Stain Deconvolution for Deep Learning Based Multiclass Diagnosis of Oral Cancer in Liquid-Based Cytology

Sungju Park
New York University College of Dentistry
sp8262@nyu.edu

ABSTRACT

Background: Oral cancer remains a major public health challenge, with prognosis and survival rates highly dependent on early detection and accurate diagnosis. While conventional diagnosis relies on invasive tissue biopsies and histological examination, these approaches can be resource-intensive, time-consuming, and impractical for routine screening or continuous monitoring of at-risk populations. Liquid-based cytology (LBC) combined with Papanicolaou staining provides a non-invasive, cost-effective alternative that enables large-scale screening of oral mucosal cells. However, accurate interpretation of cytological slides still depends heavily on expert pathologists, and the absence of detailed cell-level annotations limits the development and evaluation of reliable AI-based diagnostic tools. This study aims to improve AI-assisted oral cancer detection by introducing additional diagnostic information through stain deconvolution and systematically comparing the performance of deep learning models trained on original RGB images and stain-separated channels. By leveraging targeted stain information, the proposed approach seeks to enhance classification accuracy and interpretability, supporting more robust and scalable computer-assisted screening for early oral cancer detection.

Methods: We analyze H&E-stained liquid-based cytology (LBC) slides using a convolutional neural network trained on both original RGB images and stain-separated channels (Haematoxylin, Eosin, and DAB) extracted via color deconvolution. The dataset comprises 962 digitized LBC images spanning four diagnostic classes (NILM, LSIL, HSIL, SCC). Images are preprocessed with resizing, normalization, and augmentation to enhance model generalizability. Given the absence of detailed cytological annotations, we adopt a fully supervised deep learning approach using only image-level diagnostic labels. Model performance is evaluated using a stratified 4-fold cross-validation, with metrics including F1 score, accuracy, recall, precision, and ROC AUC. Statistical comparisons between models are conducted using the Friedman test with Conover's post hoc analysis (Holm correction) to assess the impact of stain separation. Grad-CAM visualizations are generated to interpret the discriminative regions contributing to classification decisions.

Results: Our experiments demonstrate that: (i) the DAB stain-deconvoluted channel consistently yields the highest diagnostic performance among all tested configurations, achieving an F1 score of 95.8% and an accuracy of 98.2%, with improved true positive detection and fewer false negatives for clinically significant classes; (ii) the Eosin channel provides significantly different but weaker diagnostic information compared to RGB and other channels ($p < 0.05$, Conover's post hoc test), underscoring its limited contribution when used in isolation. Thus, stain deconvolution offers valuable complementary diagnostic information beyond conventional RGB imaging, supporting its utility for improving the reliability and interpretability of automated oral cancer classification.

Conclusion: This pilot study advances automated oral cytology by integrating deep learning with stain deconvolution to improve the multi-class classification of liquid-based Pap smear images. Our results show that targeted stain separation, particularly isolating the DAB channel, can significantly enhance diagnostic performance by improving true positive detection rates and reducing false negatives for high-grade lesions and carcinoma, which are critical for early intervention. This approach demonstrates that leveraging complementary stain information can increase the reliability and interpretability of AI-assisted cytological screening without adding unnecessary complexity to the laboratory workflow. Moreover, this framework has broader potential for adaptation to immunocytochemistry applications, where alternative staining protocols could similarly be optimized to support more accurate and explainable automated diagnosis.

KEYWORDS

Color deconvolution, Convolutional neural network, Liquid-based cytology, Oral cancer, Papanicolaou staining

---

* This paper presents a pilot study conducted using a publicly available cervical cancer dataset, prior to the main project, which will utilize an oral cytology dataset from Tokyo Dental College Hospital provided by Dr. Katsutoshi Kokubun.

NYU | DENTISTRY

INTRODUCTION

Oral cancer remains a significant public health concern in the United States, with an estimated 59,660 new cases and 12,770 deaths projected in 2025 alone [1]. Despite advances in treatment, the overall 5-year survival rate for oral cavity and pharyngeal cancers is only 69%, and significantly lower (57%) among Black individuals [2]. However, when detected early in stages I or II, survival can surpass 80% [3]. Despite this, up to half of oral cancer cases are identified only at advanced stages (III or IV), largely because early disease is often asymptomatic. Most patients delay seeking medical care until noticeable symptoms such as pain, bleeding, or a palpable mass, often indicating lymph node involvement, emerge [4]. A diagnostic delay of more than one month significantly increases the likelihood of presenting with advanced-stage disease [5]. These statistics underscore an urgent need for early detection strategies, as survival outcomes are considerably higher when oral cancer is diagnosed at a localized stage.

In this context, accurate and timely diagnosis becomes critical for devising effective treatment plans and improving patient outcomes. While histological examination is considered the definitive diagnostic method, it is invasive and may cause discomfort similar to a blood test [4]. In contrast, oral cytology offers a less invasive and more convenient alternative that serves as a valuable screening tool for oral mucosal conditions [6]. In particular, oral brush liquid-based cytology (LBC), a method of exfoliative cytology, has demonstrated efficacy in screening for oral squamous cell carcinoma, making it a compelling option for early detection [7]. Additionally, oral liquid-based cytology is significantly more cost-effective, with estimates showing that LBC costs less than 26% of surgical biopsy procedures in dental specialty clinics and 36% less in general dental practices. Costs rise further in cases involving large or multifocal lesions requiring multiple biopsy sites [8].

The liquid-based cytology (LBC) method, in which cells are dispersed in a fixative solution and a thin layer of cells is prepared on a slide, reduces chair-side work and improves specimen quality by minimizing cell overlap and debris [9]. In practice, brush biopsies allow non-invasive sampling of suspicious oral lesions during routine dental visits. These samples are then stained and examined by cytopathologists under brightfield microscopy. Traditionally, this diagnostic process relies on manual screening, which is labor-intensive and time-consuming. A typical clinical workflow involves rough screening by a cytotechnologist or cytopathologist, with key diagnostic areas manually marked using an ink marker, followed by secondary review by a pathologist [10]. While experienced professionals can detect cytological abnormalities linked to malignancies, the process remains technically demanding and subject to variability. Indeed, recent reports indicate that patient-level sensitivity and specificity achieved by human experts for LBC oral cytology are approximately 80% and 86%, respectively—highlighting the challenge of detecting subtle malignant changes among thousands of cells on a slide [9].

