilities by augmenting available diagnostic information in screening and early stage diagnosis of laryngeal cancers and, thereby, improve patient outcomes. When differentiating between benign and precancerous or cancerous tissue contours a recall of 99.86% was achieved. Additionally, when classifying between benign and papilloma tissues a contour level recall of 92.31% was obtained.**

Other classifications proved much more difficult for the models to learn, especially the various differences between papillomas, dysplasias and carcinomas. When performing the four class classification task of correctly distinguishing selected tissues as either *Benign*, *Papilloma*, *Dysplasia* or *Carcinoma* the recall score was 60.89%.

This investigation shows that there remain difficulties in differentiating certain tissue classes with CNNs, specifically dysplasias, and to a lesser degree other anomalous versions of papillomas and even healthy tissues, however, the ability to differentiate between healthy and either cancerous or precancerous tissue was established and these results could be interesting to those building laryngoscopic early stage diagnosis support systems and looking for a starting point for evaluating current capabilities at this task.

I. Introduction

Laryngeal cancer is a disease of the laryngeal tract [1]. It takes the form of squamous cell carcinomas (SCC) in the majority of cases and is diagnosed by histopathological examination of tissue samples extracted with biopsy. Like most cancers, early stage diagnosis is critical to improving patient outcomes.

A. Pathology and Diagnosis

Visual analysis is used to prescreen for early-stage development of precancerous regions. The precancerous status refers to epithelial lesions which feature cellular abnormalities and structural alterations ranging from simple hyperplasia to *in situ* carcinoma also known as dysplasia. [1].

These visual analyses of tissues were - and still are - conducted using classic white light (CWL) imaging techniques. However, one solution that has emerged as a more effective means in detection and screening is narrow-band imaging (NBI), whereby special filters in a sequential red, green, and blue illumination system are used to better identify underlying structural tissues [2]. This works because the utilized illumination wavelengths of 415 nm (blue) and 540 nm (green) correspond to the peak absorption spectrum of hemoglobin, and consequently, serve to sharpen and delineate underlying capillary vessels on surface mucosa more effectively than CWL [2]. This aids in detection because altered vascular patterns can be clear signs of tumor onset [1]. See Fig. 1 for a side by side comparison of CWL and NBI images.

In summary, cancerous and precancerous tissues are structurally and visually differentiable from benign ones and NBI imaging enhances the view of underlying vascular patterns which themselves can be indicative of tumor onset.

B. Deep Learning in Medical Imaging

Since the impressive deep learning solution by Google using ImageNet burst onto the scene in 2012 [3], the effectiveness of deep learning and, specifically, convolutional neural networks (CNNs) for image classification tasks has been demonstrated in a plethora of contexts that have extended into medical image analysis [4].

Many unique applications of CNNs and deep learning techniques to medical imaging problems have been explored. Some of those relevant to this study include using deep learning for lesion detection in mammograms [4] and CNNs for detecting gastric cancers in endoscopic images [5].

One of the major benefits of deep learning is that it allows the computer to learn abstract concepts, such as the distinction between cancerous or benign tissues, from simpler ones such as edges, contours, and textural vascular patterns without the need for hand-engineering of even the simplest features by experts [6].

CNNs can therefore automatically learn the features of a dataset whereas with other algorithms these features must be

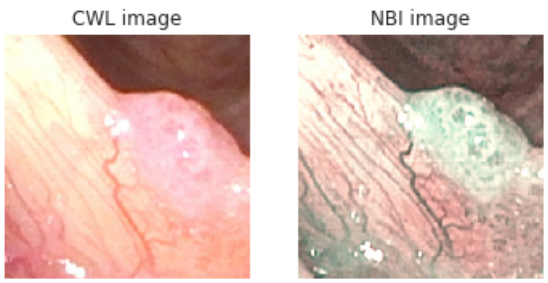

Fig. 1: Comparison between a CWL laryngoscopic image and its NBI counterpart. Note the increased clarity and sharpness of vascular structures and tissue patterns in the NBI version.

hand-engineered by experts and are subject to the arduous and error-prone process of manually discovering all the factors of variation [6]. Additionally, it has been demonstrated that CNNs can outperform humans on a variety of image classification tasks, suggesting that they may be able to discover features even the experts cannot detect as they consequently outperform them [7].

Succinctly, laryngoscopic precancerous structures have a multitude of distinguishing characteristics which, when augmented by NBI imaging, provide fertile ground for the application of CNNs to aid professionals in early stage diagnosis and thus improve patient outcomes.

C. Research Goals

This research provides an initial exploration of the application of deep learning techniques, specifically convolutional neural networks, in the domain of computer aided laryngoscopic diagnosis with a focus on tissue differentiation and classification. To the best of my knowledge, this is the first attempt at applying CNNs to a laryngoscopic dataset and the goal is to provide useful information for other researchers to build upon going forward.

In this paper, mainly two questions are explored.

1) What differentiations between tissue samples are difficult for a CNN to learn and what causes these difficulties?
2) What are the general parameters and overall design of an effective CNN for laryngoscopic tissue classification?

The intention is to answer these questions in as much relevant detail as possible. Finally, the research is concluded by assessing the experimental results against an independently developed laryngoscopic tissue dataset created by Moccia, De Momi, and Mattos [8].

