

A WHITE BOX METHOD TO CLASSIFY THE HEALTH STATUS OF ENDOMETRIAL TISSUE

Vidya Shree Suresh, Thea Yu, Nelly Madani, Kiran Naidoo-Edwardson

Department of Electrical and Computer Engineering, University of Waterloo

ABSTRACT

This report proposes a white box technique to classify disease conditions in endometrial tissue and achieves an accuracy of 80% on a set of 240 images. The approach first normalizes images into a similar color space, then performs semantic segmentation, then extracts relevant medical features that a pathologist would consider, and finally reports a classification using a logistic regression model. While this approach works reasonably well, it needs further improvement to address underlying artifacts present in endometrial tissue images.

Index Terms— Endometrial Cancer, Feature Extraction, Autoencoder, Segmentation

1. INTRODUCTION

Diagnosing endometrial tissue disease is difficult as tissue morphology changes throughout the menstrual cycle [1]. Furthermore, the distinguishing features for diagnosis are subtle and require experienced pathologists for diagnosis [1]. Additionally, they are prevalent - endometrial cancer alone affected over 300,000 women in 2018 [2]. The diagnosis of endometrial disease represents a significant bottleneck for pathologists, particularly in high-population countries [3]. To tackle this, researchers have built deep learning algorithms to automate the classification of conditions [4, 5]. While successful, none of these algorithms have reached the clinic possibly due to over-fitting and limited data which are common challenges in medical image use-case. Therefore, an image analysis tool that classifies tissue based on image features would be a powerful tool in diagnoses.

To make a diagnosis, a pathologist must distinguish between the following tissue states: (1) endometrial cancer (EC), (2) endometrial hyperplasia which is a pre-cancerous condition (3) endometrial polyps (EP), another possibly pre-cancerous condition and (4) normal endometrium [1]. To distinguish between the various conditions pathologists look at the features within the epithelial, glandular, and stromal areas of an image. These regions are shown as regions 1, 2, and 3 respectively in Fig. 1. In healthy cases, the glands made up of the epithelial and glandular areas, are close to

Thanks to Dr. Padma for her guidance in defining the features associated with healthy and abnormal H&E slides.

circular in morphology, the stromal area is sparse with white space and the nuclei within the epithelium, which are the dark purple circles are in a single layer columnar organization and all point towards the center of the gland. In unhealthy cases such as EH, EP and EC, pathologists look for several features when making a diagnosis. The first is the enlarged gland architecture as a sign of budding or finger-like projections, as shown in Fig. 1. This feature is shared across all unhealthy cases but the more extreme deviations point to more severe cases [1]. Another indication of unhealthy tissue is the increase in density and in extreme cases, the disappearance of the stromal area, to diagnose EH the gland to stroma area is generally greater than 55% [1] To differentiate between EH and EC, pathologists examine nuclei patterns within the epithelium [1]. In mild EH, nuclei form a single columnar layer, while in EC and extreme EH, they become stratified and disorganized, forming multiple layers [1]. Nuclei color changes, appearing smudged and darker in extreme EH and EC, are also considered. Lastly, if epithelial cells invade the stroma, it is a clear sign of EC [1]. Therefore, the remainder of this paper describes a pipeline to extract these features and classify the tissues based on them.

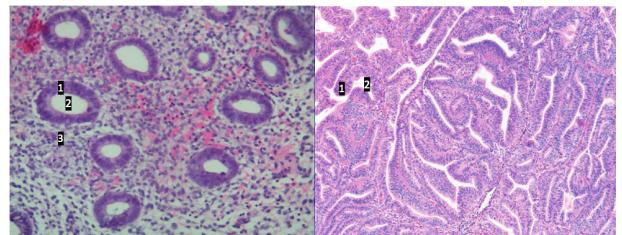


Fig. 1. Left: Healthy Follicular Phase Endometrial Tissue - (1) Epithelial Region, (2) Glandular Region (3) Stromal Region [6] Right: Endometrial Tissue with EC - (1) Epithelial Region, (2) Glandular Region [1]

2. DATASET AND PIPELINE OVERVIEW

The dataset comprises 3,302 images of the endometrium, derived from 498 hematoxylin and eosin (H&E) stained endometrial specimens. Each image is assigned a class label based on a diagnosis given by a group of three expert pathol-

ologists [4]. To ensure the quality of the dataset and to reduce the computational intensity, 240 images were selected from the original dataset due to artifacts such as tissue tears, blurred images, and other factors. To ensure the dataset was balanced, an equal number of images were taken for each class.