Therefore, oral cytology is primarily used to screen patients and determine whether a definitive histopathological diagnosis is warranted [11]. To overcome current limitations and reduce diagnostic workload, integrating artificial intelligence (AI) with liquid-based cytology (LBC) has emerged as a promising solution. Automated, AI-assisted cytology screening systems have the potential to enhance diagnostic accuracy and standardize evaluation, thereby reducing the time and labor needed for cytological examination. Recent advances in deep learning (DL) have demonstrated significant promise in cytological applications by enabling automated feature extraction directly from digital slide images. Notably, unlike traditional manual interpretation, no human bias is associated with the feature extraction component of DL systems, which can help standardize screening and minimize inter-observer variability. DL methods can efficiently handle large volumes of digital image data and help avoid bias in cell selection during analysis [12].

Although there is increasing interest in AI-based decision support for cancer detection in cytology [13], the application of color deconvolution within deep learning frameworks for oral cancer detection remains limited. Previous studies in histopathology have shown that color preprocessing and stain separation can improve the performance of CNN-based models by isolating diagnostically meaningful features[14, 15]. However, color deconvolution has mostly been used as a preprocessing step for stain normalization, and explicit comparisons of how individual stain channels contribute to classification performance in cytology have not been explored. In particular, no prior work has systematically quantified the diagnostic impact of each stain channel for the automated multi-class classification of oral liquid-based cytology images. Accordingly, this study provides a systematic evaluation of how stain separation using color deconvolution influences the performance and interpretability of deep learning models for oral cancer screening.

MATERIALS AND METHODS

Study Design

The aim of this study was to develop and systematically evaluate deep learning models for the automated multi-class classification of liquid-based cytology (LBC) Pap smear images, with a particular focus on comparing the diagnostic performance

of models trained on original RGB images with stain-separated channels (Haematoxylin, Eosin, and DAB) using color deconvolution.

Sample Acquisition

This study utilizes the Liquid-Based Cytology Pap Smear (LBCP) dataset, which was developed as a benchmark resource for automated multi-class classification of pre-cancerous and cervical cancer lesions [16]. The LBCP dataset is designed in accordance with the diagnostic standards outlined in The Bethesda System (TBS) for cervical cytological reporting. A total of 962 liquid-based cytology (LBC) images are included in the dataset, which are categorized into four diagnostic classes: NILM (Negative for Intraepithelial Lesion or Malignancy), LSIL (Low-Grade Squamous Intraepithelial Lesion), HSIL (High-Grade Squamous Intraepithelial Lesion), and SCC (Squamous Cell Carcinoma). The detailed class distribution is presented in **Table 1**.

The cervical smear samples were collected from a cohort of 460 patients attending gynecological screening at three established medical diagnostic centers in Northeast India: Babina Diagnostic Pvt. Ltd., Imphal; Gauhati Medical College and Hospital, Guwahati; and Dr. B. Borooah Cancer Research Institute, Guwahati. The liquid-based cytology (LBC) method using *BD Sure Path* was employed to ensure uniform cell fixation and a clear background, which offers significant advantages over conventional Pap smear techniques. The cervical cell samples were stained using the standard Haematoxylin and Eosin (H&E) staining protocol.

Prepared slides were examined and imaged by experienced pathologists to ensure accurate diagnostic categorization. Images were acquired using a Leica DM 750 microscope integrated with an ICC50 HD camera, under 400× magnification (combining a 40× objective lens and a 10× eyepiece). Images were captured at a resolution of 2048×1536 pixels in JPG format. Slides were scanned systematically to reduce overlap and ensure distinct coverage of cellular regions.

Table 1. Number of images per class in the Pap smear dataset.

Class	Number of images
NILM	612
LSIL	113
HSIL	163
SCC	74
Total	**962**

CNN Model Architecture

In this study, we adopted EfficientNet, a state-of-the-art convolutional neural network (CNN) architecture for image classification. EfficientNet was proposed to achieve superior accuracy and efficiency by introducing a novel *compound scaling* method that uniformly scales network depth, width, and input resolution with a fixed ratio [17]. Compared to earlier CNNs such as ResNet, DenseNet, Inception, and Xception, EfficientNet provides significantly better performance with fewer parameters and lower computational cost, making it well-suited for high-resolution image analysis tasks. EfficientNet scales up from the base B0 model to B7 step by step, with image recognition accuracy improving as the base size increases. However, because the parameters and computational complexity increase exponentially as EfficientNet scales up, we utilized the lightest variant, EfficientNet-B0, as our classification model, with its parameters initialized by pretrained ImageNet weights.

Table 2. shows the detailed baseline architecture configuration for EfficientNet-B0 used in our study, starting with an input image resolution of 512×512 pixels. The model begins with an initial 3×3 convolution layer, followed by a series of *mobile inverted bottleneck convolution* (MBConv) blocks with different kernel sizes (3×3 and 5×5), strides, and expansion ratios. The resolution progressively decreases through the stages, while the number of channels (#Channels) increases to capture higher-level semantic features. The final stage includes a 1×1 convolution, global average pooling, and a fully connected layer for classification.

Color Deconvolution

Color deconvolution is a widely used digital image analysis technique that enables the separation of overlapping stains in histopathology and cytology images. In hematoxylin and eosin (H&E)-stained slides, cellular structures are visualized by

Table 2. EfficientNet-B0 baseline network configuration for 512×512 input.