II. Methods

A. The Dataset

The dataset originated from different medical institutions and professionals. It consisted of $n = 83$ NBI images, most of which contained contour regions of interest (ROIs) which had been hand selected by medical professionals. Many of these images were too low in resolution, duplicates, too poor in quality, zoomed images made with a flexible endoscope and thus too atypical, or contained labeled image regions which were too small to fit within the selected patch size of a 64x64 square image. After removing duplicates and instances of extremely low quality, a dataset was derived which consisted of 39 NBI images and 97 classified ROIs.

Constituting some of those 97 ROIs were benign ones which were hand selected by me in order to better balance the dataset. A preference was given towards selecting regions that were blatantly benign in visual appearance, nonetheless, this could be a source of error in the results since I am not a medical expert in this domain.

From this dataset, however, was further derived a prime quality dataset, which consisted only of the highest quality images (23 images, 65 ROIs). These images and contours contained little noise and provided a high degree of confidence in their quality as input data to the models. The aforementioned and initially cleaned up dataset of 39 NBI images was called *subprime*. Both of these datasets were used in the experiments. For clarity, the dataset used for each result is shown as either *prime* (P) or *subprime* (SP).

B. Definition of Gold Standard

Medical professionals defined the ROIs within the NBI images and then, based on the histological biopsy results, classified these regions into four categories: *Benign* - healthy tissues, *Papilloma* - benign epithelial tumors, *Dysplasia* - precancerous laryngeal lesions and *Carcinoma* - cancerous lesions.

In Fig. 2, one can see typical examples of each class contained within a medical professional's ROI from which a tissue sample was taken. Images had one or more ROIs, each classified into one of these four categories.

The dataset was small and relatively unbalanced as shown in Fig. 3, which is not atypical for many deep learning tasks in medical applications [9].

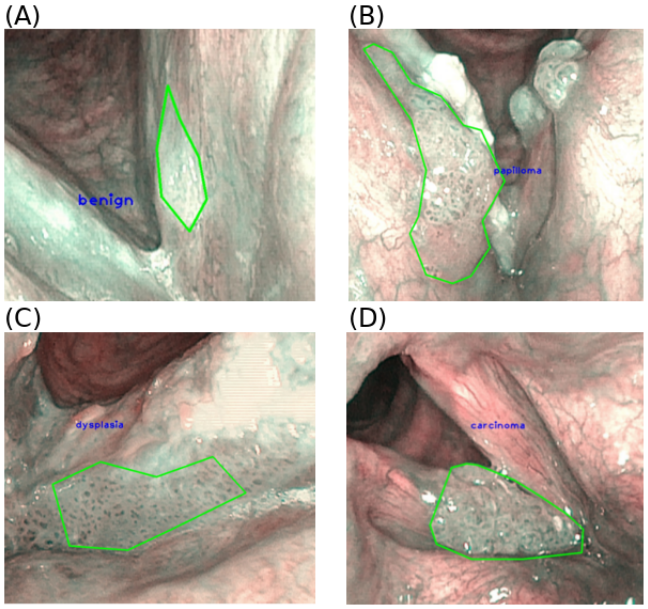

Fig. 2: Typical examples of each class contained within a medical professional's ROI. (A) *Benign*, (B) *Papilloma*, (C) *Dysplasia*, (D) *Carcinoma*.

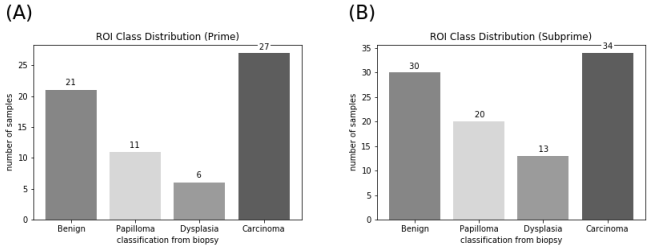

Fig. 3: The tissue ROI class distribution across the datasets. (A) Prime Dataset (B) Subprime Dataset.

C. The Data Processing Pipeline

The datasets were constructed from patches extracted from classified contours which had been split into train/test/validation sets. These patches were then used as input values for training the models. Randomness entered mainly due to our train/test/split function, which guaranteed that the test and validation ROIs consisted of samples from all classes under inspection but still allowed for those ROIs to be otherwise randomly chosen. This choice of which ROIs to use as a basis for patch generation in the test and validation sets could not be left up to pure randomness, since that could have resulted in a poor and heavily biased sample for testing and validating the models. The total patches in each of the train, test, and validation sets also varied depending on dataset and experiment and the fact that the chosen patch size of 64x64x3 meant that larger ROIs provided more patches than did smaller ones.

After the train/test/validation ROIs were selected the patch generation process began. This started by generating a bounding box around the polygonal ROI, generating patches inside the bounding box, and then keeping all patches whose area intersected by at least 50% with the polygonal ROI. Fig. 4 demonstrates this part of the patch extraction process.

After some initial testing, an additional problem which surfaced was that of poor and noisy patch quality. In order to address one aspect of this, patches which originated from extremely specular surfaces, were thus almost pure white and consequently provided little if no beneficial information to the model, were automatically detected and removed.

D. Experiment Design and Structure

Six experiments were conducted using this dataset. Each experiment investigated a slightly different tissue classification question. A seventh baseline experiment was conducted using an independently developed and recently published laryngeal tissue dataset [8]. The first six experiments were conducted using the same structure, which was as follows.