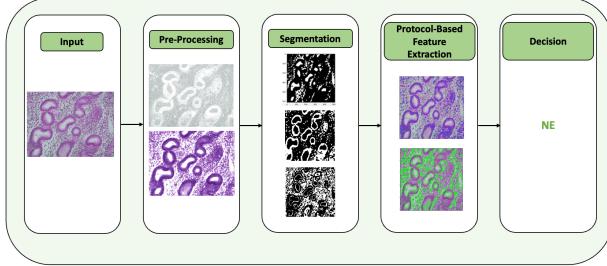


Fig. 2. Classification Pipeline

This report proposes a pipeline shown in Fig. 2: first, sparse stain separation followed by the structure-preserving color normalization (SPCN) [7] normalizes the image colors. Then, semantic region segmentation is performed by an autoencoder [8] combined with a hierarchical agglomerative clustering method (HAC) [9]. Next, features are extracted following a pathologist's protocol [1]. Finally, images are classified using logistic regression. [10].

3. PRE-PROCESSING

Pre-processing is performed to correct the image color variations present due to differing staining protocols. SPCN[7] was selected for its ability to preserve structural information [7] which is important for applications where structural details are critical for classification. Sparse stain separation and SPCN[7] were carried out using the code release from Vahadane [11]. This technique uses a reference image which for this report is taken from a histopathology textbook [1].

The stain separation process uses SPAMS [12] MATLAB optimization toolbox which solves unsupervised sparse estimation problems. Each image was decomposed into a stain density map (matrix H with i th row H_i , for each H and E stains ($i=2$)) and a stain matrix W using sparse non-negative matrix factorization of the Beer-Lambert transformation (V) of the RGB input image ($V = WH$). The stain matrix W represents the spectral profiles of the stains where values in each row correspond to the absorbance of light of each stain at different wavelengths (color bases for each stain). The goal is to minimize the Frobenius norm [13] between V and the approximation WH , subject to the constraints that W and H are non-negative and sparse. The non-convex optimization problem is as follows [7] (λ is the sparsity parameter):

$$\min_{W,H} \frac{1}{2} \|V - WH\|_F^2 + \lambda(\alpha\|W\|_1 + (1 - \alpha)\|H\|_1) \quad (1)$$

The sparsity constraint forces most pixels to respond to a single stain by penalizing large values in the matrices. This allows an accurate separation of stains. The mixing parameter α determines the balance between the sparsity of the W and H matrices. However, the implemented code uses the mexLasso function from SPAMS to estimate density maps by applying a least angle regression (LARS) algorithm with Cholesky-based implementation [14] to solve the optimization problem. This allows the algorithm to explore the entire regularization path without defining a fixed α value, leaving sparsity the only fixed hyperparameter. It was set at $\lambda = 0.03$ based on the sensitivity analysis by [7].

After decomposition of reference and source images into their color bases (W) and stain density maps (H), the normalization process is done by combining a scaled version of the density map of the source (H_s) with the color base of the reference (W_r instead of W_s) and an inverse Beer-Lambert transformation as the final step to obtain the normalized source image. This preserves the structure of the source images in terms of stain concentration maps.

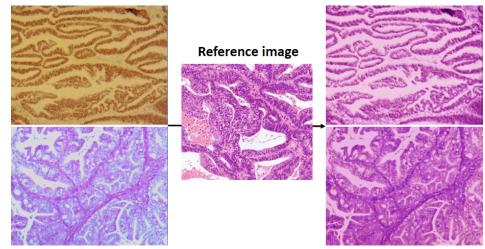


Fig. 3. SPCN[7] applied to EA_54 (top) and EA_69 (bottom). The reference is an EA case taken from [1].

4. AUTOENCODER

Autoencoder [8] is a neural network that learns to encode and decode data, where the bottleneck layer [15] represents important image features for reconstruction. The HAC [9], can be employed for image segmentation by clustering the essential features extracted by the bottleneck layer, thus enabling the division of the image into distinct regions based on these characteristics. The autoencoder in this report [8] was trained on a stain-separated subset of the training images to perform region-map segmentation. The encoding layers of the autoencoder [8] extract key image features, which are clustered into 4 masks using Hierarchical agglomerative clustering (HAC) [16] to obtain the gland, stroma, and epithelium regions of interest. HAC [9] is an unsupervised algorithm that groups similar data based on their pairwise distances. The mask-to-region correspondence is then decided by the color intensity and Euclidean distance between masks. Next, feature extractions are performed on each region mask.

