This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

High-magnification Multi-views Based Classification of Breast Fine Needle Aspiration Cytology Cell Samples using Fusion of Decisions from Deep Convolutional Networks

Hrushikesh Garud*, S. P. K. Karri, Debdoot Sheet, Jyotirmoy Chatterjee, Manjunatha Mahadevappa, and Ajoy K. Ray
Indian Institute of Technology Kharagpur, 721 302, INDIA
*hrushikesh.garud@ti.com

Arindam Ghosh
Sub-divisional Hospital,
Kharagpur, 721 302, INDIA

Ashok K. Maity
Midnapur Medical College & Hospital,
Midnapur, 721101, INDIA

Abstract

Fine needle aspiration cytology is commonly used for diagnosis of breast cancer, with traditional practice being based on the subjective visual assessment of the breast cytopathology cell samples under a microscope to evaluate the state of various cytological features. Therefore, there are many challenges in maintaining consistency and reproducibility of findings. However, digital imaging and computational aid in diagnosis can improve the diagnostic accuracy and reduce the effective workload of pathologists. This paper presents a deep convolutional neural network (CNN) based classification approach for the diagnosis of the cell samples using their microscopic high-magnification multi-views. The proposed approach has been tested using GoogLeNet architecture of CNN on an image dataset of 37 breast cytopathology samples (24 benign and 13 malignant), where the network was trained using images of 54% cell samples and tested on the rest, achieving 89.7% mean accuracy in 8 fold validation.

1. Introduction

Cancer is the third leading cause of death worldwide preceded only by cardiovascular, infectious and parasitic diseases. Breast malignancy is the second most common type of malignancy after lung cancer and the fifth most common cause of cancer death worldwide. According to the World Health Organization projections, breast cancer caused 559,000 deaths worldwide in the year 2008. The incidence rate of breast cancer is increasing rapidly in developing countries. By the year 2030, South-East Asia will witness nearly 186,000 breast cancer deaths annually, more than the higher income countries (176,000 deaths). Worldwide mortality due to breast cancer is expected to stand at 790,000 [22][1].

Fine needle aspiration cytology (FNAC) is one of the most commonly used pathological investigations for screening, and diagnosis of breast cancer. The traditional practice of breast FNAC is based on subjective assessment. Here, microscopic appearance of the aspirates is visually evaluated on various cytological criteria. Therefore, there are many challenges in maintaining consistency and ensuring reproducibility in findings are inevitable. Moreover, there exists an overlap in the state of various cytological criteria in benign and malignant lesions, and inadequate or non-representative sampling may lead to equivocal diagnosis [18, 7]. With reported rates of $6.9 - 20\%$ [18] for equivocal diagnosis against the real gray zone of 2% [2], there exists scope to reduce the rate of equivocal diagnosis.

Digital imaging and computational aid in diagnosis can help to improve the diagnostic accuracy and to reduce the effective workload of a pathologist. In this regard, researchers and practitioners of pathology have been using quantitative analysis for computer-aided diagnosis(CAD) of pathology samples including breast FNAC [16, 6]. This paper presents a deep convolutional neural network (CNN) based classification approach for the diagnosis of the breast FNAC cell samples using their microscopic high-magnification multi-views. The proposed approach of cell sample classification uses majority voting based decision scheme over class prediction obtained for individual views. The Figure 1 summarizes the contribution of this paper in-

[1] GBD 2004 Summary Tables, compiled and published by Health statistics and informatics Department, World Health Organization, Geneva, Switzerland

cluding the creation of image dataset of high-magnification multi-views of breast cytopathology samples and CNN based breast FNAC cell sample classification.

The rest of the paper is organized as- Section 2 describes the prior art for computer vision, and machine learning techniques used in breast cancer diagnosis by FNAC image analysis and deep learning techniques used in histopathology image analysis, Section 3 describes the image dataset developed by us and used during experimentation for this paper, Section 4 describes our methodology, Section 5 presents details of the experimental setup used, Section 6 presents classification results obtained along with discussion of the findings. Conclusions for the study are presented in Section 7.