1) Split the training, test and validation ROIs from either the *prime* or *subprime* datasets.
2) Generate patches from each of the sets of ROIs.

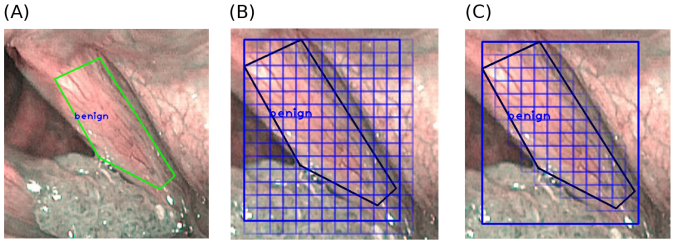

Fig. 4: Illustration of patch extraction process. (A) Original ROI with *Benign* classification (B) Generated bounding box and containing patch set (C) Filtered patch set after removing patches whose area was less than 50% inside the ROI. Notice that some of the *Benign* patches on the bottom border overlap with and hence contain *Carcinoma* tissue. This illustrates a general source of error in model training.

Name	Class Type	Classes
E1	Binary	Benign v. Papilloma
E2	Binary	Benign v. (Dysplasia+Carcinoma)
E3	Binary	Papilloma v. (Dysplasia+Carcinoma)
E4	Binary	Dysplasia v. Carcinoma
E5	3-Class	Benign, Papilloma, (Dysplasia+Carcinoma)
E6	4-Class	Benign, Papilloma, Dysplasia, Carcinoma
E7	4-Class	He, Hbv, Leukoplakia, IPCL

TABLE I: Experiments organized by name, problem type and classes examined. The class *(Dysplasia+Carcinoma)* was a merger of *Dysplasia* and *Carcinoma* classes which denotes all cancerous or precancerous tissues. E7 uses the laryngeal dataset which utilizes different tissue classifications [8].

3) Randomize the patches and corresponding labels so each training batch sees enough different classes.
4) Apply mean subtraction normalization.
5) Start with a CNN containing 2 hidden layers and no regularization to find parameters such as batch size and learning rate.
6) Incrementally add complexity and features such as more hidden layers or increased regularization such as dropout.
7) Select the best performing model.
8) Evaluate the model predictions at both the ROI and extracted patch level.

E. The Experiment Description

The experiments are listed in Table I. For experiments E2, E3 and E5 the *Dysplasia* and *Carcinoma* classes were merged. This makes sense because they are very similar in visual appearance and dysplasias can be indicators of precancerous tissue. This, combined with experiment E4 which explicitly inspected the relationship between dysplasias and carcinomas, served to help evaluate the difficulty of differentiating these two similar tissue classes.

Another focus which can be seen in these experiments is evaluating the difficulty of differentiating tissue - either benign or carcinoma/dysplasia - from papillomas. This was the purpose of experiments E1 and E3.

E6 was an attempt at classifying all the labels in the dataset, while E1-E5 allowed for an exploration of relationships between the other tissue types in more fine-grained detail.

Binary classification problems (E1-E4) used binary cross-entropy as their loss function and the sigmoid function $\sigma(x) = {}^{1}/_{1+e^{-x}}$ as the model's final activation function, whereas E5-E7 used categorical cross entropy as a loss function and a softmax output layer activation. All tested models used standard ReLUs $f(x) = max(0, x)$ for all other activation functions between layers.

In terms of validating the results, K-fold validation was not conducted due to limited time and computing resources. Instead after training and testing the models, they were run once against the validation set. The results which are reported in this paper are the scores of those final runs against the validation datasets.

For the evaluation scores of precision and recall at a patch level I used scikit-learn's corresponding functions with macro

Model	Depth	Parameters
Model 1 (M1)	4-Conv Layers, 1-Dense, 1-Output	8,455,714
Model 2 (M2)	6-Conv Layers, 1-Dense, 1-Output	17,659,896
Model 3 (M3)	5-Conv Layers, 2-Dense, 1-Output	56,366,786

TABLE II: Experiment models. Parameters given are for binary classification models only.

Metric	Score
Patch Precision (specificity)	90.41%
Patch Recall (sensitivity)	92.54%
Patch Accuracy	92.04%
ROI Precision (specificity)	86.16%
ROI Recall (sensitivity)	92.31%
ROI Accuracy	88.74%

TABLE III: Experiment E1: Benign v. Papilloma. Patch and ROI scores (model M2, dataset SP).

averaging. This calculates the metrics for each label but ignores imbalance in the label distribution. It was decided to not weight the results since the data was already imbalanced and weighting would have artificially inflated the scores on these metrics. For ROI level scoring these statistics were computed with self-implemented python functions since scikit-learn does not currently provide functions for measuring multiclass continuous outputs such as those which result from the softmax function.

Experiment E7 used the laryngeal dataset which is a different dataset consisting of patches created from NBI laryngeal images. These had been put into four classes with a total of 330 patches per class. The classes were labeled as *He* (healthy tissue), *Hbv* (tissue with hypertrophic vessels), *Le* (tissue with leukoplakia) and *IPCL* (tissue with intrapapillary capillary loops). This dataset was used to evaluate the best performing models against a similar set of images.

Variations on the three models in Table II were evaluated. Model 1 (M1) was a simple two hidden layer model which was used to start probing in each experiment. Model 2 (M2) was a slightly more advanced variant on M1 which added another hidden layer and more filters per layer. Model 3 (M3) had been suggested from previous studies and had many more parameters than the other two.