The autoencoder [8] network is defined using the Keras

API in Python. The input image is a grayscale image with dimensions 512x512. Out of 240 images the train and test splits were 70% and 30% respectively. The encoder consists of three convolutional [15] layers with 128, 64, and 32 filters, respectively with Relu activation [17]. The encoded representation has a dimension of 16x16x32. The decoder consists of three transpose convolutional layers with the same number of filters as the corresponding encoder layers, followed by a sigmoid activation layer [15] to reconstruct the input image. The model is trained using the mean squared error (MSE) loss and the Adam optimizer [18]. Hyperparameter tuning was performed to fix the number of layers and neurons through grid search by specifying the value range manually. The training and test MSE are reported in Fig. 6.

HAC [9] is then used to group the image pixels into four clusters, which are assigned to different tissue components based on green channel intensity. The gland is typically observed to have the highest green channel reflection. Thus, gland label is assigned to the brightest cluster. HAC [9] was chosen over K-means [19] and DBSCAN [16] due to their limitations. K-means [19] assumes spherical clusters, which may not be appropriate for irregularly shaped data. DBSCAN requires the specification of hyperparameters, which can be difficult to scale to large datasets. In contrast, HAC [9] is flexible and can handle various data types [16].

The binary masks obtained from the clustering were processed further using morphological operations. Specifically, use erosion to remove small regions and dilation to fill mask gaps through OpenCV [20] library. The distance between the gland and other clusters was then calculated to generate an epithelium mask by assigning the pixels closest (in Euclidean distance) to the gland cluster to the epithelium mask.

5. FEATURE EXTRACTIONS

5.1. Gland Extractions

The characteristics of the gland reflect the health status of the sample 1. After identifying the binary mask corresponding to the gland (as described in Section 4), the objects of interest were highlighted by detecting the contours using the cv2.findContours() function [20]. For each contour (4), the area, perimeter, and aspect ratio were collected as features.

Another important feature is the percentage of overlap between a detected gland and a circle of the same area. This is calculated by creating two binary masks of the glandular structure and the circle and then taking their bitwise intersection. The overlap percentage is then calculated by dividing the sum of the intersection by the sum of the glandular structure mask and multiplying by a hundred as shown in Fig. 4.

To quantify the shape of the glands, thresholds were applied to the aspect ratio and circularity, which classifies the gland as tubular or branched. Then the count of tubular or circular glands was added to the feature set.

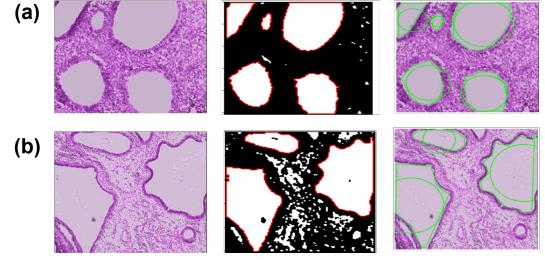


Fig. 4. The gland feature extraction process is shown above. Panel a) displays the raw image, segmented image, and circularity approximation (left to right) of the glandular structure in EH_Simple_47. Panel b) displays the same for the glandular structure in EH_Simple_14.

5.2. Nuclei Extractions

All nuclei features are calculated based on ellipse-approximated nuclei contours, which were extracted by using K-means [19] to segment the nuclei from the epithelium [19] and then finding contours and fitting ellipses using the OpenCV Python library section [20]. This is shown in Fig. 5 below.

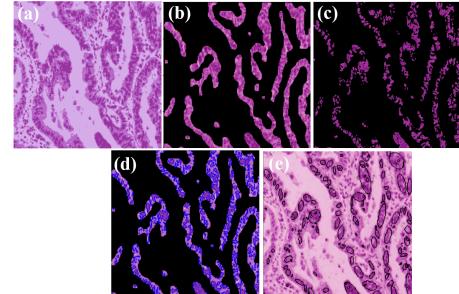


Fig. 5. Nuclei a) raw image b) epithelium mask c) K-means [19] clustered Nuclei d) Nuclei Contours e) Fit Ellipses

First, the standard deviation and mean of nuclei color in each RGB channel were extracted as features. As discussed in section 1, nuclei appear darker and vary more in color in unhealthy cases [1].

