Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2019 May ; 2019: . doi:10.1109/bhi.2019.8834632.

Improved Prediction on Heart Transplant Rejection Using Convolutional Autoencoder and Multiple Instance Learning on Whole-Slide Imaging

Yuanda Zhu,

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

May D. Wang,

Dept. of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA

Li Tong,

Dept. of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA

Shriprasad R. Deshpande

Pediatric Cardiology, Children's National Health System, Washington, DC, USA

Abstract

Heart transplant rejection is one major threat for the survival of patients with a heart transplant. Endomyocardial biopsies are effective in showing signs of heart transplant rejection even before patients have any symptoms. Manually examining the tissue samples is costly, time-consuming and error-prone. With recent advances in deep learning (DL) based image processing methods, automatic training and prediction on heart transplant rejection using whole-slide images expect to be promising. This paper develops an advanced pipeline for quality control, feature extraction, clustering and classification. We first implement a stacked convolutional autoencoder to extract feature maps for each tile; we then incorporate multiple instance learning (MIL) with dimensionality reduction and unsupervised clustering prior to classification. Our results show that utilizing unsupervised clustering after feature extraction can achieve higher classification results while preserving the capability for multi-class classification.

Keywords

heart transplant rejection; pathological whole-slide imaging; stacked convolutional autoencoder; multiple instance learning; weakly-supervised learning

yzhu94@gatech.edu.

I. INTRODUCTION

Cardiac allograft rejection is the primary factor limiting long-term survival for patients with heart transplant [1]. In order to prepare for a treatment plan after the surgery, early detection is required to identify rejection types and grades. One common rejection type is acute cellular rejection (ACR) [2]. Usually, this happens in the first three to six months when part of the immune system, T-cells, attacks the cells of the new heart. The other non-chronicle rejection type is humoral rejection, also known as acute antibody-mediated rejection. This type of rejection occurs since antibodies may injure the blood vessels of patients' body. Each type of rejection has three grades, based on the level of rejection. Identifying the type and grade of rejection is essential in diagnosis.

Heart biopsies can effectively detect heart transplant rejection before the patients have any symptoms [3]. Manually examining pathological whole-slide images (WSI) by the expert pathologist is time-consuming, costly and error-prone while computer-aided diagnosis using whole-slide imaging has proven its usefulness in the prediction of heart transplant rejection.

Feature extraction is one major focus of pathological image processing. Spatially localized features of color, shape, and texture are widely utilized for non-DL based feature extraction on pathological whole-slide images. Barker et al. [4] extracted a total of 227 color and shape-based features from brain tumor pathological WSI tiles prior to WSI-level classification using the elastic net and weighted voting. Tong et al. [5] extracted 461 features such as nuclear density and grey level co-occurrence matrix (GLCM) from heart transplant WSIs before merging and classifying these features using deep neural networks with dropout. Dooley et al. [6] pointed out that additional object-level and pixel-level features could be extracted from the heart transplant WSIs. These features include the locations of foci of infiltrates and percent of monocyte inflammation, which specifically describe the different types and grades of heart transplant rejection. However, understanding these features could be theoretically challenging and extracting them could be time-consuming. Some of these features can only be applied to pathological WSIs of a certain disease.

Meanwhile, the DL based, convolutional autoencoder (CAE) has also been used for feature extraction. Makhzani and Frey proposed the convolutional winner-take-all autoencoder in [7] for feature representation. By using a non-symmetric, hierarchical autoencoder whose encoder consists of several convolutional layers and decoder consists of a shallow, large-size and linear deconvolutional layer, both lifetime (temporal) sparsity and spatial sparsity can be learned for each feature map. Competitive classification performance was achieved on MNIST, CIFAR-10, and ImageNet. Zerhouni et al proposed DictiOnary Learning Convolutional autoEncoder (DOLCE) [8] to generate a signature representative of pathological images. The architecture of CAE is similar to that in [7], consisting of a multi-layer encoder and a one-layer decoder. By soft assigning each tile to a set of dictionary elements, this deep-learning based feature extraction method outperformed state-of-art results on tumor imaging classification.

