Breast cancer image classification with machine learning and GLRLM in metastasis detection

Chang Li, Jiaqi Wu and Ying Zhou

Abstract: This report is mainly based on the CAMELYON17 Challenge where competitors should use machine learning or deep learning algorithms to automate metastasis detection in whole-slide images(WSI) and produce a patient-level classification about the pN-stage of patients. This research aims to test whether traditional computer vision features like Gray level co-occurrence matrix(GLCM), Gabor filters are useful in metastasis detection. Besides, a variant of GLCM, Gray level run length matrix(GLRLM) was creatively proposed based on proliferation properties of cancer cells. As a result, Color, GLCM, Gabor filters and GLRLM are used as features in Support Vector Machine(SVM) with radial basis function(RBF) kernel. The model achieves 94% accuracy in 64x64 windows classification and with a threshold set at 40%, as high as 99.5% accuracy could be achieved on 512x512 patches classification. After analysis, such good performance could possibly be attributed to the usage of GLRLM.

Keywords: CAMELYON17 Challenge; Metastasis Detection; Support Vector Machine; Gray Level Run Length Matrix; Biomedical Image Analysis

1. INTRODUCTION

Breast cancer is one of the most common and deadly cancers among women worldwide [1]. Although the prognosis of breast cancer patients is generally good, with an average 5-year overall survival rate of 90% and 10-year survival rate of 83%, the prognosis deteriorates significantly when breast cancer metastases [2] . Although the five-year survival rate of local breast cancer is 99%, the case of regional (lymph node) metastasis drops to 85%, and the case of distant metastasis is only 26% [3]. Therefore, it is important to determine whether there is metastasis for adequate treatment and the best chance of survival.

The first step in determining whether there is metastasis is to check the regional lymph nodes. The presence or absence of metastasis in these lymph nodes is not only a poor prognostic factor in itself but also an important predictor of the existence of distant metastasis [4].In breast cancer, the most common way to assess the regional lymph node status is the sentinel lymph node procedure [5, 6].With this procedure, a blue dye and/or radioactive tracer is injected near the tumor. The first lymph node reached by the injected substance, the sentinel node, is most likely to contain the metastasized cancer cells and is excised. Subsequently, it is submitted for histopathological processing and examination by a pathologist.a key challenge for pathologists in assessing lymph node status is the large area of tissue that has to be examined to identify metastases that can be as small as single cells, which is time-consuming, and pathologists may miss small metastases [7].

A retrospective study showed that expert pathological examinations changed the status of lymph nodes in 24% of patients [8].It has been shown that deep learning algorithms can identify transfers in SLN slides with 100% sensitivity, while 40% of slides without transfers can be identified in this way. This may significantly reduce the workload and time consumption of the pathologist [9].Slide imaging is a technique in which a high-speed slide scanner digitizes a slide with a very high resolution (for example, 240 nm per pixel). This results in an image size of approximately 10 gigapixels, and is commonly referred to as a full slide image (WSI). Such a large amount of data makes WSI very suitable for analysis using machine learning algorithms [10].

Therefore, we tried to train SVM to extract features from the data set provided in the CAMELYON17 challenge and use the SVM classifier to distinguish tissues containing cancer cells from normal cell tissues.Preliminary classification of tissues containing cancer cells and tissues with normal cells through the SVM classifier can reduce the time required to analyze a complete pathological section (which may contain millions or more cells), and reduce the pathologists early workload. The probability of fatigue and misdiagnosis. At the same time, the manual classification of breast cancer by pathologists is a subjective and tedious process, and the diagnosis results of different pathologists are also very different[11].

2. RELATED WORK

In the use of CAMELYON17 data, almost all teams have carried out extensive data augmentation to increase the variation in the training set [12] [13] [14]. Data enhancement algorithm has been applied to increase the size of training set, so that large-capacity learners can benefit from more representative training data [15].Rozsa et al [16]. Noisy processing or adding slight disturbance to the original data makes them more challenging in classification or segmentation.Other strategies include randomly cropping patches and applying affine transformations (e.g. scaling) [17].

All participants used a preprocessing step to identify tissue regions in the WSIs. All participants used simple filtering and thresholding algorithms, mostly Otsu's adaptive threshold at a low resolution level [18].Differences between the methods were mainly found in which color space the thresholds were applied, for example RGB (red-greenblue), HSV (hue-saturation-value), or HSI (hue-saturationintensity), and the type of morphological operations that were used to refine the thresholded image.