III. Results

A. Experiment E1: Benign v. Papilloma

This was the first experiment which was run and helped to learn general parameters which worked well in the remainder of the experiments. It was discovered that the models learned best with low batch sizes between $16 - 64$ and by using the Adam Optimization Algorithm which yielded better results than RMSprop. Also, keeping the learning rate between 0.001 and 0.0001 and cutting it in half once the loss had plateaued all proved useful, not only for this but for the rest of the tests as well.

The dataset for the results in this experiment consisted of 50 ROIs, 38 for training and 6 each for testing and validation. The patch distribution for each set is shown in Table IV.

Dataset	T	B	P	B%	P%
Train	1555	1088	467	69.97%	30.03%
Test	177	95	82	53.67%	46.33%
Validation	201	134	67	66.67%	33.33%

TABLE IV: Experiment E1: Benign v. Papilloma. Patch distribution across datasets. T = *Total*, B = *Benign*, P = *Papilloma*

In terms of tissue differentiation at both a patch and ROI level, the results are illustrated in Table III and the corresponding confusion matrix for the patches is shown in Fig. 5. Many of the falsely classified patches were those around tissue boundaries, which was a problem that repeated itself in the remainder of the experiments. At the boundaries of any selection, for instance an ROI labeled *Papilloma*, there are contour edges where the generated patches were largely on top of what visually appeared to be benign tissue. This made up a significant portion of the inspected errors but by no means all of them. Earlier models also had trouble differentiating between greenish benign tissue and greenish papillomas. Overall, the CNN had the fewest problems identifying papillomas when those had a clear outgrowth geometrical shape. Benign tissue was most correctly identified when it was not in a green spectrum meaning when there was little underlying vascular structure.

The six ROIs in the validation dataset were all correctly classified when selecting the output class with the highest probability score. The most error filled prediction had $P(benign) = 0.62$ and $P(papilloma) = 0.38$. However, since false negatives are more important in assessing the utility of the classifications an example which exhibits a false negative patch is shown in Fig. 6.

There was no data augmentation performed in this experiment, but the model did experience a large improvement in the performance when moving from the prime dataset, which contained 32 ROIs after selecting only *Benign* and *Papilloma* classes, to the subprime one with 50 ROIs. In this case, the

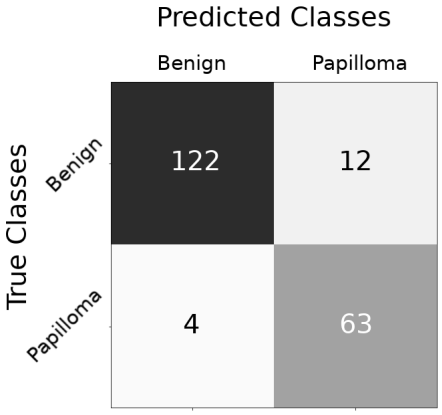

Fig. 5: Classification performance of experiment E1. Patch level confusion matrix on 201 validation patches (model M2, dataset SP).

Metric	Score
Patch Precision (specificity)	99.37%
Patch Recall (sensitivity)	99.17%
Patch Accuracy	99.28%
ROI Precision (specificity)	96.15%
ROI Recall (sensitivity)	99.86%
ROI Accuracy	97.93%

TABLE V: Experiment E2: Benign v. Dysplasia+Carcinoma. Patch and ROI scores (model M3, dataset SP).

Dataset	T	B	C	B%	C%
Train	3844	1047	2797	27.24%	72.76%
Test	366	149	217	40.71%	59.29%
Validation	278	121	157	43.53%	56.47%

TABLE VI: Experiment E2: Benign v. Dysplasia+Carcinoma. Patch distribution across datasets. T = *Total*, B = *Benign*, C = *Dysplasia+Carcinoma*

Metric	Score
Patch Precision (specificity)	73.83%
Patch Recall (sensitivity)	74.32%
Patch Accuracy	83.90%
ROI Precision (specificity)	58.80%
ROI Recall (sensitivity)	91.53%
ROI Accuracy	63.70%

TABLE VII: Experiment E3: Papilloma v. Dysplasia+Carcinoma. Patch and ROI scores (model M3, dataset SP).

Dataset	T	P	C	P%	C%
Train	3143	409	2735	13.01%	86.99%
Test	256	134	122	52.34%	47.66%
Validation	388	73	315	18.81%	81.19%

TABLE VIII: Experiment E3: Papilloma v. Dysplasia+Carcinoma. Patch distribution across datasets. T = *Total*, P = *Papilloma*, C = *Dysplasia+Carcinoma*

best performing model on the prime dataset achieved a patch level accuracy of 67% and a recall of 66%. This demonstrates that even moderate increases in the size of the dataset can yield significantly better results.

B. Experiment E2: Benign v. Dysplasia+Carcinoma

The dataset used in this experiment consisted of 77 ROIs, 65 for training and 6 each for testing and validation, the patch distribution for each set is shown in Table VI.