On the other hand, since tile-level labels are not provided, we cannot directly perform classification on each tile. With only image-level labels, we categorize the problem into the

second type of weakly-supervised learning that only coarse-grained labels are given [9]. Thus, we need to incorporate multiple instance learning before classification. Durand et al. [10] proposed the WELDON model to aggregate the extracted tile-level features (local feature descriptors) into image-level features (global feature descriptors) and selected top instances and negative evidence for the classifier. Courtiol et al. [11] modified the WELDON model into CHOWDER by applying one-dimensional convolution on local feature descriptors to generate global feature descriptor. By selecting the top instances and negative evidence from the global feature vector, CHOWDER is successful in binary classification through end-to-end training, yet its capability of multi-class classification was not proven.

In this project, we proposed an automated pipeline (shown in Figure 2) to perform WSI-level weakly-supervised classification after extracting features from unlabeled tiles. Even on a small sample size of labeled WSIs, through using stacked CAE for feature extraction and unsupervised clustering prior to classification, this work increased the classification accuracy. Besides, this pipeline can be extended to multi-class classification.

The rest of paper is organized as follows: in Section II, we briefly introduce our data set. In Section III, we explain the details of our quality control, stacked CAE, clustering and classification as our methodology. In Section IV, we show the experimental results. In Section V, we discuss the results from the previous section, the advantage of our pipeline and possible improvements in future work.

II. DATASET

Children's Healthcare of Atlanta (CHOA) provided us with 43 digitized whole-slide images from endomyocardial biopsies stained with hematoxylin and eosin (H&E) obtained from pediatric heart transplant patients. These whole-slide images are broken into smaller tiles of the size of 512 × 512 pixels. Each whole-slide image has 20,000 to 80,000 tiles.

Annotation grades of antibody-mediated rejection (AMR) and acute cellular rejection (ACR) were implemented by an expert cardiologist for each whole-slide image. Out of 43 whole-slide images, 18 were assigned to no rejection class; for the remaining 25 whole-slide images that rejection presents, both AMR and ACR present in 14 biopsies, only AMR presents in 9 biopsies and only ACR presents in 2 biopsies. We decided to perform binary classification, rejection vs. no rejection, other than multi-class rejection grade classification due to the limited size of dataset.

III. METHODS

A. Quality Control and Color Normalization

Our quality control consists of an empty tile check step and a color normalization step (see Figure 3). The empty tile check is introduced before applying color normalization. A tile will be eliminated from further analysis if its mean and deviation of each color channel fail to reach a certain threshold. This step is to remove "all white background" empty tiles that do not capture any tissues.

Then we adopted “VB-Reinhard-Weighted” color normalization proposed by Magee et al. [12]. The core algorithm is the Variational Bayesian Gaussian Mixture (VBGM) model for pixel-level color segmentation. The variational (EM-like) estimation process was implemented to maximize the posterior probability that each pixel of the tile belongs to one of the three components (two stains and background). Figure 4 shows color normalization results.

B. Convolutional Autoencoder (CAE)

Fully connected autoencoder ignores the 2D image structure, introducing redundancy in the parameters and causing features to be global. Convolutional autoencoder, however, preserves spatial locality through sharing weights among all locations of input [13].

Our stacked CAE architecture shows in Figure 5: the encoder consists of four sets of 2D convolution layer, activation layer (“Relu”) and max-pooling layer; the decoder consists of four sets of 2D convolution layer, activation layer (“Relu”) and up-sampling layer. For each convolution layer, a kernel of size 3×3 is adopted to capture non-linear features of given tiles. The use of max-pooling layers increases the desired nonlinearity in feature extraction. Cross-entropy is used as the loss function when reconstructing each tile; feature map after 50 iterations has the size of 32×32×2.