Subsequently, in order to perform metastasis detection on each slide, all 12 teams trained a convolutional neural network structure (CNN) and extracted image patches from the identified tissue regions (normal and metastatic regions). Then apply the trained network to the test image to obtain the transfer probability map. In the post-processing step, most participants thresholded the likelihood map, collected several features from the identified cancer area, and used a separate classifier (such as random forest) to determine the WSI category: negative, ITC, micro, or Macro [17].

With respect to the different types of algorithms, all participants used CNNs.Specifically, they used variants of common network architectures: ResNet, GoogLeNet/Inception , VGG-Net, U-Net, and one team used DenseNet.

Although, CNN has strong nonlinear fitting ability, can map arbitrarily complex nonlinear relations, and the learning rules are simple, which is convenient for computer implementation. It has strong robustness, memory ability, nonlinear mapping ability and strong self-learning ability. However, CNN cannot explain its reasoning process and reasoning basis. It requires a large amount of sample data, which is easy to fall into the local optimum, and the training result is not stable. SVM have strict theoretical and mathematical foundations. Based on the principle of structural risk minimization, the generalization ability is better than the former, and the algorithm has global optimality. It is a theory for small sample statistics.

3. METHODOLOGY

Our overall process is to first use image segmentation to cut 512*512 patches and perform image enhancement, then use GLCM, Gabor filter and other methods to extract image features and use Principal component analysis to reduce the dimension of data, and finally apply SVM to do the classification.

3.1. Patches Extraction

3.1.1. ***Data description***

The CAMELYON dataset is a combination of the WSIs of sentinel lymph node tissue sections collected for the CAMELYON16 and CAMELYON17 challenges, which contained 399 WSIs and 1,000 WSIs, respectively.Geert et al. [19] have made many contributions to this data set and their research methods. This resulted in 1399 unique WSIs and a total data size of 2.95 terabytes. The dataset is currently publicly available after registration via the CAMELYON17 website [20]. It has been licensed under the Creative Commons CC0 license.We selected some images from C17 data set for research, and enhanced the images. 600 patches containing tumor cells and 800 patches all normal cells were obtained in dataset.

3.1.2. ***Tools for data use***

Several tools are available to visualize and interact with the CAMELYON dataset [19]. Here, dataset developers show examples of how to use the data with an open-source package developed by them, ASAP(Automate Slide Analysis Platform) [21]. We also used other open-source packages, such as OpenSlide [22] to help us cut patches. But OpenSlide does not contain functionality for reading annotations or storing image analysis results.

ASAP contains several components, of which one is a viewer/annotation application. This can be started via the ASAP executable within the installation folder of the package. After opening an image file from the dataset, one can explore the data easily. The provided reference standard can be loaded via the annotation plugin. In addition, new annotations can be made with the annotation tools provided.

The annotations are provided in human-readable XML format and can be parsed using the ASAP package. Annotations are stored as polygons. Each polygon consists of a list of (x, y) coordinates at the highest resolution level of the image.

3.1.3. ***Image segmentation***

The image of the data set consists of many layers, and the size of each pixel in the lowest layer is 0.25 microns. We take a 4-fold enlarged image(Figure 1 left) and make a 512 * 512 pixel slice image data set. Here is our operation flow.We use ASAP software to label a rectangular cell group area of the image. In the left side of Figure 1, the lower rectangle is the area of cancer cell group, and the upper rectangle is the area of normal cell group. ASAP software provides the vertex coordinates of the relevant rectangle and the width and height of