The M3 model obtained a recall score of 99.86% for ROIs and a patch recall score of 99.17%. A full breakdown of the statistical measures can be found in Table V. These were obtained using data augmentation which substantially increased the ability of the models to correctly classify the *Dysplasia+Carcinoma* patches. There was still some misclassification around the edges of ROIs which was observed in many other experiments and contributed to the lower patch recall score. At the false positive end, a typical mistake made by the model was a misclassification originating from a ground truth patch or ROI which was benign and green in texture. An example can be found in Fig. 7, and this illustrated a problem seen throughout the experiments, namely, that when benign patch tissues or contour regions were misclassified they were often smooth and green or demonstrative of vascular underlying tissue that most often had a greenish texture.

This could have been a problem with the size of the dataset, that is, there simply were not enough benign greenish samples for the models to learn to differentiate such tissues from other benign ones, or, and this warrants discussion with medical professionals, there could be characteristic structures of such tissues which make them uniquely difficult to differentiate. Beyond this error zone, however, there was not a significant problem distinguishing between benign and cancerous or pre-cancerous tissues.

C. Experiment E3: Papilloma v. Dysplasia+Carcinoma

This experiment proved very difficult on many levels. Due to the low number of *Papilloma* classified ROIs, the dataset was highly unbalanced. It consisted of 67 ROIs, 55 for training and 6 each for testing and validation. This overall imbalance

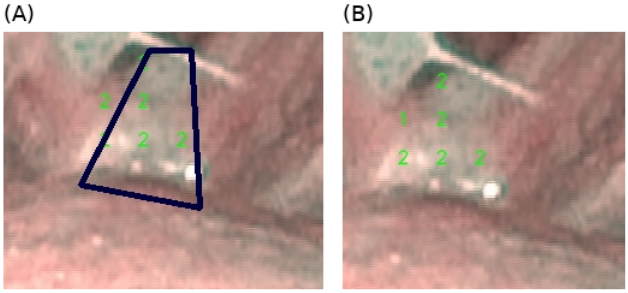

Fig. 6: Experiment E1: Benign v. Papilloma. Example of false negative boundary region patch classification. (A) Ground truth labels (B) Model M2 predictions. *Benign* and *Papilloma* patches are marked as 1 and 2 respectively. Notice the false negative on the left outer edge of the contour. To the naked eye it appears as if the model is correctly classifying the tissue though one does not know for sure.

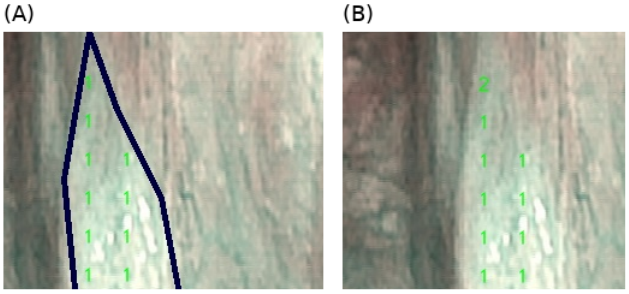

Fig. 7: Experiment E2: Benign v. Dysplasia+Carcinoma. Misclassified greenish benign tissue at tip of ROI. (A) Ground truth (B) Model M3 predictions. *Benign* and *Dysplasia+Carcinoma* are marked by labels 1 and 2. The misclassified tissue patch at the top of the contour demonstrates an occurrence of false positives and is potentially informative for building higher precision models.

can be seen in Table VIII where total patches per class are shown.

The best obtained results are shown in Table VII, however, due to the imbalance in the dataset the results are statistically insignificant. Still, there is some useful information one can glean from an examination of the experimental outcomes.

First, when choosing the class with the highest probability score, four out of the six ROIs were correctly classified. The only contour level misclassifications were in a safe direction. That is to say, these were tissues labeled as *Papilloma* which were classified as *Dysplasia+Carcinoma*. Also when inspecting the contour misclassification regions, one of them was in a tissue area which also contained dysplasias. This image is shown in Fig. 8.

This indicates that there is some visual and structural similarity between papillomas and dysplasias which the model did learn, perhaps more samples would enable the model to learn the specific differences underlying this degree of surface similarity.

D. Experiment E4: Dysplasia v. Carcinoma

Many attempts were made to build models that could differentiate between dysplasias and carcinomas across all the datasets. Variations on M1 through M3 were tried, some with regularization via dropout and some without, some with data augmentation and some without. Whatever the combination no significant progress could be made and there are, consequently, no results worth reporting since none of the models could learn to differentiate meaningfully between these two classes.

There could be several causes for this which should be mentioned because they will need to be addressed in future research.

1) The datasets were far too biased for carcinomas. Dysplasias made up only 9.2% and 13.4% of total classified ROIs (see: Fig. 3).
2) Of those labeled dysplasias, several were annotated as severe. It is possible there is little to no visual distinction between a severe dysplasia and carcinomas.

To test the legitimacy of the first item one need only get more examples of dysplasias, ideally of both the severe and normal varieties. Noteworthy as well, is that it could also be the case that given further samples the models could have learned to distinguish between a generic form of dysplasia and carcinomas. After these experiments, the open question remains as to whether visually or structurally distinguishable forms of dysplasias exist which can be learned by CNN models.

As for the second item, it can be noted that the goal of computer aided diagnosis is to assist professionals in the early stage diagnosis process in order to improve patient outcomes. Since a significant number of lesions that are dysplasias progress to invasive carcinoma [10], it makes sense to discuss in detail this differentiation process with medical practitioners. Here the goal would not only be to delineate differences between severe and normal dysplasias (as well as perhaps other subclassifications) and carcinoma itself, but also to evaluate the benefits, particularly as early stage diagnostic metrics - of distinguishing these tissues in the first place.