Stage	Operator	Resolution	#Channels	#Layers
1	Conv3x3	512×512	32	1
2	MBConv1, 3x3	256×256	16	1
3	MBConv6, 3x3	256×256	24	2
4	MBConv6, 5x5	128×128	40	2
5	MBConv6, 3x3	64×64	80	3
6	MBConv6, 5x5	32×32	112	3
7	MBConv6, 5x5	16×16	192	4
8	MBConv6, 3x3	16×16	320	1
9	Conv1x1 + Global Avg Pooling + FC	1×1	1280	1

combining multiple dyes, each with distinct absorbance properties in the red, green, and blue (RGB) channels. This overlap complicates automated analysis because each pixel's RGB value represents the combined effect of multiple stains. To address this, dye unmixing mathematically estimates the contribution of each stain by leveraging the fact that the optical density (OD) at a given pixel is linearly related to the concentration of the stain, according to the Lambert-Beer law [18].

In this study, the classical method proposed by Ruifrok and Johnston [19] was adopted for color deconvolution, which transforms RGB pixel intensities into OD space and applies an orthonormal matrix transformation to isolate the concentration of each stain component. Specifically, the original H&E images were decomposed into separate channels representing Haematoxylin, Eosin, and DAB (3,3-Diaminobenzidine). Haematoxylin primarily highlights cell nuclei and chromatin structure, Eosin emphasizes cytoplasmic and extracellular regions, and DAB serves as an immunohistochemical marker for protein expression.

Each stain is characterized by its specific absorbance across the RGB channels. In this work, the OD transformation for each pixel is defined as:

$$OD_C = -\log_{10}\left(\frac{I_C}{I_{0,C}}\right), \quad C \in \{R, G, B\}.$$

Here, I_C and $I_{0,C}$ represent the transmitted and incident light intensities for each channel C. The normalized OD vector for each stain enables an orthonormal transformation to unmix stains. The color deconvolution relies on a stain-specific OD matrix M, where each row represents the optical density profile of a pure stain across RGB channels. For H&E with DAB staining, the following matrix was used:

$$M = \begin{bmatrix} 0.18 & 0.20 & 0.08 \\ 0.01 & 0.13 & 0.01 \\ 0.10 & 0.21 & 0.29 \end{bmatrix}.$$

To estimate the relative abundance $\mathbf{C}$ of the three stains at each pixel, the OD vector $\mathbf{y}$ is multiplied by the inverse of M:

$$\mathbf{C} = M^{-1} \cdot \mathbf{y}.$$

This yields orthogonal concentration estimates for the Haematoxylin, Eosin, and DAB channels, producing separate images in which overlapping stain contributions are minimized. The results of this separation are presented in **Figure 4.**, which illustrates the extracted (a) Haematoxylin, (b) Eosin, and (c) DAB channels in the top row color format. The bottom row shows representative grayscale images that highlight regions of interest for each stain component: (d) Haematoxylin, (e) Eosin, and (f) DAB.

Each original H&E-stained liquid-based cytology (LBC) image was processed to generate three separate channel images corresponding to Haematoxylin, Eosin, and DAB. These stain-specific images were organized as independent input datasets, and a total of four convolutional neural network (CNN) models were developed and trained: one for the original RGB images and one each for the Haematoxylin, Eosin, and DAB channels.

EXPERIMENTAL SETUP

Data Partitioning

For data separation, a stratified 4-fold cross-validation procedure was employed [20]. Formally, the dataset $\mathcal{D}$ was randomly divided into $k = 4$ mutually exclusive subsets (folds) $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_k$ of approximately equal size. Stratification was applied to

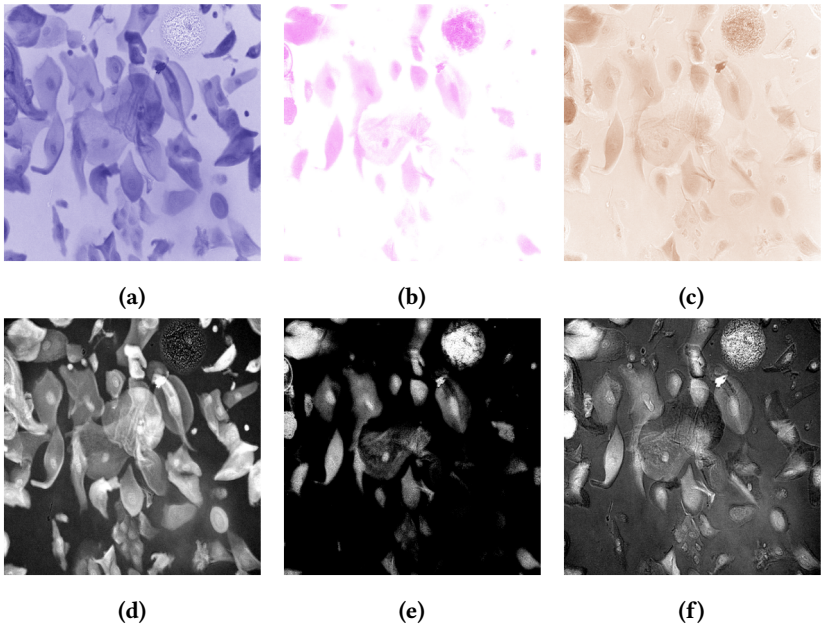


Figure 1. Color deconvolution results showing stain-separated channels: (a) Haematoxylin, (b) Eosin, (c) DAB in color format; and their corresponding grayscale concentration maps: (d) Haematoxylin, (e) Eosin, (f) DAB.

ensure that the proportion of each diagnostic class was maintained in every fold. For each iteration $t \in \{1, 2, \ldots, k\}$, the model was trained on $\mathcal{D} \setminus \mathcal{D}_{(t)}$ and validated on the held-out fold $\mathcal{D}_{(t)}$. Each instance thus appeared exactly once in the test set and $k - 1$ times in the training set. The cross-validation estimate of accuracy is defined as:

$$\text{acc}_{\text{CV}} = \frac{1}{n} \sum_{(x_i, y_i) \in \mathcal{D}} \delta\left(I(\mathcal{D} \setminus \mathcal{D}_{(t)}, x_i),\ y_i\right)$$

where:

- n is the total number of instances in the dataset,
- $\delta(\cdot)$ is an indicator function equal to 1 if the predicted label matches the true label,
- $I(\mathcal{D} \setminus \mathcal{D}_{(t)}, x_i)$ denotes the prediction for instance x_i by the model trained on the corresponding training folds.