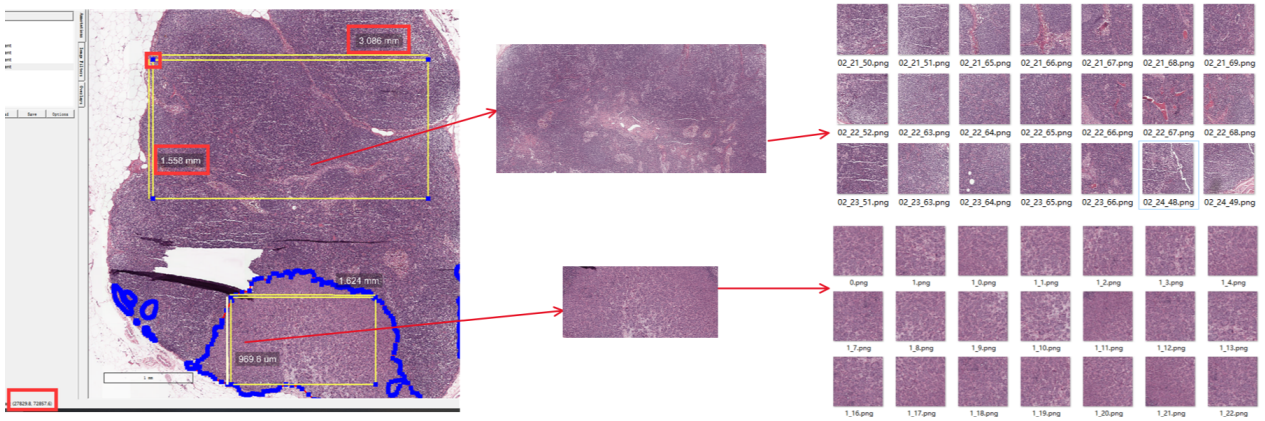

FIGURE 1. Patches extraction process; left: the original image; middle: segmented non-standard normal cell group (top) and cancer cell group (bottom); right: patches set of normal cell group (top) and cancer cell group (bottom) after data enhancement

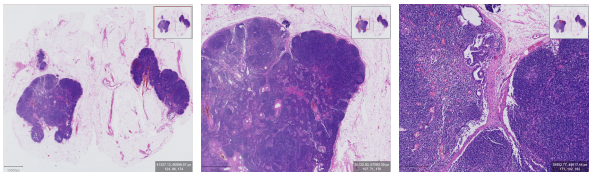

FIGURE 2. Image examples of different downsampling rates of WSI from the CAMELYON17 dataset.

the rectangle, such as the upper left vertex coordinates of the upper rectangle (27829.8, 72857.6), with a width of 3.086 microns and a height of 1.558 microns. Then we use the openslide library to segment the rectangle into a PNG image(Figure 1 middle). We perform the same operation on the original data many times to get some larger images. These images do not meet the requirements of 512 * 512, so we need to enhance them to get a usable data set.

3.1.4. ***Data augmetation***

Data augmetation is also called data amplification, which means that limited data can produce value equivalent to more data without substantial increase of data.

For the pictures we obtained before, we randomly generate X and Y coordinates, and then pick out the 512 * 512 small image under this coordinate as well as doing the data enhancement operations shown as Algorithm 1.

After the data enhancement operations, we get a larger training set: 600 pictures(512 * 512) of cancer cells(Figure 1,right,bottom); 500 * pictures(512 * 512) of normal cells(Figure 1,right,up).

3.2. Feature Extraction

Regarding image feature extraction, we mainly use three methods plus color to collect image features.

Algorithm 1 Framework of Data augmetation.

Require: The original image;

1: Original rotation: 90 degrees, 180 degrees, 270 degrees;
2: The original image is mirrored along the Y axis;
3: Blur the original drawing;
4: Adjust the illumination according to the original drawing;
5: Add noise to the original drawing (Gaussian noise, salt and pepper noise);

Ensure: 512*512 pixel patches;

3.2.1. ***GLCM***

GLCM refers to a common method of describing texture by studying the spatial correlation characteristics of gray levels. In 1973, Haralick et al. proposed a GLCM to describe texture features. Since the texture is formed by the repeated occurrence of grayscale distribution in the spatial position, there will be a certain grayscale relationship between two pixels separated by a certain distance in the image space, that is, the spatial correlation characteristics of the grayscale in the image. We summarise the GLC matrix by six indexes : Contrast, Angular Second Moment, Correlation, Dissimilarity, Homogeneity, Energy, these indicators reflect the uniformity of image gray distribution, texture thickness, visual clarity and other texture features.

3.2.2. ***Gabor Filters***

Gabor filters are used to extract relevant features in different scales and directions of patches in the frequency domain.

