An Ensemble-based Active Learning for Breast Cancer Classification

Sanghoon Lee
Weisberg Division of Computer Science
Marshall University
Huntington, USA
leesan@marshall.edu

Mohamed Amgad
Department of Biomedical Informatics
Emory University
Atlanta, USA
mtageld@emory.edu

Mohamed Masoud
Department of Neurology
Emory University
Atlanta, USA
mohamed.masoud@emory.edu

Rajasekaran Subramanian
Department of Neurology
Emory University
Atlanta, USA
rajasekaran.subramanian@emory.edu

David Gutman
Department of Neurology
Emory University
Atlanta, USA
dgutman@emory.edu

Lee Cooper
Department of Pathology
Northwestern University
Chicago, USA
lee.cooper@northwestem. edu

***Abstract*—Ensembled machine learning paradigms enable base learners to provide more accurate predictions than a standard approach using a single learner. Though the ensemble learning decreases variance or bias, improving predictions, limited literatures have been reported with an active learning strategy narrowing uncertainty in prediction. We present an ensemble based active learning approach for breast cancer detection, averaging predictions from the start of the art machine learning models on histopathology images. We demonstrate that the ensemble based active learning approach outperforms other approaches on breast cancer detection.**

Keywords—active learning, ensemble learning, machine learning, whole slide images, breast cancer

I. Introduction

Notable advances in whole-slide imaging technology have been made in the field of histopathology image analysis over the past decade. Microscope slide scanners enable reliable and high-resolution images of tissue samples in a few minutes and specialized software potentially aid biological scientists in interpreting histological images. Detecting cancers such as breast cancer, lung cancer, and prostate cancer, however, has been a challenging in delivering an accurate cancer detection in the histopathological interpretation of whole-slide images [1]-

Active learning is a well-known learning technique that aims to improve the learning performance in a target domain through training and querying datasets [2]. When it comes to learning a new model, few labeled data are required to train the model in the beginning. The trained model is then used for predicting unlabeled data providing a new suggestion through a certain type of queries to be corrected at the beginning again. Active learning enhances the learning performance by iteratively correcting the data, thereby increasing the accuracy. Major components of the active learning are shown in the Fig. 1.

Many literatures have been reported on the use of an active learning in histopathology image analysis. Ibrahim et al. demonstrated that the active learning process increases the classification accuracy by selecting the discriminative genes based on some types of cancers such as prostate cancer, colon cancer, and ovarian cancers [3]. Nalisnik et al. presented a web-based machine learning framework that uses the active learning to enhance the brain tumor identification [4]. However, these approaches have been limited to a machine learning algorithm chosen by their empirical research without the consideration of other algorithms. Stacking models from various algorithms has been recently received a quite attention on the machine learning communities. A typical challenge here is to increase the accuracy by combining predictions from the models.

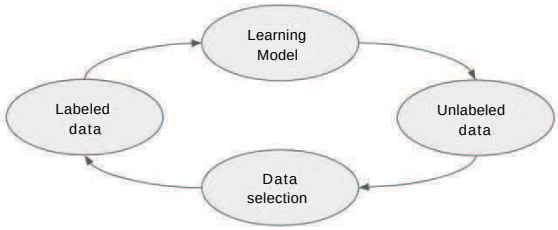

Fig. 1 Major components of Active learning

In this paper, we present an ensemble based active learning approach to classify breast cancer, stroma, and lymphocytes on histopathology images, generalizing multiple state of the art machine learning models. The advantages of the proposed method are as below. First of all, the ensemble based active learning approach is unbiased to a certain type of machine learning model so that this approach can avoid duplicated tasks for optimizing hyper-parameters. Second, the benefits of the active learning can be obtained through the entire process of the learning; the number of training samples is significantly reduced during the active learning process while increasing the accuracy on the prediction.