Repeating this process for all folds and averaging the results provides an estimate of the model's generalization performance that is less biased than a single holdout set. An overview of this cross-validation strategy is presented in **Figure 2.**, which depicts the iterative partitioning and evaluation scheme.

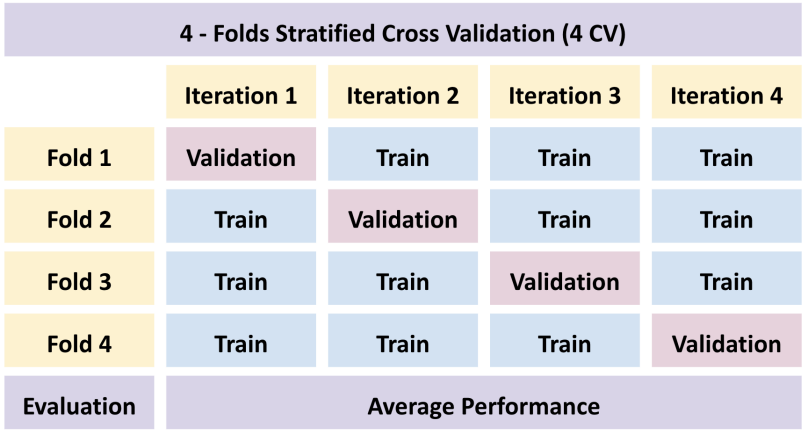
Data Augmentation

All images were resized to 512×512 pixels prior to training. Using larger input dimensions has been shown to improve image recognition performance by preserving more relevant morphological features [21]. Pixel intensities were normalized to the [0,1] range for consistent input scaling. Training data were augmented using a combination of geometric and photometric transformations, including random horizontal and vertical flipping, random rotations within a ±10-degree range, and random brightness variations up to ±10%. These augmentations were performed during training using TensorFlow. Images in the validation sets were processed only with resizing and normalization, without augmentation, to ensure unbiased performance evaluation.

Training Details

All models were trained using the EfficientNet B0 architecture with input images resized to 512×512 pixels. Model weights were initialized with ImageNet-pretrained weights. The optimizer used was Adam, an adaptive first-order gradient-based optimizer that combines momentum and RMSProp principles for efficient parameter updates in deep learning models [22].

Figure 2. Schematic overview of the stratified 4-fold cross-validation procedure used for training and evaluation. In each iteration, three folds were used for training the CNN model and the remaining fold for validation. This process was repeated four times so that each fold served once as the validation set.



The initial learning rate was set to 1×10^{-5}. A *ReduceLROnPlateau* learning rate scheduler was applied to adaptively reduce the learning rate when no improvement in validation loss was observed for three consecutive epochs. Specifically, the learning rate update follows:

$$\eta_{t+1} = \begin{cases} \eta_t \times \gamma, & \text{if no improvement for } p \text{ epochs} \\ \eta_t, & \text{otherwise} \end{cases} \qquad \text{with} \quad \gamma = 0.5, \quad p = 3.$$

The minimum learning rate was constrained to 1×10^{-7} to avoid excessively small updates. To prevent overfitting, *Early Stopping* with a patience of 10 epochs was employed, restoring the best weights observed during training. The maximum number of training epochs was set to 30, with the final model selected based on the lowest validation loss.

A batch size of 16 was used to balance computational efficiency and stable gradient estimation. The loss function used for this four-class classification task was the sparse categorical cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log p_{y_i}, \quad y_i \in \{0, 1, 2, 3\},$$

where p_{y_i} denotes the predicted probability for the true class label of sample i. All methods were implemented using TensorFlow and Keras, along with Python libraries including NumPy, OpenCV, Pandas, SciPy, and scikit-posthocs. Experiments were conducted using Google Colab Pro with an NVIDIA Tesla T4 GPU.

RESULTS AND ANALYSIS

Model Evaluation

We first conducted a comparative evaluation of the classification performance of the original RGB images and the stain-deconvoluted channels (Haematoxylin, Eosin, and DAB) using 4-fold cross-validation. As presented in **Table 3.**, the DAB channel achieves the highest diagnostic performance across all reported metrics, with an average F1 score of 0.958 ± 0.019 and accuracy of 0.982 ± 0.008. The performance suggests that the immunohistochemical contrast captured by the DAB stain provides highly discriminative features that aid in distinguishing subtle morphological differences among oral cancer subtypes.

The non-deconvoluted RGB model also maintains strong performance, with a lower F1 score (0.952 ± 0.048) and accuracy (0.979 ± 0.020) compared to DAB. However, its larger standard deviation indicates that its predictions are less consistent across folds, suggesting that mixed stain signals may introduce more variability compared to focused, stain-specific features. The Hematoxylin channel achieves competitive results (F1 score: 0.937 ± 0.038; accuracy: 0.972 ± 0.016; ROC AUC: 0.996 ± 0.002). Its lower precision and recall compared to DAB suggest that it introduces more false positives and false negatives when used alone. The Eosin channel consistently shows the weakest performance across all metrics (F1 score: 0.649 ± 0.040; recall:

0.628 ± 0.039), indicating that cytoplasmic and stromal staining alone provides insufficient discriminatory signal for robust classification. Notably, its relatively higher precision compared to recall suggests that while its positive predictions tend to be correct, many true positives are missed, a pattern that could lead to under-detection of abnormal cells in practice.

Table 3. Performance of the non-deconvoluted RGB channel model and stain-deconvoluted channel models under 4-fold cross-validation. Reported metrics are average F1 score, accuracy, ROC AUC, recall, and precision (± standard deviation). The highest averages for each metric are highlighted in bold.