C. Multiple Instance Learning (MIL)

We incorporated multiple instance learning before classification on the unlabeled tiles. For baseline, we aggregated local features and selected the top instances as well as the negative evidence; for the proposed method, we applied K-means clustering to group feature vectors of all tiles from the same WSI into several clusters and generated a normalized distribution vector.

The pipeline of our baseline is depicted in Figure 1. Similar to WELDON in [10], we aggregated all local feature descriptors of the same tile to be the global features. Then, we adopted the top instances and negative evidence by selecting the largest and smallest R entries from the global feature vector. These 2R top instances and negative evidence are features for our classifier.

Prior to cluster all local feature descriptors, we performed dimensionality reduction to only keep the top ranking features. Here, we used principal component analysis (PCA) to keep the principal components in each tile’s feature vector that explained 95% of variance.

We performed K-means clustering on the reduced feature vectors. We first stacked these local feature vectors from all whole-slide images in the training set. We then calculated the Calinski and Harabaz score [14] to find the optimal number of clusters. This score is the ratio between the within-cluster dispersion and the between-cluster dispersion. A larger score indicates a better clustering performance. Thus, we chose the number of clusters that generated the largest score.

We applied the same K-means clustering model with the chosen number of clusters on the local feature vectors of each whole-slide image. Consequently, all whole-slide images in the

training set and testing set shall have a certain number of tiles (local feature vector) in each cluster. We normalized the number of tiles to the total number of tiles for each whole-slide image to generate a normalized distribution vector, indicating the percentage of tiles in each cluster. These distribution vectors are the features for the classification described in section 3.3.3.

We used a multi-layer perceptron classifier for classification with five-fold cross validation. The four hidden layers have 200, 100, 50 and 25 hidden neurons respectively.

IV. RESULTS

Similar to CHOWDER, we selected 1, 10, and 100 top instances and negative evidence from the local feature descriptors. Baseline tends to perform better on classification with more instances selected. As shown in Table 1, our clustering method has higher mean value and lower standard deviation of classification accuracy and AUC score.

V. DISCUSSION

In [6], 461 hand crafted features led to the highest classification accuracy of 70%. Utilizing convolutional autoencoder for feature extraction gave rise to an accuracy of 72.2%. The improvement is not significant, largely due to the small data size we have. Stacked CAE, nevertheless, has shown its effectiveness in feature extraction, achieving higher accuracy with carefully designed, hand-crafted object-level and pixel-level features extraction. Consequently, this feature extraction process is more scalable to much larger data size, and more robust on different WSI datasets.

Meanwhile, the proposed clustering method achieve higher accuracy and AUC score than baseline, which was inspired by the state-of-art WELDON and CHOWDER model. Besides, the proposed clustering method preserved the capability of multi-class classification.

Improvement can be made through adopting a pre-trained deep convolutional neural network for feature extraction. Furthermore, K-means clustering can be replaced with a deep clustering algorithm. In this way, we can perform end-to-end training across the entire pipeline.

The limitations of this study are the assignment of rejection grade to the whole slide and the sample size. We anticipate the performance of the method to significantly improve with larger sample size and further refinement of the classification. We anticipate to validate the study using a prospectively acquired robust data set and to test the performance at finer granularity rather than the binary classification used here.

ACKNOWLEDGMENT

This work was supported by the grants from National Institutes of Health (NCI Transformative R01 CA163256, and National Center for Advancing Translational Sciences UL1TR000454), National Science Foundation Award NSF1651360, and CHOA-Georgia Tech collaboration grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, NSF or CHOA.