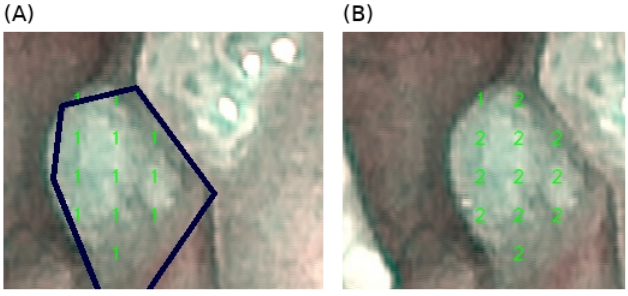

Fig. 8: Experiment E3: Papilloma v. Dysplasia+Carcinoma. Misclassified *Papilloma* in an image which contained *Dysplasia+Carcinoma* tissues. (A) Ground truth (B) Model predictions. *Papilloma* and *Dysplasia+Carcinoma* are marked by labels 1 and 2 (model M3, dataset SP).

Dataset	T	B	P	C	B%	P%	C%
Train	4217	1027	404	2786	24.35%	9.58%	66.07%
Test	352	103	61	188	29.26%	17.33%	53.41%
Validation	535	187	151	197	34.95%	28.22%	36.82%

TABLE IX: Experiment E5: Benign, Papilloma, Dysplasia+Carcinoma. Patch distribution across dataset. T = *Total*, B = *Benign*, P = *Papilloma*, C = *Dysplasia+Carcinoma*

Metric	Score
Patch Precision (specificity)	81.37%
Patch Recall (sensitivity)	64.51%
Patch Accuracy	66.36%
ROI Precision (specificity)	50.43%
ROI Recall (sensitivity)	73.12%
ROI Accuracy	50.43%

TABLE X: Experiment E5: Benign, Papilloma, Dysplasia+Carcinoma. Patch and ROI scores (model M2, dataset SP).

E. Experiment E5: Benign, Papilloma, Dysplasia+Carcinoma

Based on the previous experiments better results for this experiment were predicted than those which were actually realized. Not all possible tactics which exist for optimizing learning models were exhausted, but a large subspace of standard techniques was tried. This consisted of regularization both in the form of batch normalization and dropout as well as different combinations of data augmentation methods.

A breakdown for the dataset used for the outcomes can be found in Table IX. Those patches were derived from the subprime dataset and consisted of 79 ROIs. The training set had 9 ROIs with 3 from each class. It was the same for the validation set. One can see that imbalance was augmented to an extent by the merger of dysplasias and carcinomas under the single label *Dysplasia+Carcinoma*.

The resulting model prediction scores are shown in table X and the confusion matrix for the patches in Fig. 9. When choosing the ROI with the highest probability score, 6 out of 9 ROIs were correctly classified in the validation set, as well as all carcinoma ROIs. Interestingly, there were almost two

complete misfires as pertains to the ROI classifications. These are shown in Fig. 10 since they are illustrative of problems that have been discussed until now.

There was a lot more confusion in the model classifications than expected given previous experiments. Based off of the binary classification experiments, a better ability to differentiate between *Benign* v. *Papilloma* and *Papilloma* v. *Carcinoma* was predicted than what was actually realized.

Still, although patches and entire ROIs were misclassified all cancerous or precancerous tissues were correctly recognized and there were no false negatives at the ROI level. This reconfirms the outcomes of experiment E2 which had recall scores at patch and ROI levels of 99.17% and 99.86% respectively.

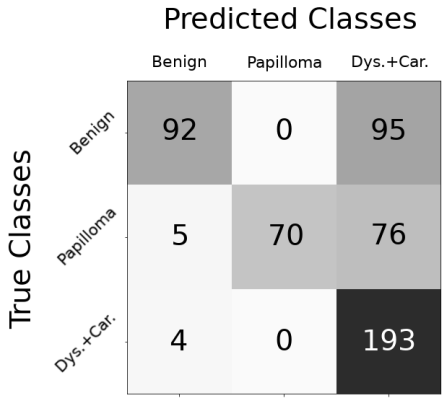

Fig. 9: Experiment E5: Benign, Papilloma, Dysplasia+Carcinoma. Patch level confusion matrix on 535 validation patches (model M2, dataset SP). Notice that many papillomas and healthy tissues are often misclassified as carcinomas.

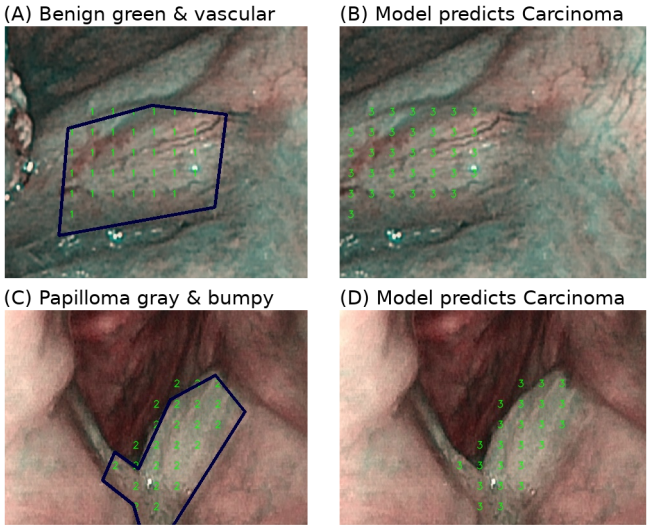

Fig. 10: Experiment E5: Benign, Papilloma, Dysplasia+Carcinoma. ROI level false positives. (A)(C) Ground truths (B)(D) Model predictions. *Benign* = 1, *Papilloma* = 2, *Dysplasia+Carcinoma* = 3. (Top) A greenish benign region misclassified completely as a *Dysplasia+Carcinoma*. (Bottom) A *Papilloma* misclassified as *Dysplasia+Carcinoma*.