Method	F1 ± std	Accuracy ± std	ROC AUC ± std	Recall ± std	Precision ± std
RGB	0.952 ± 0.048	0.979 ± 0.020	0.997 ± 0.003	0.951 ± 0.046	0.954 ± 0.050
H	0.937 ± 0.038	0.972 ± 0.016	0.996 ± 0.002	0.938 ± 0.035	0.942 ± 0.038
E	0.649 ± 0.040	0.820 ± 0.024	0.936 ± 0.022	0.628 ± 0.039	0.760 ± 0.081
DAB	**0.958 ± 0.019**	**0.982 ± 0.008**	**0.999 ± 0.001**	**0.959 ± 0.023**	**0.962 ± 0.019**

Per-Class Detection Performance

Detection performance across classes is summarized in the confusion matrix results in **Table 4.**. The DAB channel consistently achieves the highest true positive (TP) and true negative (TN) counts for every class, especially for the more clinically significant HSIL and SCC categories, which are more challenging to detect accurately. In contrast, the RGB model performs nearly as well as DAB for NILM (TP = 612) and LSIL (TP = 113) but falls behind for HSIL and SCC, where precise detection is crucial. The Eosin channel shows the highest false positives and false negatives across all classes indicating frequent misclassifications when used alone. Hematoxylin performs strongly for normal and low-grade classes, but does not surpass DAB for HSIL and SCC.

Table 4. Per-class confusion matrix calculated on the validation sets from 4-fold cross-validation. Shown are total True Negatives (TN), False Positives (FP), False Negatives (FN), and True Positives (TP) for each model and class (NILM, LSIL, HSIL, SCC). Higher TN and TP (↑) values indicate better performance, while lower FP and FN (↓) are better. Best values per class are highlighted in bold.

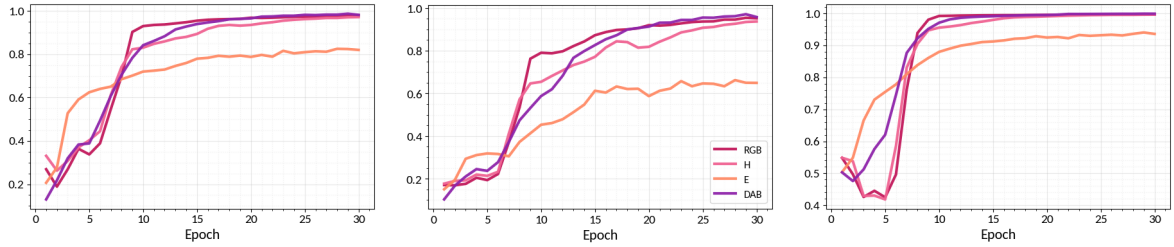
Class	**Method**	**TN↑**	**FP↓**	**FN↓**	**TP↑**
NILM	RGB	**349**	**1**	**0**	**612**
	H	348	2	1	611
	E	269	81	19	593
	DAB	**349**	**1**	**0**	**612**
LSIL	RGB	847	2	**0**	**113**
	H	846	3	**0**	**113**
	E	820	29	36	77
	DAB	**849**	**0**	**0**	**113**
HSIL	RGB	791	8	10	153
	H	790	9	14	149
	E	747	52	51	112
	DAB	**792**	**7**	**6**	**157**
SCC	RGB	879	9	10	64
	H	875	13	12	62
	E	**885**	**3**	59	15
	DAB	883	5	**7**	**67**

Learning Curve Analysis

Validation performance curves were analyzed for the best-performing fold of each model, all of which achieved their highest performance on fold 3, as shown in **Figure 3**. During the initial epochs, there is marginal variation between input configurations, with the Eosin channel even exhibiting a slightly higher mean validation accuracy and F1 score compared to the others up to approximately epochs 5–8. However, the Eosin curve gradually flattens, and a clear performance gap becomes evident as its validation metrics plateau while the RGB, Hematoxylin, and DAB channels continue to improve across accuracy, F1 score, and ROC AUC.

The RGB, Hematoxylin, and DAB channels all maintain positive learning curves. Notably, the DAB curve exhibits stable, monotonic improvement with minimal fluctuation, and it shows a steeper ascent, leading in both accuracy and ROC AUC. The DAB input demonstrates a steady increase from the outset, highlighting the immediate benefit of its clear IHC-based class separation. In contrast, the RGB and Hematoxylin channels show minor early variance before converging upward. During the early training phase (epochs 0–5), both inputs undergo a slight local minimum in ROC AUC and accuracy before improving.

Figure 3. Validation performance curves for each input configuration across 4-fold cross-validation: (a) Accuracy, (b) F1 score, and (c) ROC AUC.



(a) **(b)** **(c)**

Statistical Evaluation

To evaluate the significance of the performance differences observed in the validation learning curves, we applied the non-parametric Friedman test to evaluate whether there are significant differences between models across multiple folds [23]. The test statistic was computed as:

$$Q = \frac{12N}{k(k+1)} \left[\sum_{j=1}^{k} \bar{R}_j^2 \right] - 3N(k+1)$$

where:

- N is the number of blocks (folds, $N = 4$),
- k is the number of groups (models, $k = 4$),
- $\bar{R}_j$ is the average rank of the j-th model across folds.

For this study, the test statistic was $Q = 8.10$ with a p-value of 0.044, indicating that there is a statistically significant difference among the models ($p < 0.05$).