Gabor filter is a linear filter with a Gaussian kernel which is modulated by a sinusoidal plane wave. Frequency and orientation representations of the Gabor filter are similar to those of the human visual system. Gabor filter banks are commonly used in computer vision and image processing. They are especially

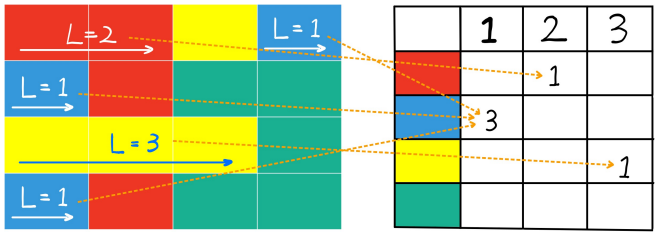

FIGURE 3. illustration of GLRLM

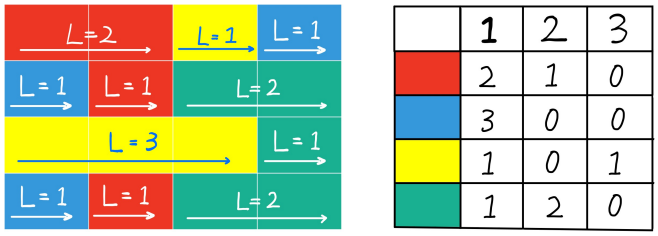

FIGURE 4. Example of a finalized GLRLM in 0 degree

suitable for texture classification.

In practice, a combination of parameters (frequency = 0.8, 1 and θ = 30, 45, 120, 135 degree) are taken for Gabor filters to be applied. Values in each 20 bins of histogram are collected as features. A total of $2x4x20$ = 160 features are extracted.

3.2.3. ***GLRLM***

The GLRLM is a variant of GLCM, which counts the frequency of a sequence of pixels with the same color value in ROI given direction for each color and run length combination. An example is provided in Figure3 about how gray level run length matrix is calculated. For example, there is only one instance for purple color with length 3(last row, last three cells in Figure3), so that 1 is marked in the purple row, column3 of GLRL matrix. The final calculation result of this example patch can be seen in Figure 4.

In Figure 5, cancer cells can be spotted at the top-left and bottom-right areas of the patch example. This patch example inspired us to use GLRLM because from this image we can explicitly see the difference between cancer cells and normal cells at the bottom left. Cancer cells, since they proliferate endlessly, therefore are squeezed together in limited local area. Also, congestion changes their shapes from **circle-like to ovals**. In conclusion, if we can derive the information about **increased cell length or cell congestion** from the image, the classification should work very well. From our perspective, it seems to be the perfect job for GLRLM, And this research aims to test whether GLRLM is useful in metastasis detection.

In practice, a median filter with disk=1 is first applied to the image 50 times to smooth the image. Then Length and direction need to be specified when calculating GLRLM. A combination of parameters(direction equals to 0,45,90,135 degrees and length equals to 2,3,4,5,6,7) are taken. Values in each 26 bin of histogram are collected. A total of $4*6*26 = 624$ features are extracted.

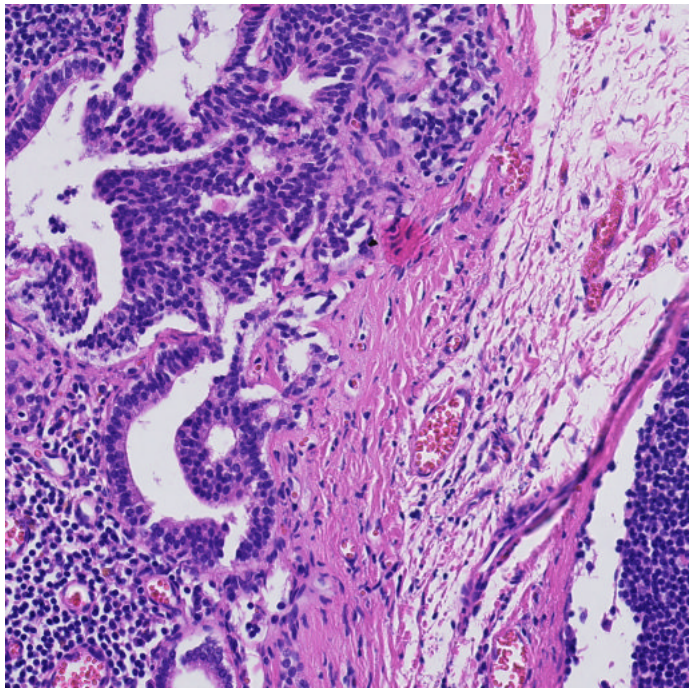

FIGURE 5. Patch example

	Precision	Recall	f1-score	Count
Normal cell	0.93	0.93	0.93	6440
Cancer cell	0.95	0.94	0.94	7660
Total				14100