References

[1]. Usta E, Burgstahler C, Aebert H, Schroeder S, Helber U, Kopp AF and Ziemer G, “The challenge to detect heart transplant rejection and transplant vasculopathy non-invasively - a pilot study,” Journal of Cardiothoracic Surgery, vol. 4, p. 43–2009. [PubMed: 19682394]

[2]. Patel JK, Kittleson M, Kobashigawa JA, “Cardiac allograft rejection,” The Surgeon, vol. 9, no. 3, pp. 160–167, 2011. [PubMed: 21550522]

[3]. Stewart S, Winters GL, Fishbein et el MC, “Revision of the 1990 Working Formulation for the Standardization of Nomenclature in the Diagnosis of Heart Rejection,” The Journal of Heart and Lung Transplantation, vol. 24, no. 11, pp. 1710–1720, 2005. [PubMed: 16297770]

[4]. Barker J, Hoogi A, Depeursinge A, Rubin DL, “Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles,” Medical Image Analysis, vol. 30, 2016.

[5]. Tong L, Hoffman R, Deshpande SR and Wang MD, “Predicting heart rejection using histopathological whole-slide imaging and deep neural network with dropout,” in IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2017.

[6]. Dooley AE, Tong L, Deshpande SR and Wang MD, “Prediction of heart transplant rejection using histopathological whole-slide imaging,” in IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2018.

[7]. Makhzani A and Frey B, “Winner-Take-All Autoencoders,” in Advances in Neural Information Processing Systems, 2015.

[8]. Zerhouni E, Prisacari B, Zhong Q, Wild P and Gabrani M, “Disease grading of heterogeneous tissue using convolutional autoencoder,” in IEEE International Symposium on Biomedical Imaging (ISBI), 2017.

[9]. Zhou ZH, “A brief introduction to weakly supervised learning,” National Science Review, vol. 5, no. 1, pp. 44–53, 2017.

[10]. Durand T, Thome N and Cord M, “WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks,” in The IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[11]. Courtiol P, Tramel EW, Sanselme M and Wainrib G, “Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach,” in Computer Vision and Pattern Recognition, Salt Lake City, 2018.

[12]. Magee D, Treanor D, Crellin D, Shires M, Smith K, Mohee K, and Quirke P, “Colour normalisation in digital histopathology images,” Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy, vol. 100, 2009.

[13]. Masci J, Meier U, Cireşan D, Schmidhuber J, “Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction,” in Artificial Neural Networks and Machine Learning – ICANN, 2011.

[14]. Calinski T, Harabasz J, “A dendrite method for cluster analysis,” Communications in Statistics, 1974.

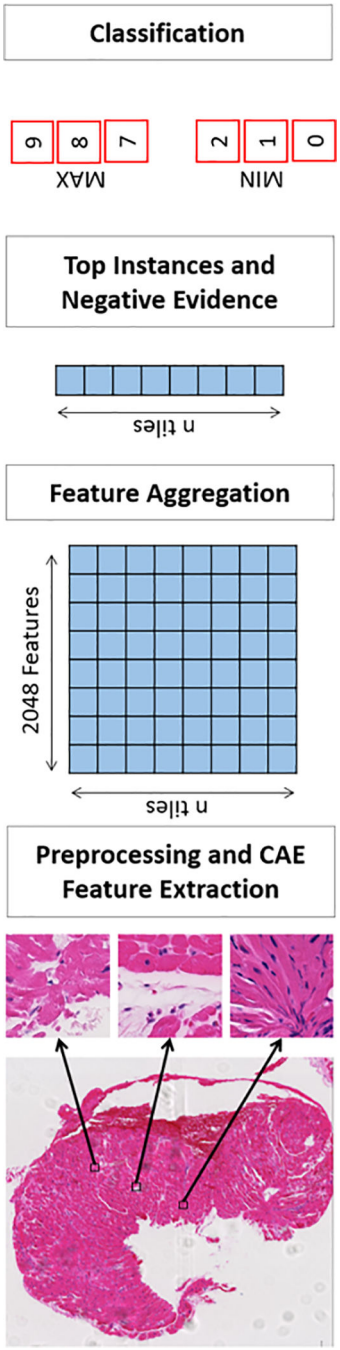

Fig. 1.
Pipeline for baseline approach through feature aggregation and top instance selection.