Next, the angle between the nuclei and the center of the gland was extracted. This feature was calculated by fitting ellipses to the contours of the nuclei and then calculating the difference in angle from the vertical axis to the direction of the major axis fit ellipse to the angle generated from drawing a vector from the center of the nearest gland to the nuclei. Here, it was expected that this difference in angle would be low for well-ordered nuclei and high for highly disordered nuclei based on the features discussed in section 1 [1].

Another feature is the standard deviation of gland-to-nuclei distances. To extract it, the minimum distances between every nucleus center and the contour of its associated gland were found. Then, for each gland, all the gland-to-

nuclei distances were normalized by the range of the distances. Then the standard deviation of normalized distances was calculated. Finally, the feature was recorded as a flag where it would be set to true if the proportion of glands with a standard deviation greater than a threshold (empirically determined as 20%).

The final feature extracted was the regularity of the shape. Healthy nuclei look much more circular and distinct. To get an idea of the shape, the contour of the nuclei was overlaid with a circle equal in area to the contour and the percentage of overlap area was calculated. It was expected that more irregular shapes would have lower overlap percentages compared to regularly shaped nuclei based on the unhealthy image case [1]. All these features were summarized for each image by taking global statistics of each feature like the mean, median, and standard deviation. This resulted in a 22-dimensional feature vector to describe each image.

6. CLASSIFIER

The labeled 22-dimensional feature vector is then scaled such that all features have the same scale and then passed into a logistic regression classifier with an l2 regularization parameter of 0.4 and 10-fold cross-validation to prevent overfitting. Logistic regression is a widely used classification model, often applied in tasks such as image recognition in which the posterior probabilities of a class are computed by applying a logistic function to the output of a linear predictor. SKlearn's logistic regression [21] with a one-vs-rest (OvR) [15] approach was implemented in which a separate binary classifier (with sigmoid logistic function [15]) is trained for each class. During testing, the classifier with the highest predicted probability is selected as the output class. A test size of 0.3 was used for our dataset of 240 images. The average four-class cross-validation accuracy was recorded as 0.81. The classification summary of the fit model by labels is shown in Table 1 for the test data and training and test accuracies shown over 200 trials in Fig. 6. In an OvR case, Precision, Recall, and F1-score are measured for a single class against all the other classes [21].

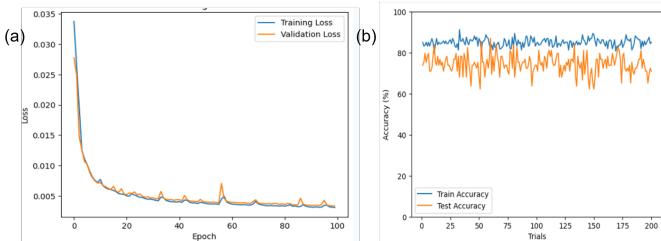


Fig. 6. a) MSE of training vs test of Autoencoder for 100 epochs b) Accuracy of training vs test of Multilabel Logistic regression based on 200 trials

7. DISCUSSION

After fitting the logistic regression model, feature importance was determined by examining the feature coefficients, which represent the importance of features. Larger coefficients represent correlates to increased importance. In this model, the most critical nuclei features were the mean of nuclei color in the blue ($m_1=2.23$) and green channels ($m_2=1.62$) and the shape and orientation of the nuclei with respect to the gland center ($m_3=0.52$). The most critical gland features were the gland area means ($m_4=0.3$) and the proportion of area occupied by the gland, and epithelium in the image ($m_5=0.2$). The findings align with the significance that pathologists attribute to the color and arrangement of nuclei, glandular shape, and the shift in gland-to-stroma ratio resulting from gland invasion into the stroma.

However, according to Table 1, the precision of the NE class (80%) is lower due to tissue morphology differences between the luteal and follicular phases [1]. In the luteal phase, an increased amount of fluid is present between the tissues which appears as more white space. Since a doctor would report the patient's menstrual phase, reference images for each phase could be selected for SPCN [7] to reduce this variation. The EH class has a precision of 60% due to ill-defined epithelial layers contributing to discrepancies in nuclei extraction, impacting the model's performance.