TABLE 1. Window-level Classification Performance

4. RESULT & ANALYSIS

4.1. Result

Ahead of model training and testing, we extracted around 600 patches(512*512) whose most areas contain metastasis and 500 patches(64*64) of the same size containing mostly normal cells. Then, each patch is split into 64 64x64 small windows for later classification. Although not all of the windows purely contain normal cells or metastasis, they share the same label as the patches they are extracted from. Then, Color (8), GLCM (8), Gabor filters (160) and GLRLM (624) around 800 features are extracted from each small window and stored in a CSV file. Finally, data are split into training data and testing data, and a model is built. Cross-validation is also performed to select the best parameters. In the end, SVM with RBF kernel and C=20 outperforms. The result is shown in Table 1.

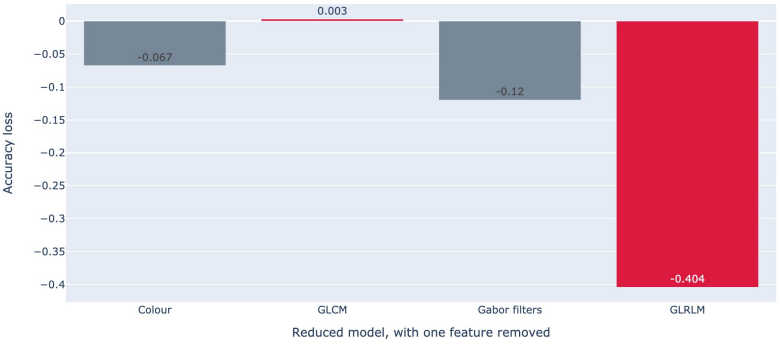

FIGURE 6. reduced model performance

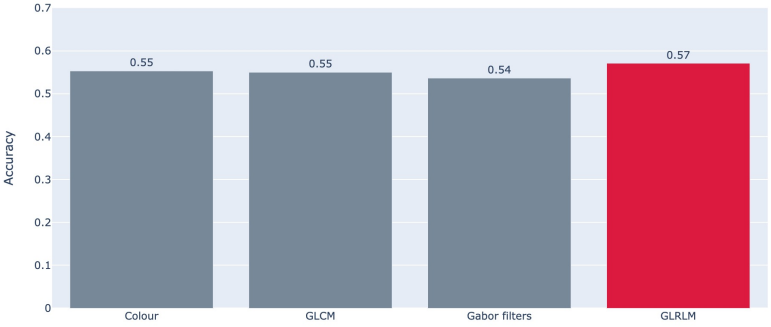

FIGURE 7. Model with single feature

4.2. Feature Analysis

Feature analysis is performed to see which features are helpful in classification while which are not. Since there are mainly four subset of features, in the first round, we remove one subset of features from the model and see how accuracy would change. The result is shown in Figure 6. Accuracy loss is largest when GLRLM is removed from the model, with accuracy decreasing by 40%. Accuracy decreases modestly when color and Gabor filters are removed from the model. Surprisingly, Accuracy increases by 0.3% if GLCM feature is removed.

In the second round, only one subset of feature is used in the model. As the result is shown in Figure 7, we can see that with only one set of features, models do not have significantly better performance than a model producing random guesses.

4.3. Classification aggregation

After model construction and classification for each 64x64 windows, we produce the classification result for each 512x512 patches. A threshold need to be set to determine the lowest percentage of windows classified as 1(metastasis) a patch should contain when it is classified as 1(metastasis). For each patch, truth label is attached when they are extracted. Compared with truth table, we derive the patch-level classification accuracy given a certain threshold, which is shown in Figure 8.