II. Ensemble Learning on Histopathology Images

Ensemble learning is a well-established machine learning strategy solving a problem by training multiple learners [5-6]. While traditional machine learning models use one base learner, ensemble learning uses multiple individual learners to improve prediction performance. In this section, 6 base learners generated from randomized training data are briefly described and the pseudo-code of the ensemble learning method used in this paper is shown Fig. 2.

A. Support Vector Machine

Support Vector Machine (SVM) is a machine learning algorithm based on the statistic learning theory. The idea of

XXX-X-XXXX-XXXX-X/XX/$XX.OO ©20XX IEEE
978-1-7281-1867-3/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on November 16,2022 at 08:49:41 UTC from IEEE Xplore. Restrictions apply.

the SVM is producing an optimal hyperplane to categorize test data given a set of labeled training data. SVM has been widely used to detect cancers in the field of histopathology image analysis [7-8].

B. *K-nearest neighbors*

K-nearest neighbors (KNN) is a non-parametric classification algorithm that takes into account the similarities of the elements in a dataset. In KNN, n-dimensional feature space is defined by the input parameters and new test data are classified based on the majority of k-nearest neighbors [9-10].

C. *Naive Bayes*

Naive Bayes (NB) is a probability-based prediction algorithm that employs Bayes theorem assuming that all the features in a dataset are conditionally independent. NB maximizes the posterior probability to determine the new test dataset given a set of labels training dataset [11-12].

D. *Random Forest*

Random Forest (RF) is a well-known machine learning algorithm that combines multiple decision trees to improve the predictive accuracy and avoid an overfitting problem. Each tree is randomly built by a new training samples and averaged for a final prediction [13].

E. *Gradient Boosingting*

Gradient Boosting (GB) is a scalable machine learning algorithm that utilizes gradient-based optimization and boosting. GB updates parameters by minimizing a loss function and constructs a model from a set of decision trees used as weak prediction models [14-15].

F. *Logistic Regressioin*

Logistic Regression (LR) is a statistical machine learning method that classifies a categorical variable. LR builds a model discriminating between samples by estimating probabilities of a certain class [16-17].

Input : Data set D = {(i , ?/i), (22,3/2), •••, (z_p , y_p)}
Base learner £
Number of base learners *B*

Process :
for *b* = 1 to *B* :
hb — *L(T>b)11*Train a base learner *hb*
end

Output : *H(x)* = I 'y *hb(.x')* **I/-S**
\ 6=1 /'

Fig. 2 An ensemble learning algorithm

III. An Active Learning Framweork on Breast Cancer Detection using an Ensemble Learning

A. *Whole slide image preprocessing*

Whole slide images are large, multi-resolution and multi-layered images created by digitizing glass slides and used for biomedical research or clinical diagnosis. Easy layer on the whole slide images represents different resolutions and the typical size of the whole slide images can vary from 2 to 8 GB at 20x magnification or 40x magnification.

We used HistomicsTK[18] to conduct the preprocessing of the whole slide images. In order to process the whole slide images, it is necessary to decompose a whole slide image into small tiles so that avoiding an issue with limited hardware capacity. Tiling of the whole slide image is a fundamental approach for parallelization of whole slide image analysis, avoiding computation burden on processing the entire image. After then, features from the tile images can be extracted in pre-processing steps. In this paper, however, we use the region of interests (ROIs) extracted from the whole slide images to verify the effectiveness of the ensemble-based active learning rather than using the whole slide images. We leave the work of adopting the ensemble based active learning to the whole slide images for the future work. In addition to the tilting, a color normalization method for histopathology images was performed to adjust the intensities of the images since the color normalization is a practical approach to improve the color appearance. Thus, for the histopathology images, Reinhard color normalization is used and then 128x128 sized image patches are segmented to be used as both training and test objects to demonstrate the effectiveness of the ensemble based active learning approach.