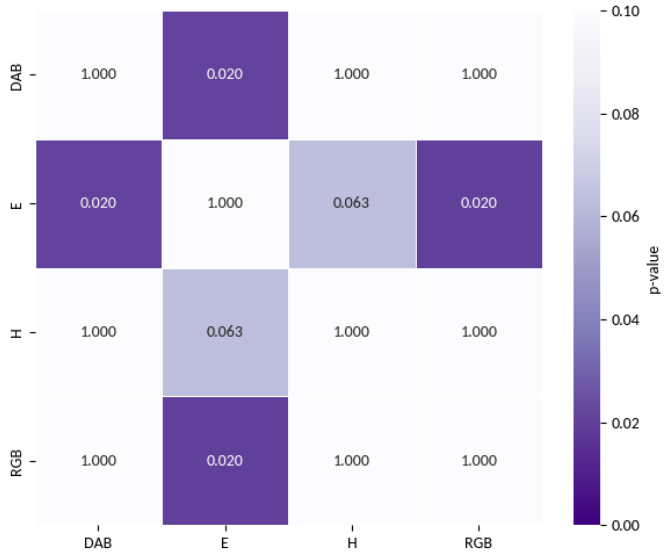
Following a significant Friedman result, pairwise comparisons were performed using Conover's post hoc test, which compares the absolute difference between the average ranks of each pair of models [24]:

$$T_{ij} = \frac{|\bar{R}_i - \bar{R}_j|}{SE}, \quad SE = \sqrt{\frac{k(k+1)}{6N}}$$

where SE is the standard error of the differences between ranks.

To control for the increased risk of Type I errors due to multiple pairwise comparisons, the resulting p-values were adjusted using Holm's sequential step-down procedure. Holm's method works by ordering the unadjusted p-values in ascending order

Figure 4. Conover's Post Hoc Pairwise Comparison Matrix



$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ and comparing each $p_{(i)}$ to $\alpha/(m-i+1)$, where m is the total number of comparisons and α is the desired familywise error rate. This provides greater statistical power than the Bonferroni method while still controlling the familywise error rate. The final pairwise adjusted p-values are visualized in the Conover's post hoc matrix shown in **Figure 4**.

To assess the relative performance of the different models, the mean F1 scores obtained on the validation sets from each of the four cross-validation folds were compared using non-parametric statistical testing. The Friedman test was first applied to evaluate whether there were overall significant differences in F1 performance across the RGB, H, E, and DAB models. The test yielded a statistically significant result ($Q = 8.10$, $p = 0.044$), indicating that at least one model's performance differed significantly from the others.

Pairwise comparisons were conducted using Conover's post hoc test with Holm's sequential step-down adjustment to control the familywise error rate. The results showed that the stain-deconvoluted Eosin channel model achieved significantly lower F1 scores than both the RGB ($p = 0.020$) and DAB ($p = 0.020$) models. The difference between the E channel and the H channel did not reach statistical significance at the 0.05 level ($p = 0.063$) but suggested a trend toward lower performance. No significant differences were observed among the RGB, H, and DAB models (all $p > 0.05$), indicating that these three models performed comparably in terms of average F1 score.

Visualization of Classification using Grad-CAM

The classification of each model was visualized using gradient-weighted class activation mapping (Grad-CAM) [25]. Grad-CAM targets CNN-based image recognition models. This method provides a basis for the model's decision by weighting the gradient with respect to the predicted value. In this study, a heat map was used to highlight the regions that served as the basis for classification according to their importance. Grad-CAM uses the final convolutional layer of the EfficientNet model. **Figure 5** shows the visualization of the area of interest for classification decisions in the deep learning model.

The Grad-CAM heatmaps confirm that all input configurations correctly concentrate on the uniform, small nuclei typical of normal squamous epithelial cells, with no misdirected activation in the background or cytoplasm for NILM label. The RGB, Haematoxylin and DAB channels show the clearest and most sharply defined nuclear focus, demonstrating strong recognition of chromatin uniformity and nuclear spacing. The Eosin channel provides similar nuclear alignment but with more diffuse emphasis.

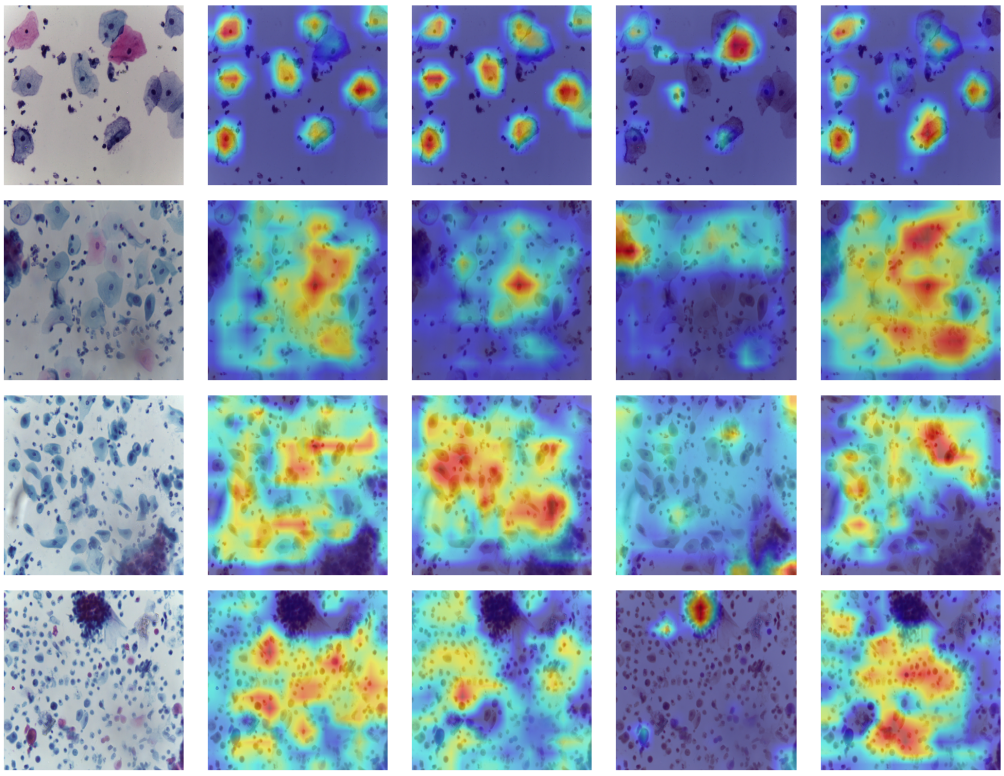
For LSIL, the original image shows mild nuclear enlargement, irregular chromatin, and an inflammatory background typical of low-grade squamous intraepithelial lesions. The RGB, Haematoxylin, and DAB models consistently focus on the clusters of dysplastic nuclei. The RGB heatmap spreads across multiple nuclei, suggesting it captures the overall pattern of mild dysplasia. Haematoxylin shows a more concentrated focus on a single prominent nucleus with intense chromatin, strength in isolating nuclear detail. DAB distributes heat across multiple nuclei with greater intensity than RGB, indicating stronger signal

for identifying atypical nuclei. The Eosin channel does not focus on the nuclei and instead highlights irrelevant background areas, demonstrating poor localization for LSIL features.