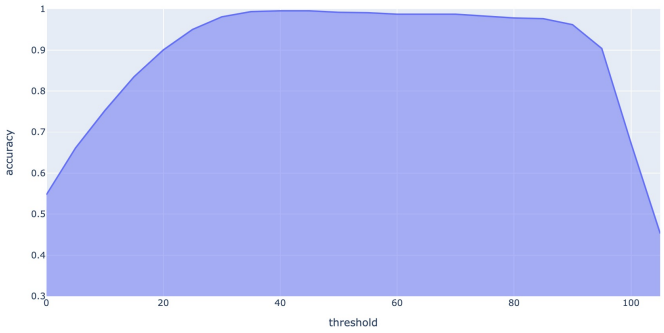

FIGURE 8. Patch-level Classification Accuracy

The accuracy, starting at 54.7% when threshold is 0, reaches the highest, 99.5% when threshold is 40%. Then accuracy remains approximately unchanged until threshold hits 90%. It decreases rapidly to 45.3% when the threshold is set to 100%.

5. DISCUSSION

From Table 1, we are confident to say our model tells normal cells and metastasis apart. Now it is very interesting to answer our question: **is it GLRLM that makes the whole classification work?** The answer is both **yes and no**. The answer is yes not only because the classification accuracy of the saturated model is as high as 94% , but also if we have a look at Figure 6, when the GLRLM is removed from the model, the performance suddenly drops by 40% to around 50% classification accuracy. It means the reduced model would perform just a little bit better than a random guesser without GLRLM. However, the answer is also no because if we have a look at figure 7, a model that only uses GLRLM has roughly the same accuracy as with other single-feature models. It means, with GLRLM alone, the model could not function well, even though GLRLM has the biggest number of features extracted(624 out of 798).

What we can derive from this research is that, in the task of classification, the relationship between subsets of extracted features is interdependent. GLRLM is useful when in the presence of other features. In addition, we also hypothesize that the reason why our model could achieve such a good result is that the ratio of each feature subset is suitable. Features like color and six properties of GLCM have very few degrees of freedom, this is unlikely to separate the classes. But these features could provide a succinct generalization of each data point, and thus could partition them into different categories. Then it is easy for GLRLM to find a fine cut that separates two classes.

Finally, since, it is very inconvenient to classify on the whole slide image, Doing classification on smaller patches and then aggregating results to get higher-level classification results would be a common practice. In the process of aggregation, selecting the right threshold

is quite significant. If we have a look at Figure 8, selecting threshold to be zero means every patch will be classified as metastasis, then result would be the percentage of metastasis patches in the dataset. Similarly, Selecting the threshold to be 100% would get result roughly close to the percentage of patches containing normal cells. To get the highest accuracy, the threshold number selection depends on the case.

6. LIMITATION & FUTURE WORKS

All of our data are extracted from patient045 in the CAMELYON17 dataset. Therefore, the metastases seen by the model are very similar. We notice d that different patients in CAMELYON17 dataset have different cell shapes and metastasis appearances. If given more time, more patches from other patients would be extracted so that our model could be robust.

The usefulness of GLRLM in metastasis detection could be further analysed. More work could be done to see whether a specific length or color group provides significant aid for classification.

Different ways of classification result aggregation could be tried. Classification result of mother patch could be determined by the child patch that has the highest confidence level.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA: A Cancer Journal for Clinicians*, vol. 66, no. 1, 2016.

[2] N. Howlader, A. Noone, M. Krapcho, J. Garshell, D. Miller, S. Altekruse, C. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, *et al.*, "Seer cancer statistics review," *National Cancer Institute*, vol. 2008, 1975.

[3] S. B. Edge, D. R. Byrd, C. C. Compton, and K. (Krankheit), *AJCC Cancer Staging Manual, w. CD-ROM*. AJCC Cancer Staging Manual, w. CD-ROM, 2008.

[4] A. C. Voogd, M. Nielsen, J. L. Peterse, M. Blichert-Toft, H. Bartelink, M. Overgaard, G. Van Tienhoven, K. W. Andersen, R. J. Sylvester, and J. A. Van Dongen, "Differences in risk factors for local and distant recurrence after breast-conserving therapy or mastectomy for stage i and ii breast cancer: pooled results of two large european randomized trials.," *Journal of Clinical Oncology Official Journal of the American Society of Clinical Oncology*, vol. 19, no. 6, pp. 1688–97, 2001.

[5] T. Fehm and D. Wallwiener, "[axillary dissection vs. no axillary dissection in women with invasive breast cancer and sentinel node metastasis: implications for the radiation oncologist].," 2012.