B. *Ensemble based Active learning*

Active learning performs well on a very small number of labeled samples in the beginning because samples are chosen from the prediction results by actively participating in the collection of the training samples. Reducing the number of training samples but enhancing the performance of the machine learning models. Active learning has been widely used for obtaining the optimized prediction results with a few training samples.

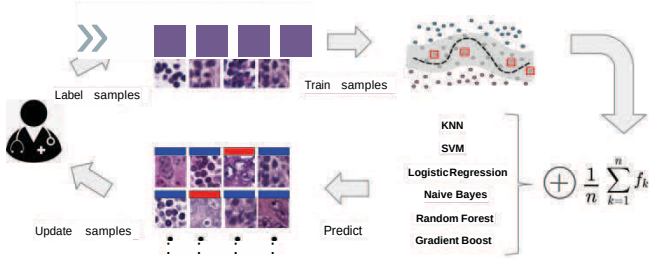

Fig. 3 Active learning framework using an ensemble learning

The key idea of the active learning is enabling users to iteratively select training samples to classify an unknown decision boundary, correcting prediction errors improving the accuracy. In Fig. 3, the iterative process of the active learning framework using an ensemble learning is presented. 128x128 sized image patch is shown as an example. A blue color bar over each image patch indicates breast cancer and a red color bar indicates lymphocytes. 6 different machine learning models are used for ensemble learning. The average of the predictions is computed to predict all remaining samples. In this stage, machine learning models trained by labeled samples predict unlabeled samples and the prediction probabilities are averaged over the entire samples. Next, the predicted samples after the predicting stage are considered as next samples to be trained. Correcting the predicted samples through the iterative process increases the accuracy of cancer

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on November 16,2022 at 08:49:41 UTC from IEEE Xplore. Restrictions apply.