Class	Precision	Recall	F1-score
NE	0.80	0.80	0.80
EP	1.00	0.94	0.97
EH	0.61	0.69	0.65
EA	0.81	0.77	0.79

Table 1. Precision, recall, and F1-score for each class

8. CONCLUSION

The pipeline demonstrated in this study presents a novel approach to creating a human-interpretable diagnostic tool. This model uses a Neural Network to perform semantic segmentation, allowing region-specific features to be extracted and passed to a simple classifier. This approach is advantageous compared to deep learning classification approaches as it makes white-box decisions and does not cut images and features to tiles [5]. The model was tested on 240 images, and achieved an accuracy of 80%, with low F1 scores for the EH class (1) which is comparable to the result (76.91%) achieved by the state-of-the-art deep learning tool HIENet on the full 3302 image dataset [4]. This is attributed to the limited power of the segmentation model used, along with the quality of the dataset. To improve this work, training a state-of-the-art segmentation network such as U-Net, obtaining higher-quality image samples, or incorporating the clustering objective to the autoencoder backpropagation could all be explored.

References

- [1] R. J. Kurman, *Blaustein's pathology of the female genital tract*. Springer Science & Business Media, 2013.
- [2] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012," *International journal of cancer*, vol. 136, no. 5, pp. E359–E386, 2015.
- [3] C. Xu, Y. Li, P. Chen, M. Pan, and X. Bu, "A survey on the attitudes of chinese medical students towards current pathology education," *BMC medical education*, vol. 20, no. 1, pp. 1–7, 2020.
- [4] H. Sun, X. Zeng, T. Xu, G. Peng, and Y. Ma, "Computer-aided diagnosis in histopathological images of the endometrium using a convolutional neural network and attention mechanisms," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 6, pp. 1664–1676, 2020.
- [5] F. Zhao, D. Dong, H. Du, Y. Guo, X. Su, Z. Wang, X. Xie, M. Wang, H. Zhang, X. Cao *et al.*, "Diagnosis of endometrium hyperplasia and screening of endometrial intraepithelial neoplasia in histopathological images using a global-to-local multi-scale convolutional neural network," *Computer Methods and Programs in Biomedicine*, vol. 221, p. 106906, 2022.
- [6] X. Zeng, H. Sun, and Y. Ma, "A histopathological image dataset for endometrial disease diagnosis," 11 2018. [Online]. Available: https://figshare.com/articles/dataset/A_histopathological_image_dataset_for_endometrial_disease_diagnosis/7306361
- [7] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE transactions on medical imaging*, vol. 35, no. 8, pp. 1962–1971, 2016.
- [8] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.
- [9] P. Dahal, "Learning embedding space for clustering from deep representations," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 3747–3755.
- [10] E. Bisong and E. Bisong, "Logistic regression," *Building machine learning and deep learning models on google cloud platform: A comprehensive guide for beginners*, pp. 243–250, 2019.
- [11] a. Vahadane, "Code release/color normalization: This is the matlab code for stain separation and color normalization in computational pathology," Mar 2016. [Online]. Available: https://github.com/abhishekvahadane/CodeRelease_ColorNormalization.git
- [12] J. Mairal, "Spams (sparse modeling software) optimization toolbox," Feb 2017. [Online]. Available: <http://thoth.inrialpes.fr/people/mairal/spams/>
- [13] B. Noble and J. W. Daniel, "Applied linear algebra. 2nd ed," Englewood Cliffs, New Jersey: Prentice-Hall, Inc. XVII, 477 p. \$ 11.85 (1977),, 1977.
- [14] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," 2004.
- [15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [16] M. K. Gupta and P. Chandra, "A comparative study of clustering algorithms," in *2019 6th International Conference on Computing for Sustainable Global Development (INDIACOM)*, 2019, pp. 801–805.
- [17] A. F. Agarap, "Deep learning using rectified linear units (relu)," *CoRR*, vol. abs/1803.08375, 2018. [Online]. Available: <http://arxiv.org/abs/1803.08375>
- [18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.
- [19] K. Alsabti, S. Ranka, and V. Singh, "An efficient k-means clustering algorithm," 1997.
- [20] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.
- [21] "Sklearn.linear-model.logisticregression." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html