The visualizations for HSIL show that the RGB and DAB models reliably capture the irregular, enlarged nuclei with hyperchromasia, with well-dispersed focus across multiple dysplastic cells. The Haematoxylin input produces the strongest overall signal, correctly enhancing hyperchromatic regions but spreading attention more diffusely to other nuclei, which may reduce specificity. The Eosin channel does not localize the nuclear abnormalities, instead highlighting scattered, non-diagnostic areas.

For SCC, the Grad-CAM heatmaps demonstrate that the RGB, Haematoxylin, and DAB models all successfully highlight the abnormal nuclear pleomorphism and hyperchromasia that characterize squamous cell carcinoma, with well-dispersed focus across multiple irregular nuclei. The DAB and RGB models show the greatest intensity and coverage, indicating strong detection of both nuclear and cytoplasmic irregularities within the inflammatory background. The Eosin channel again fails to localize meaningful nuclear features.

Figure 5: Grad-CAM visualizations for all diagnostic classes (NILM, LSIL, HSIL, SCC) top to bottom; and input configurations (a) Original, (b) RGB, (c) Haematoxylin, (d) Eosin, (e) DAB.



CONCLUSION

In this study, we present a comprehensive comparison of stain-separation strategies for deep learning-based multi-class classification of Pap smear LBC images. Our experiments showed that the DAB channel yielded the highest F1 score (0.958±0.019) and accuracy (0.982±0.008), outperforming the non-deconvoluted RGB (F1: 0.952±0.048) and Haematoxylin (F1: 0.937±0.038) channels. Confusion matrix results confirmed that DAB achieved the highest true positive and true negative detection for higher-grade lesions of HSIL (TP=157) and SCC (TP=67). In contrast, the Eosin channel showed significantly lower F1 (0.649±0.040) and higher misclassifications, consistent with its poor Grad-CAM localization. These results suggest that applying color deconvolution to focus on distinct stain-specific features, particularly the DAB channel, can play an important role in supporting automated models for primary screening in the oral cytological diagnostic environment.

DISCUSSION

Our research focused on classification performance using color deconvolution techniques to isolate each stain channel and systematically evaluate their individual contributions. However, this study has some limitations.

First, we did not consider the effects of other color preprocessing techniques such as color normalization (both color transfer and color constancy) and color augmentation. Future work should systematically compare these preprocessing methods alongside stain deconvolution to clarify their individual and combined impacts on model performance.

Second, further investigation is needed on the effects of data augmentation strategies. Both geometric and photometric augmentations should be explored in detail. In particular, the impact of photometric augmentation, including variations in illumination conditions, should be carefully considered. Techniques for background microscopy illumination correction — such as polynomial fitting, rolling ball algorithms, Gaussian blur, or entropy minimization — can help reduce unwanted background signals in RGB images. It is crucial to examine how these corrections interact with stain deconvolution by comparing model performance with and without background correction.

Finally, our work did not address the impact of hyperparameter tuning and optimizer selection. Previous research has shown that the Sharpness-Aware Minimization (SAM) optimizer can improve performance in liquid-based cytology diagnosis [26]. Therefore, comparative evaluations of different optimizers (e.g., SGD, Adam, and SAM) and learning rate scheduling strategies should be conducted to understand their effects in the context of stain-specific deep learning models.

Modifications for Actual Project

- The pilot study used H&E staining on a publicly available dataset; however, the dataset to be obtained from Dr. Kokubun's group uses Papanicolaou (Pap) staining, which includes dyes such as OG-6 and EA, and therefore requires different optical density (OD) values for accurate stain separation. Consequently, a major modification for the actual project will be to adjust the OD values and potentially the color deconvolution method. Papanicolaou staining employs up to five dyes, which makes simple linear color unmixing infeasible with only RGB channels. Therefore, additional methods, such as multispectral spectral unmixing, should be considered [27]. The observation model for Pap-stained images also needs to account for a nonlinear transform based on the Lambert–Beer law, which differs from the standard approach used for H&E staining.
- It is also important to note that the pilot dataset was cervical cytology, which may differ in cellular morphology and diagnostic patterns from oral cytology. Although this pilot project used the SurePath method for staining and a cervical dataset, it served as a useful preliminary step to validate our pipeline and minimize research delays. Insights gained will inform the adjustments needed for the actual study. The new dataset from Dr. Kokubun's group will follow diagnostic criteria specific to oral cytology, categorizing samples into five diagnostic classes: NILM (Negative for Intraepithelial Lesion or Malignancy), OLSIL (Oral Low-Grade Squamous Intraepithelial Lesion), OHSIL (Oral High-Grade Squamous Intraepithelial Lesion), SCC (Squamous Cell Carcinoma), and IFN (Indefinite for Neoplasia) according to the JSCC 2015 diagnostic guidelines [28].
- One concern is the computational limitation of using Google Colab Pro. The pilot study, with fewer than 1,000 images, was manageable in terms of GPU memory. However, the planned dataset is expected to exceed 15,000 images. It remains unclear whether Google Colab will be sufficient for larger-scale experiments, particularly when performing extensive hyperparameter tuning. Additional computing resources may be required to handle larger data volumes and more complex experiments efficiently.
- Therefore, we plan to request detailed information on the specific staining dyes used (e.g., OG, EA, EY, hematoxylin), including the manufacturer and the corresponding OD values for each RGB channel. This will enable appropriate color deconvolution for the Pap-stained slides. Additionally, for each classified class, we aim to conduct detailed Grad-CAM analyses and will seek Dr. Kokubun's expert advice to ensure the analysis aligns with professional diagnostic standards.