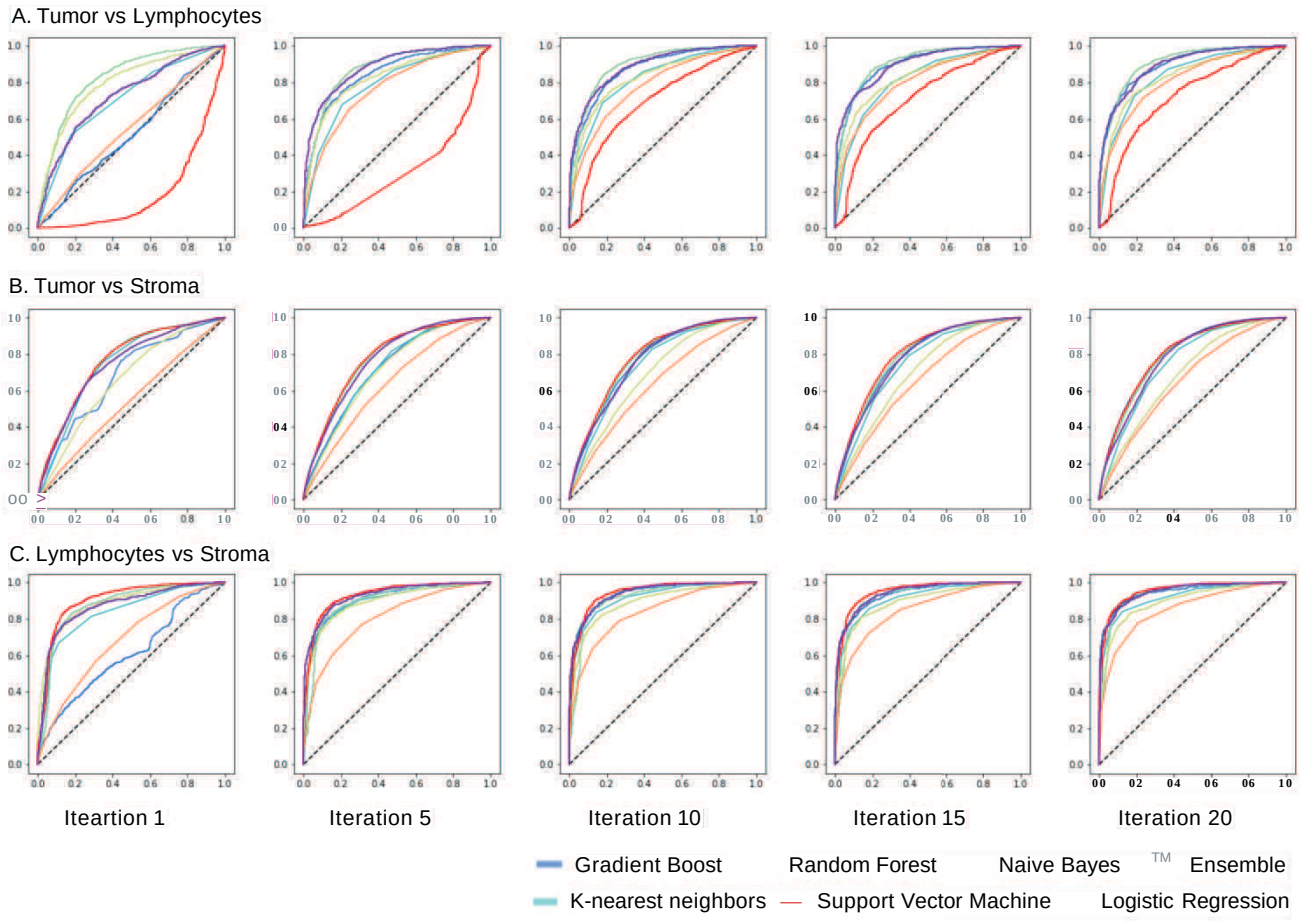

Fig. 3 Performance results of breast cancer, lymphocytes, and stroma without active learning

detection. Our paper aims to increase the accuracy by averaging of the different machine learning models. This iterative process is conducted by correcting sample labels

Input : Data set = $\{(x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p)\}$

Base learner £

Number of base learners B

Process :

do

$2?' = \{(x_1^t, y_1^t), (x_2^t, y_2^t), \ldots,$ **(a:*,y*)}**

$T >= T >— D^*$ LtRemove training samples

for b — 1 to B :

$h_b = \mathcal{L}(\mathcal{D}_b{}^t)$

end

until iteration is done or |7? —P' | <= 0

Output : $H(x') = \mathbf{I}\ '\ / hb(x)\ \mathbf{j}\ /B$

Fig. 2 Ensemble based active learning algorithm

updating all models. As shown in Fig. 4, an ensemble learning is adopted in the active learning process. Training samples selected from whole dataset are removed from the dataset and all baser learners are trained using the training samples. New training samples are stacked on the previous training samples and removed from the whole dataset until the iterative process is done or no samples in the dataset is remained to be removed.

IV. Experimental Results

In order to demonstrate the effectiveness of the proposed ensemble based active learning, two experiments were conducted under three different binary classification problems: breast cancers versus lymphocytes, breast cancers versus stroma, and stroma versus and lymphocytes. 7 machine learning methods including an ensemble learning are compared with each other in the first experiment, while 7 machine learning methods are compared with each other in the second experiment using an active learning. Total 63 annotated images are extracted from the whole slide images and each of them includes a set of image patches labeled with positive or negative values to perform the two experiments.

A. Dataset

The original images used for the dataset were obtained from the whole slide images offered by The Cancer Genome Atlas (TCGA, https://tcgadata.nci.nih.gov/tcga/) archive and are annotated by human experts to conduct the experiments. Total 27,397 image patches are used for a dataset and each image patch is considered as positive if it contains 75% of the positive area. For example, for the classification of breast cancer and lymphocytes, an image patch is positive if it occupies 75% of breast cancer region while it is negative if it occupies lymphocytes region.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on November 16,2022 at 08:49:41 UTC from IEEE Xplore. Restrictions apply.