Metric	Score
Precision (specificity)	94.96%
Recall (sensitivity)	95.01%
Accuracy	95.46%

TABLE XI: Cross Verification E7: Tissue classification scores for the laryngeal dataset. These are patch scores since that is all the laryngeal dataset offers.

F. Experiment E6: Benign, Papilloma, Dysplasia, Carcinoma

Finally, the four class classification experiment was run with, as expected, similar results to E5. The extraction of dysplasias into their own class decreased the model scores. The obtained patch-level recall dropped to 61.41% and the ROI level recall dropped down to 60.89%.

There were 8 total ROIs with 2 examples from each class for both the test and validation datasets. When selecting the highest scored prediction, the model M2 succeeded in classifying both *Carcinoma* ROIs as well as both *Benign* ROIs. This was to be expected from the previous experiments, since differentiating purely benign and carcinoma tissue, though still not perfect, provided much less difficulty for the models compared to differentiating between other tissue types.

The specifics of every metric are left out since no new information surfaced that added to the understanding of the problem in a meaningful way. The only new piece of information that potentially sheds some light on the problem is shown in Fig. 11. There one can clearly see that in the middle of an otherwise benign tissue region, the model M2 suddenly predicts that a few patches are dysplasias. If not an outright error in the model, there could be some underlying similarity between red and to some extent vascular benign tissues and potentially precancerous regions i.e. dysplasias. Until this point, misclassifications for benign tissues had been observed being concentrated in greenish regions as discussed in, for example, experiment E2.

G. Cross Verification E7: He, Hbv, Leukoplakia, IPCL

In order to check whether the models M2 and M3 were in a general sense effective at classifying laryngeal type

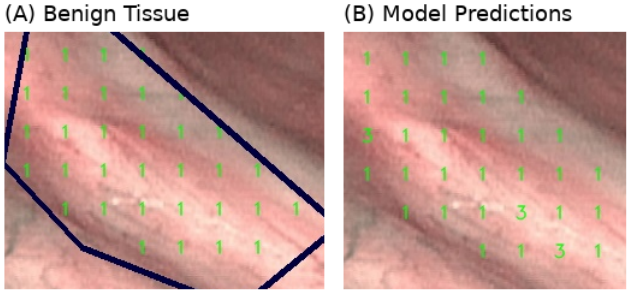

Fig. 11: Experiment E6: Benign, Papilloma, Dysplasia, Carcinoma. Misclassified patches possibly hinting at structural similarity between red benign tissues and some types of dysplasias. (A) ground truth (B) predictions. *Benign* and *Dysplasia* are marked by labels 1 and 3 (model M2, dataset SP).

tissues, they were evaluated on the laryngoscopic tissues dataset created by Moccia, De Momi, and Mattos [8]. The resulting scores of this evaluation are listed in Table XI and were obtained using data augmentation and 30 epochs. The corresponding confusion matrix is shown in Fig. 12. The recall score reported in their paper was a median of 93% which came from the best performing feature and 98% when excluding low-confidence pathces [11]. Thus this quick verification experiment demonstrated that CNNs can reach similar performance capabilities to hand-crafted features for general purpose laryngoscopic tissue classifications. It also shows that the developed models, particularly M2 and M3, are well suited starting structures for approaching this problem domain.

H. Model Assessment Summary

Overall, the best performing models were M2 and M3 and there was little significant difference between them. Most of the differences observed were likely due to randomness since there was often less than a few percentage points difference between the best performance of M2 and the best performance of M3. There was a noticed boost in performance by going from two to at least three hidden layers and also by increasing filters in those layers. This made either M2 or M3 clearly better choices than M1 for approaching any of the classification tasks. There was no significant difference when comparing one to two dense layers before the final softmax output function. In almost all cases adding data augmentation and regularization via dropout improved overall scores and when employing data augmentation, at a minimum scores remained constant and training time increased. Using the Adam Optimization Algorithm and halving the learning rate when plateauing all improved training performance and results. Keeping the batch sizes small and between 16-64 also improved the ability of the models to learn as well as training time. No models were trained for over 30 epochs.

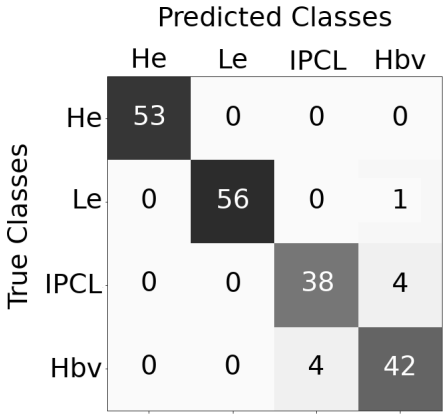

Fig. 12: Cross Verification E7: Confusion matrix computed from classifications of laryngeal tissue patches from the laryngeal dataset. (He) Healthy (Le) Leukoplakia (IPCL) intrapapillary capillary loops (Hpv) hypertrophic vessels. Uses a version of M2 adapted for the larger input patch size of the new dataset.