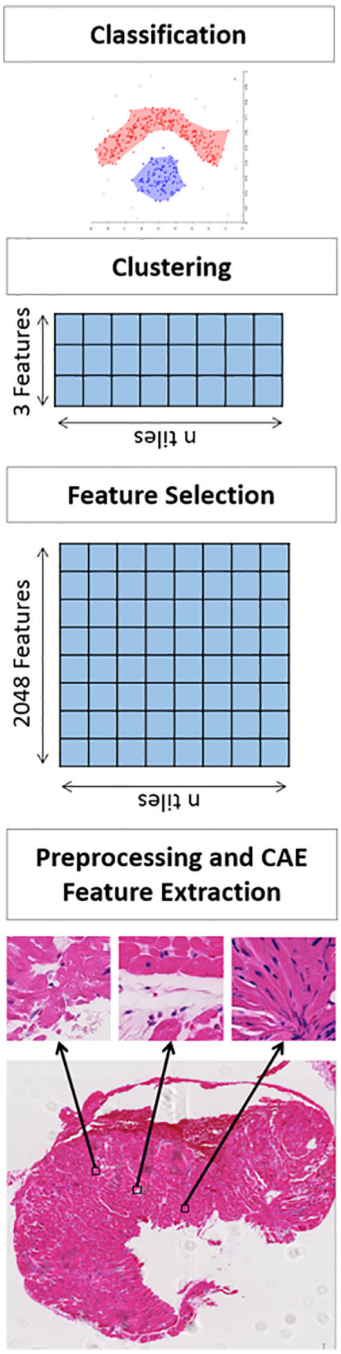

Fig. 2.
Pipeline for proposed approach through feature selection and clustering.

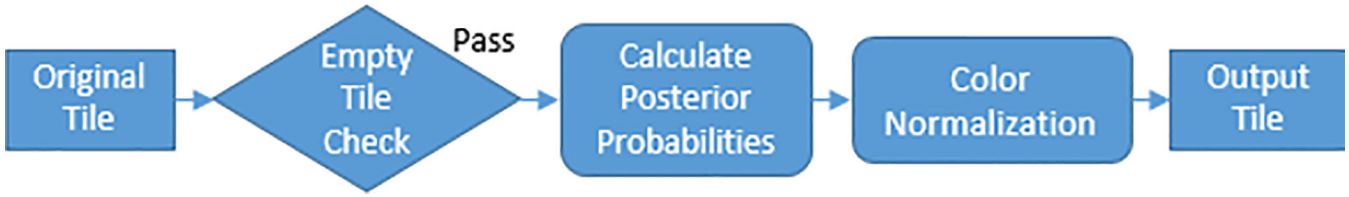

Fig 3.
Preprocessing, including empty tile check and color and normalization.