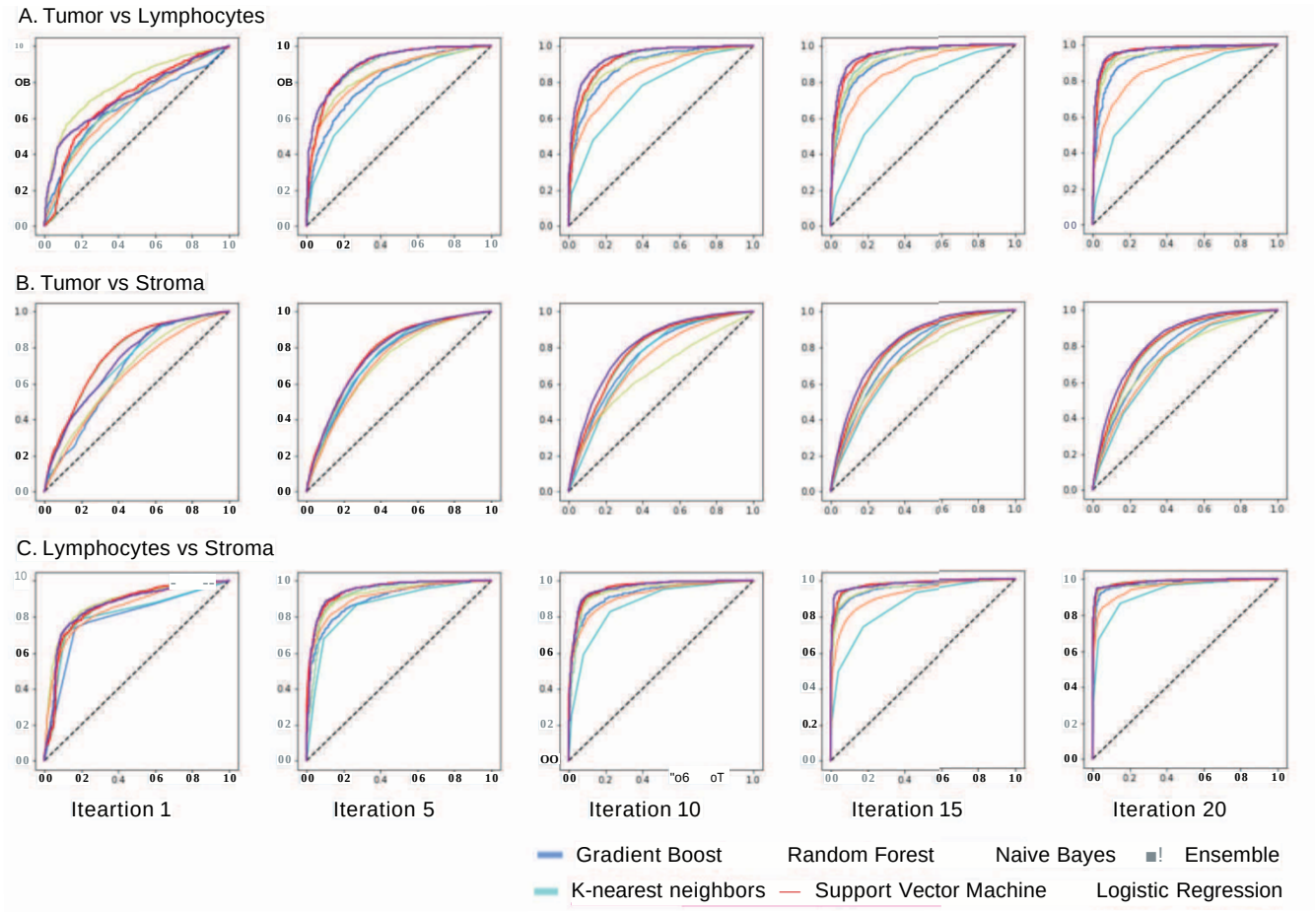

Fig. 4 Performance results of breast cancer, lymphocytes, and stroma with active learning