Data Availability

The dataset used in this study is publicly available on Mendeley Data at https://data.mendeley.com/datasets/zddtpgzv63/4.

Code Availability

All source code used for data preprocessing, color deconvolution, model training, and performance evaluation in this study is publicly available at https://github.com/psj03283/oral-cancer.

REFERENCES

[1] Rebecca L. Siegel, Taylor B. Kratzer, Annah R. Giaquinto, Hyuna Sung, and Ahmedin Jemal. Cancer statistics, 2025. *CA: A Cancer Journal for Clinicians*, 75(1):10–45, 2025.

[2] American Cancer Society. Cancer facts and figures 2025, 2025. American Cancer Society.

[3] Sol Silverman, A. R. Kerr, and J. B. Epstein. Oral and pharyngeal cancer control and early detection. *Journal of Cancer Education*, 25:279–281, 2010.

[4] Michael McCullough, G. Prasad, and C. Farah. Oral mucosal malignancy and potentially malignant lesions: An update on the epidemiology, risk factors, diagnosis and management. *Australian Dental Journal*, 55(suppl. 1):61–65, 2010.

[5] Isabel Gómez and et al. Is diagnostic delay related to advanced-stage oral cancer? a meta-analysis. *European Journal of Oral Sciences*, 117:541–546, 2009.

[6] Hadeel Alsarraf, Asgeir Kujan, and Omar C.S. Farah. The utility of oral brush cytology in the early detection of oral cancer and oral potentially malignant disorders: A systematic review. *Journal of Oral Pathology & Medicine*, 47:104–116, 2018.

[7] Katsutoshi Kokubun and et al. Evaluation of oral brush liquid-based cytology for oral squamous cell carcinoma: A comparative study of cytological and histological diagnoses at a single center. *BMC Oral Health*, 23(1):Article 145, 2023.

[8] Majdy Idrees and et al. Oral brush biopsy using liquid-based cytology is a reliable tool for oral cancer screening: A cost-utility analysis. *Cancer Cytopathology*, 130(9):740–748, 2022.

[9] S. Sukegawa and et al. Clinical study on primary screening of oral cancer and precancerous lesions by oral cytology. *Diagnostic Pathology*, 15:1–6, 2020.

[10] S. Sornapudi and et al. Comparing deep learning models for multi-cell classification in liquid-based cervical cytology images. In *U.S. National Library*, pages 820–827, Missouri University of Science and Technology Rolla MO USA; Lister Hill National Center for Biomedical Communications, 2019.

[11] J. Sekine, E. Nakatani, K. Hideshima, T. Iwahashi, and H. Sasaki. Diagnostic accuracy of oral cancer cytology in a pilot study. *Diagnostic Pathology*, 12:27, 2017.

[12] Pranab Dey. The emerging role of deep learning in cytology. *Cytopathology: Official Journal of the British Society for Clinical Cytology*, 32(2):154–160, 2021.

[13] MS Landau and L Pantanowitz. Artificial intelligence in cytopathology: A review of the literature and overview of commercial landscape. *Journal of the American Society of Cytopathology*, 8(4):230–241, 2019.

[14] Massimo Salvi, Filippo Molinari, and Alessandra Bert. The impact of pre-and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Computers in Biology and Medicine*, 128:104129, 2021.

[15] Zhu He, Zexuan Liu, Bin Song, and Xuelong Li. Deconv-transformer (dect): A histopathological image classification model for breast cancer based on color deconvolution and transformer architecture. *Information Sciences*, 608:1093–1112, 2022.

[16] E. Hussain, L.B. Mahanta, H. Borah, and C.R. Das. Liquid-based cytology pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. Available dataset.

[17] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[18] P. Haub and T. Meckel. A model based survey of colour deconvolution in diagnostic brightfield microscopy: Error estimation and spectral consideration. *Scientific Reports*, 5:12096, 2015.

[19] A. Ruifrok and D. Johnston. Quantification of histochemical staining by color deconvolution. *Analytical and Quantitative Cytology and Histology*, 23:291–299, 2001.

[20] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, volume 14, pages 1137–1143, 1995.

[21] Kaori Oya, Yasushi Oya, Hiroshi Yoshimura, Noriyuki Ikeda, Kazuya Oka, Kohei Morita, and Ken-ichi Mukaisho. Oral squamous cell carcinoma diagnosis in digitized histological images using convolutional neural network. *Journal of Dental Sciences*, 18(1):322–329, 2023.

[22] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, pages 1–15, 2015.

[23] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

[24] W. J. Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, New York, 3rd edition, 1999.

[25] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2020.

[26] S. Sukegawa, Kazuhiro Takabatake, Hideki Kawai, Kazuhiko Nakano, Hiroko Nagatsuka, Hiroshi Fujita, and Hideki Nagatsuka. Effective deep learning for oral exfoliative cytology classification. *Scientific Reports*, 12(1):13281, 2022.

[27] Saori Takeyama, Kazuyoshi Iwata, Tatsuya Abe, Yoshihiro Ohtake, Motohiko Nagayama, Tatsuya Nagai, and Yoichi Tanaka. Dye amount quantification of papanicolaou-stained cytological images by multispectral unmixing: spectral analysis of cytoplasmic mucin. *Journal of Medical Imaging*, 12(1):017501, 2025.

[28] Motohiko Nagayama, Akinori Ihara, and Yoichi Tanaka. Oral cytology in japan: Its useful approach and criteria for early detection of carcinoma and precursor lesions. In *Inflammation and Oral Cancer*, pages 43–54. IntechOpen, 2022.