IV. Conclusion

This research establishes that convolutional neural networks can achieve state-of-the-art results in tissue classification of laryngoscopic NBI images when differentiating between benign and cancerous or precancerous tissues, though even here and particularly for further tissue classes there still remains much work to be done.

The achieved patch and ROI recall scores of 99.17% and 99.86% demonstrate that relatively simple CNNs can differentiate between healthy tissue and either cancerous or precancerous tissues. The experiments also showed with patch and ROI recall scores of of 92.54% and 92.31% respectively that a binary classifier can also distinguish between benign tissue and papillomas.

Other experiments achieved lower scores for classifications between more specified breakdowns of the tissue classes and the causes for this were given in the corresponding results sections. An open question remains as to what tissue classifications and differentiations are most vital for medical experts in deciding for biopsy during screening. Going forward, clarifying this ambiguity will be extremely important for developing datasets and building robust classification systems that aid experts in early stage diagnosis.

Future research should resolve this by working closely with medical experts in order to establish the visual, structural and early stage outcome differences between various stages of dysplasias as well as papillomas, though the latter likely to a much lesser degree. This will be important for developing correctly labeled datasets that are tailored for recognizing relevant early stage precancerous tissues. In order to evaluate the viability of developed models in practice, one would also need to know the diagnostic metrics of professionals during laryngoscopy.

Another extensive problem, and one which was only passingly mentioned in the paper was that of finding quality images, especially due to the specular and unstable nature of the NBI images produced during laryngoscopy. Finding quality images is an error prone, manual and time consuming process. Future research should focus on automating this stage of quality image detection.

In terms of neural network topology, there is still much ground that could be explored such as transfer learning and assessing deeper and higher parameter models than those which were used in this paper's experiments. However, this research shows that relatively simple models with at least three hidden layers, dropout regularization and either one or two fully connected dense layers can achieve quality results, especially when paired with data augmentation.

Finally, I think verifying these experiments on a more balanced laryngoscopic tissues dataset and evaluating methods for dealing with unbalanced and noisy data in laryngoscopy would both be beneficial research directions to take.

Acknowledgement

I would like to thank Prof. Dr. Jörg Lohscheller for taking the time and having the patience to give his expert advice and opinion without which I would not have been able to achieve any of these results.

References

[1] P. Schultz, "Vocal fold cancer," *European Annals of Otorhinolaryngology, Head and Neck Diseases*, vol. 128, no. 6, pp. 301–308, 2011.

[2] T. Hayashi, M. Muto, R. Hayashi, K. Minashi, T. Yano, S. Kishimoto, and S. Ebihara, "Usefulness of Narrow-Band Imaging for Detecting the Primary Tumor Site in Patients with Primary Unknown Cervical Lymph Node Metastasis," *Japanese Journal of Clinical Oncology*, vol. 40, no. 6, pp. 537–541, 2010.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[4] S. K. Zhou, H. Greenspan, and D. Shen, Eds., *Deep Learning for Medical Image Analysis*. Elsevier/Academic Press, 2017.

[5] T. Hirasawa, K. Aoyama, T. Tanimoto, S. Ishihara, S. Shichijo, T. Ozawa, T. Ohnishi, M. Fujishiro, K. Matsuo, J. Fujisaki, and T. Tada, "Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images," *Gastric Cancer*, pp. 1–8, 2018. DOI: 10.1007/s10120-018-0793-2. [Online]. Available: https://doi.org/10.1007/s10120-018-0793-2.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016, ISBN: 0262035618, 9780262035613.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15, Washington, DC, USA: IEEE Computer Society, 2015, pp. 1026–1034, ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.123. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.123.

[8] S. Moccia, E. D. Momi, and L. S. Mattos, *Laryngeal dataset*, Oct. 2017. DOI: 10.5281/zenodo.1003200. [Online]. Available: https://doi.org/10.5281/zenodo.1003200.

[9] T. Shaikhina and N. A. Khovanova, "Handling limited datasets with neural networks in medical applications: A small-data approach," *Artificial Intelligence in Medicine*, vol. 75, pp. 51–63, 2017, ISSN: 0933-3657. DOI: 10.1016/j.artmed.2016.12.003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365716301749.

[10] M. Sadri, J. McMahon, and A. Parker, "Management of laryngeal dysplasia: A review," *European Archives of Oto-Rhino-Laryngology and Head & Neck*, vol. 263, no. 9, pp. 843–852, 2006, ISSN: 1434-4726. DOI: 10.1007/s00405-006-0078-y. [Online]. Available: https://doi.org/10.1007/s00405-006-0078-y.

[11] S. Moccia, E. De Momi, M. Guarnaschelli, M. Savazzi, A. Laborai, L. Guastini, G. Peretti, and L. Mattos, "Confident texture-based laryngeal tissue classification for early stage diagnosis support," *Journal of Medical Imaging*, vol. 4, pp. 4 –4 –10, 2017. DOI: 10.1117/1.JMI.4.3.034502. [Online]. Available: https://doi.org/10.1117/1.JMI.4.3.034502.