B. Performance results

The performance results of breast cancer, lymphocytes, and stroma classification without active learning are shown in the Fig. 5. Each row indicates Receiver Operating Characteristics (ROC) curves of breast cancer vs. lymphocytes, breast cancer vs. stroma, and lymphocytes vs. stroma and each column indicates the number of iterations performed on the models. In each iteration, training samples are randomly selected from the dataset and the predictions are averaged by the ensemble learner. ROC curves are measured based on the patches not the regions by comparing the traditional machine learning models with the ensemble learning. As shown in Fig. 5, the ensemble learning without an active learning does not provide better performance than others. Area Under Curve (AUC) results are shown in Table 1. The experiment shows that a logistic regression model provides the highest AUC result on breast cancer vs. lymphocytes and a gradient boost model provides the highest AUC result on breast cancer vs stroma while a support vector machine model provides the highest AUC results on stroma vs. lymphocytes. An ensemble learning without an active learning does not provide any significant results on the breast cancer detection. Performance results of breast cancer, lymphocytes, and stroma with active learning is shown in the Fig. 6. While the first experiment selects training samples randomly for each iteration, the second experiment augments training samples based on the recommended samples through the active learning. In each iteration, a set of samples which are the most uncertain is considered as the next training samples, thereby increasing the accuracy in prediction on the active learning. As shown in the Fig. 6, ROC curves with the active learning are significantly increased and the ensemble learner outperforms other machine learners on all the classification results.

V. Conclusion and Future Works

An ensemble-based active learning method is presented on a breast cancer classification problem. An active learning narrowing uncertainty in prediction is adopted during the classification and an ensemble learning averaging the predictions from the base learners is used for the final prediction. Two experimental results show that the active learning with the ensemble learning increases significantly the performance of the classification results. The ensemble-based active learning method can be applied to the whole slide image analysis. For example, the proposed method can be imported in HistomicsML[4] framework to improve its performance since it has only adopted a random forest algorithm in the framework to the classification problem.

References

[1] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," pp. 170-175, 2016.

[2] B. Settles. ''Active Learning Literature Survey". Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009.

[3] R. Ibrahim, N.A.Yousri, M.A. Ismail, and N.M. El-Makky, "Multi-level gene/MiRNA feature selection using deep belief nets and active learning," In *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3957-3960. 2014

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on November 16,2022 at 08:49:41 UTC from IEEE Xplore. Restrictions apply.

TABLE 1. CLASSIFICATION RESULTS OF BREAST CANCER, LYMPHOCYTES, AND STROMA

Models	Non-Active learning			Active learning		
	Tumor vs lymphocytes	Tumor vs stroma	Stroma vs lymphocytes	Tumor vs lymphocytes	Tumor vs stroma	Stroma vs lymphocytes
Gradient Boost	0.894	**0.800**	0.946	0.933	0.772	0.970
K-nearest neighbors	0.834	0.754	0.915	0.775	0.722	0.925
Logistic Regression	**0.906**	0.789	0.937	0.960	0.804	**0.986**
Naive Bayes	0.829	0.687	0.896	0.943	0.744	0.975
Random forest	0.809	0.655	0.851	0.870	0.742	0.955
Support Vector Machine	0.716	0.798	**0.955**	0.966	0.811	**0.986**
Ensemble	0.892	0.781	0.950	**0.969**	**0.830**	**0.986**

[4] M. Nalisnik, M. Amgad, S. Lee, S.H. Halani, J.E.V. Vega, D.J. Brat, D.A. Gutman, L.A. Cooper, "Interactive phenotyping of large-scale histology imaging data with HistomicsML," *Scientific reports*, vol 7, *issue* 1, pp.14588, 2017.

[5] Z.H. Zhou, "Ensemble learning." *Encyclopedia of biometrics*, pp. 411-416,2015.