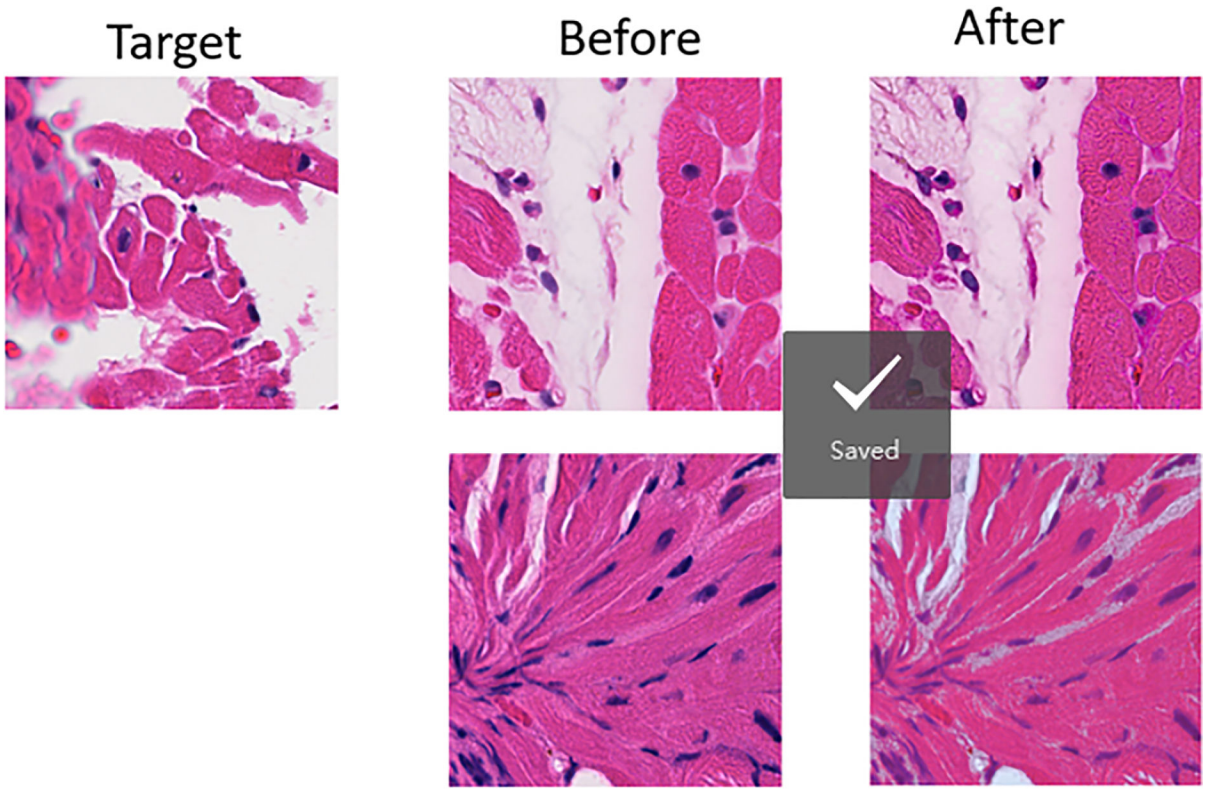

Fig. 4.
Color normalization results using “VB-Reinhard-Weighted” method.