[6] A. E. Giuliano, K. V. Ballman, L. Mccall, P. D. Beitsch, M. B. Brennan, P. R. Kelemen, D. W. Ollila, N. M. Hansen, P. W. Whitworth, and P. W. Blumencranz, "Effect of axillary dissection vs no axillary dissection on 10-year overall survival among women with invasive breast cancer and sentinel node metastasis," *JAMA*, vol. 318, no. 10, p. 918, 2017.

[7] G. Litjens, P. Bandi, B. EhteshamiBejnordi, O. Geessink, M. Balkenhol, P. Bult, A. Halilovic, M. Hermsen, R. vandeLoo, R. Vogels, Q. F. Manson, N. Stathonikos, A. Baidoshvili, P. vanDiest, C. Wauters, M. vanDijk, and J. vanderLaak, "1399 H amp;E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset," *GigaScience*, vol. 7, 05 2018.

[8] J. Vestjens, M. J. Pepels, M. D. Boer, G. F. Borm, C. V. Deurzen, P. V. Diest, J. V. Dijck, E. Adang, J. Nortier, and E. Rutgers, "Relevant impact of central pathology review on nodal classification in individual breast cancer patients," *Annals of Oncology Official Journal of the European Society for Medical Oncology*, vol. 23, no. 10, p. 2561, 2012.

[9] G. Litjens, C. Snchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, H. Christina, P. Bult, B. V. Ginneken, and V. Jeroen, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific Reports*, vol. 6, p. 26286, 2016.

[10] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates," *Cancer Letters*, vol. 77, no. 2, p. 163, 1994.

[11] Y. Xie, J. Zhang, Y. Xia, M. Fulham, and Y. Zhang, "Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest ct," *Information Fusion*, pp. 102–110, 2018.

[12] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, "Differential data augmentation techniques for medical imaging classification tasks," in *AMIA annual symposium proceedings*, vol. 2017, p. 979, American Medical Informatics Association, 2017.

[13] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, *et al.*, "Niftynet: a deep-learning platform for medical imaging," *Computer methods and programs in biomedicine*, vol. 158, pp. 113–122, 2018.

[14] S.-C. Park, J. H. Cha, S. Lee, W. Jang, C. S. Lee, and J. K. Lee, "Deep learning-based deep brain stimulation targeting and clinical applications," *Frontiers in neuroscience*, vol. 13, p. 1128, 2019.

[15] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?," in *2016 international conference on digital image computing: techniques and applications (DICTA)*, pp. 1–6, IEEE, 2016.

[16] A. Rozsa, M. Gunther, and T. E. Boult, "Towards robust deep neural networks with bang," *arXiv preprint arXiv:1612.00138*, 2016.

[17] P. Bndi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. Ehteshami Bejnordi, B. Lee, K. Paeng, A. Zhong, Q. Li, F. G. Zanjani, S. Zinger, K. Fukuta, D. Komura, V. Ovtcharov, S. Cheng, S. Zeng, J. Thagaard, A. B. Dahl, H. Lin, H. Chen, L. Jacobsson, M. Hedlund, M. etin, E. Halc, H. Jackson, R. Chen, F. Both, J. Franke, H. Ksters-Vandevelde, W. Vreuls, P. Bult, B. van Ginneken, J. van der Laak, and G. Litjens, "From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 550–560, 2019.

[18] N. Ostu, O. Nobuyuki, and N. Otsu, "A threshold selection method from gray- level histogram ieee transactions on systems," *IEEE Trans.syst.man. Cybern*, vol. 9, no. 1, pp. 62–66, 1979.

[19] L. Geert, B. Peter, E. B. Babak, G. Oscar, B. Maschenka, B. Peter, H. Altuna, H. Meyke, van de Loo Rob, and V. Rob, "1399 he-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset," *GigaScience*, no. 6, p. 6, 2018.

[20] "The camelyon17 challenge;2017." `https://camelyon17.grand-challenge.org/`. Accessed 13 November 2017.

[21] L. GJS, "Automate slide analysis platform (asap); 2017." `https://github.com/geertlitjens/ASAP/`. Accessed 17 October 2017.

[22] "Openslide; 2017." `http://openslide.org/`. Accessed 17 October 2017.