[6] LI. Kuncheva, CJ. Whitaker, "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy," Machine Learning, vol. 51, pp. 181-207, 2003.

[7] P. Khurd, C. Bahlmann, P. Maday, A. Kamen, S. Gibbs-Strauss, E.M. Genega, J.V. Frangioni, "Computer-aided Gleason grading of prostate cancer histopathological images using texton forests," *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 636-639, 2010.

[8] S. Sahran, D. Albashish, A. Abdullah, N.A. Shukor, S.H.M. Pauzi, " Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading," *Artificial intelligence in medicine*, vol. *87*, pp.78-90,2018.

[9] A. D. Belsare, M. M. Mushrif, M. A. Pangarkar, and N. Meshram. "Classification of breast cancer histopathology images using texture feature analysis." In *TENCON 2015-2015 IEEE Region 10 Conference*, pp. 1-5. IEEE, 2015.

[10] A. Gunawan, I. Wayan Supardi, S. Poniman, and Bagus G. Dharmawan. "The Utilization of Physics Parameter to Classify Histopathology Types of Invasive Ductal Carcinoma (IDC) and Invasive Lobular Carcinoma (ILC) by using K-Nearest Neighbourhood (KNN) Method." *International Journal of Electrical and Computer Engineering* 8, no. 4 pp. 2442, 2018.

[11] W. Jiang, Y. Shen, Y. Ding, C. Ye, Y. Zheng, P. Zhao, L. Liu et al. "A naive Bayes algorithm for tissue origin diagnosis (TOD-Bayes) of synchronous multifocal tumors in the hepatobiliary and pancreatic system." *International journal of cancer* 142, no. 2, pp. 357-368, 2018.

[12] M. Benndorf, J. Neubauer, M. Langer, and E. Kotter. "Bayesian pretest probability estimation for primary malignant bone tumors based on the Surveillance, Epidemiology and End Results Program (SEER) database." *International journal of computer assisted radiology and surgery'* 12, no. 3. pp. 485-491, 2018.

[13] B. E. Bejnordi, G. Litjens, M. Hermsen, N. Karssemeijer, and J. Laak, "A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images." In *Medical Imaging: Digital Pathology*, vol. 9420, p. 94200H, 2015.

[14] D. E. Kuo, M.M. Wei, K. R. Armbrust, J.E. Knickelbein, I.YL Yeung, R.B. Nussenblatt, C.C. Chan, and H.N. Sen. "Gradient boosted decision tree classification of endophthalmitis versus uveitis and lymphoma from aqueous and vitreous IL-6 and IL-10 levels." *Journal of Ocular Pharmacology' and Therapeutics'll!*, no. 4, pp. 319-324, 2017.

[15] S. Bakas, K. Zeng, A. Sotiras, S. Rathore, H. Akbari, B. Gaonkar, M. Rozycki, S. Pati, and C. Davatzikos. "GLISTRboost: combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation," In *BrainLes*, pp. 144-155, 2015.

[16] F. Muratore, L. Boiardi, A, Cavazza, R. Aldigeri, N. Pipitone, G. Restuccia, S. Bellafiore, L. Cimino, and C. Salvarani. "Correlations between histopathological findings and clinical manifestations in biopsy-proven giant cell arteritis," *Journal of autoimmunity', vol. 69*, pp. 94-101,2016.

[17] L. Nguyen, A.B. Tosun, J. L. Fine, D.L. Taylor, and S.C. Chennubhotla. "Architectural patterns for differential diagnosis of proliferative breast lesions from histopathological images," In *IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 152-155,2017.

[18] D.A. Gutman, M. Khalilia, S. Lee, M. Nalisnik, M., Z. Mullen, Beezley, D.R. Chittajallu, D. Manthey, and L.A. Cooper, "The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research," *Cancer research*, vol *77. Issue* 21, pp.e75-e78, 2017.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on November 16,2022 at 08:49:41 UTC from IEEE Xplore. Restrictions apply.