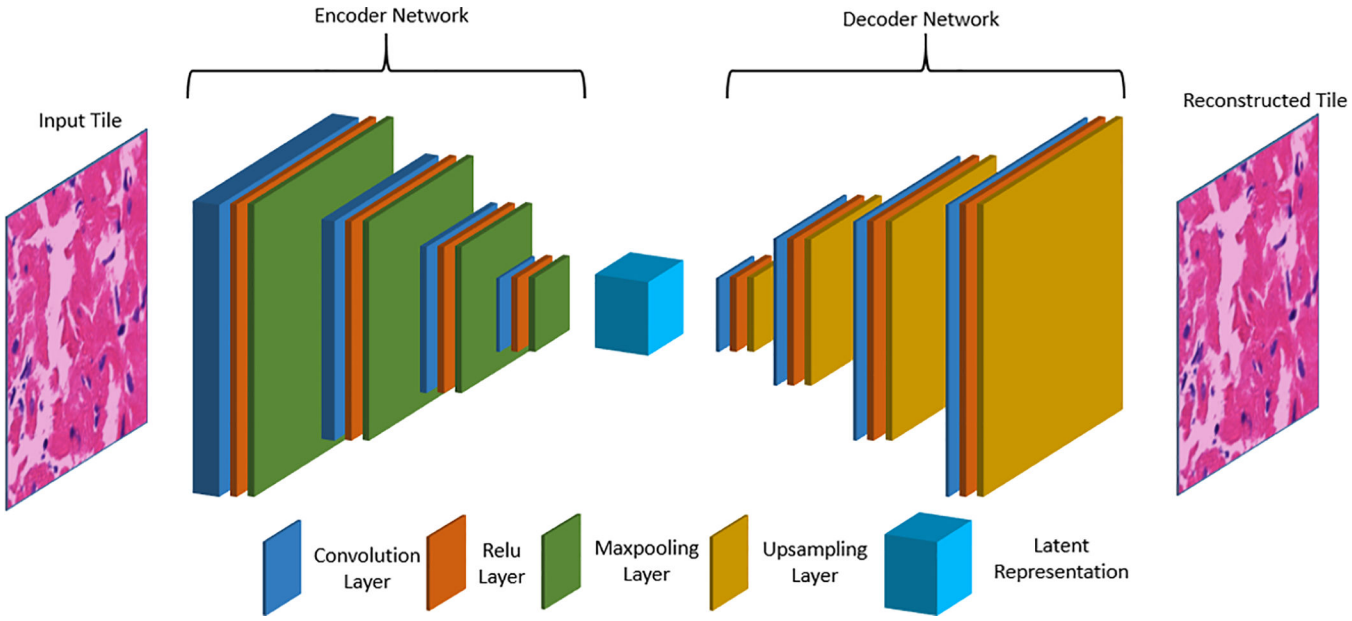

Fig. 5.

Architecture of Convolutional Autoencoder. Convolutional layers have kernel size of (3,3), and filter size 16, 8, 4, 2, and 2, 4, 8 and 16 respectively.

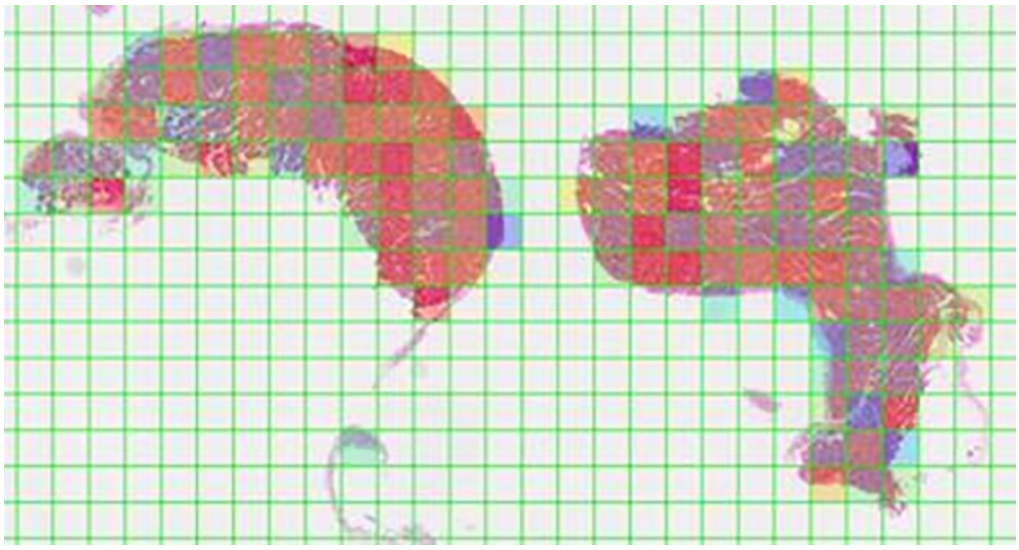

Fig. 6.
Visualization of clustering on titles.

Table 1.

Classification results for baseline and proposed method. For both accuracy and AUC, average and standard deviation values are calculated.

Method	# Instances	Accuracy	AUC
Baseline	1	0.578 ±0.130	0.508 ±0.089
	10	0.608 ±0.130	0.558 ±0.126
	100	0.697 ±0.114	0.659 ±0.137
Clustering Method		**0.722 ±0.115**	**0.713 ±0.109**