2. Prior Art

Researchers and practitioners of pathology have been using quantitative analysis to improve diagnostic accuracy and to reduce the effective workload of a pathologist. With recent developments in cost-effective and high-performance computer technology, the digital pathology has become amenable to the application of quantitative analysis in the form of decision support systems [30, 31] and computer vision and machine learning techniques based CAD systems. The CAD systems for digital pathology applications are being developed and deployed for some time now [16, 6]. Review of the prior art shows that computer vision systems commonly used in cytological diagnosis apply the bottom-up approach of diagnostic reasoning from evidence to hypothesis [23]. It involves segmentation of primitives such as clusters, cells, and nuclei using image processing and segmentation techniques [10]; quantification of diagnostically significant cytological criteria [13] using techniques that extract morphometric, densitometric, textural and structural features [24]; followed by pattern recognition techniques for prediction of abnormalities and anomalies [20].

Recently deep neural networks like autoencoders [32] are increasingly finding their way into solving whole slide histopathology image analysis challenges while jointly learning the representative feature space and classification margin. Some contributions related to radiological and interventional image analysis include [26, 1, 21]. The prior art [27] on the breast cancer histopathological image classification dataset (BreakHis database) [28] had earlier used an AlexNet [19] based CNN architecture for classifying whole slide histopathology and achieved 84-90% accuracy. Das et. al [5] in their approach combine predictions by transfer learned GoogLeNet [29] architecture of a CNN for random multiple images of a breast histopathology sample acquired at multiple magnifications to arrive at whole slide diagnosis achieving 94.67% accuracy in multifold validation over the BreakHis database.

3. Breast FNAC image dataset

To validate the efficacy of the CNNs in the representation of breast cytopathology features and tuning of the classification margin we have developed an image dataset of 37(24 Benign + 13 Malignant) cell sample slides from archives of our institute. Where cell samples were obtained from routine FNAC performed on the patients with palpable masses in the breast by an expert cytologist with freehand aspiration. The slides were prepared by wet fixation and H&E staining methods commonly used for primary diagnosis from FNAC [13]. To create the image dataset, the slides were imaged at high magnification (an image per field of view with relevant cytological information) using Leica DM750 microscope, where each cell sample was represented by 4-6 views selected by a pathologist containing cytological evidence sufficient for diagnosis by an expert. For each of the high magnification views an image of size 1024×768 pixels was captured and pre-processed for non-uniform luminance correction [33] and white balance correction [11]. Followed by the manual selection of multiple regions of interests (ROI) of size 256×256 pixels to represent the cell samples. Overall, 918 ROIs from 175 high magnification views over 37 samples constitute the image dataset.

The microscope setup involved imaging using $40\times, 0.65NA$ magnification objective lens and Leica DFC295 camera attached to the microscope via a $0.5\times$ optocoupler and housing a 1/2" 1024×768 resolution color image sensor (2x2 binning). During image acquisition, the camera was programmed to provide RGB coded pixel data without any preprocessing. Focus and field illumination settings were user defined, with variations within the acceptable range, over which images retain their diagnostic value for a human expert.

4. Training CNN for breast FNAC image classification

The deep learning method used in these experiments employed (i) training of the CNN to represent breast FNAC features and tune the classification margin; (ii) majority voting based cell sample classification using ROI classifications.Here, the GoogLeNet [29] a pre-defined architecture of CNN consisting of two convolution layers, five pooling layers, and nine *inception* layers, was trained to classify the breast FNAC ROI images. In the GoogLeNet architecture, a *inception* layer consists of 6 different sizes of filters within a single layer, with 4^{th} and 7^{th} inception layers connected to 2 auxiliary classifiers over the network. The auxiliary classifier helps to fix the vanishing gradient problem associated during error back-propagation during training of a big neural network. For experimentation, the NVIDIA Deep Learning GPU Training System (DIGITS) was used. Where,

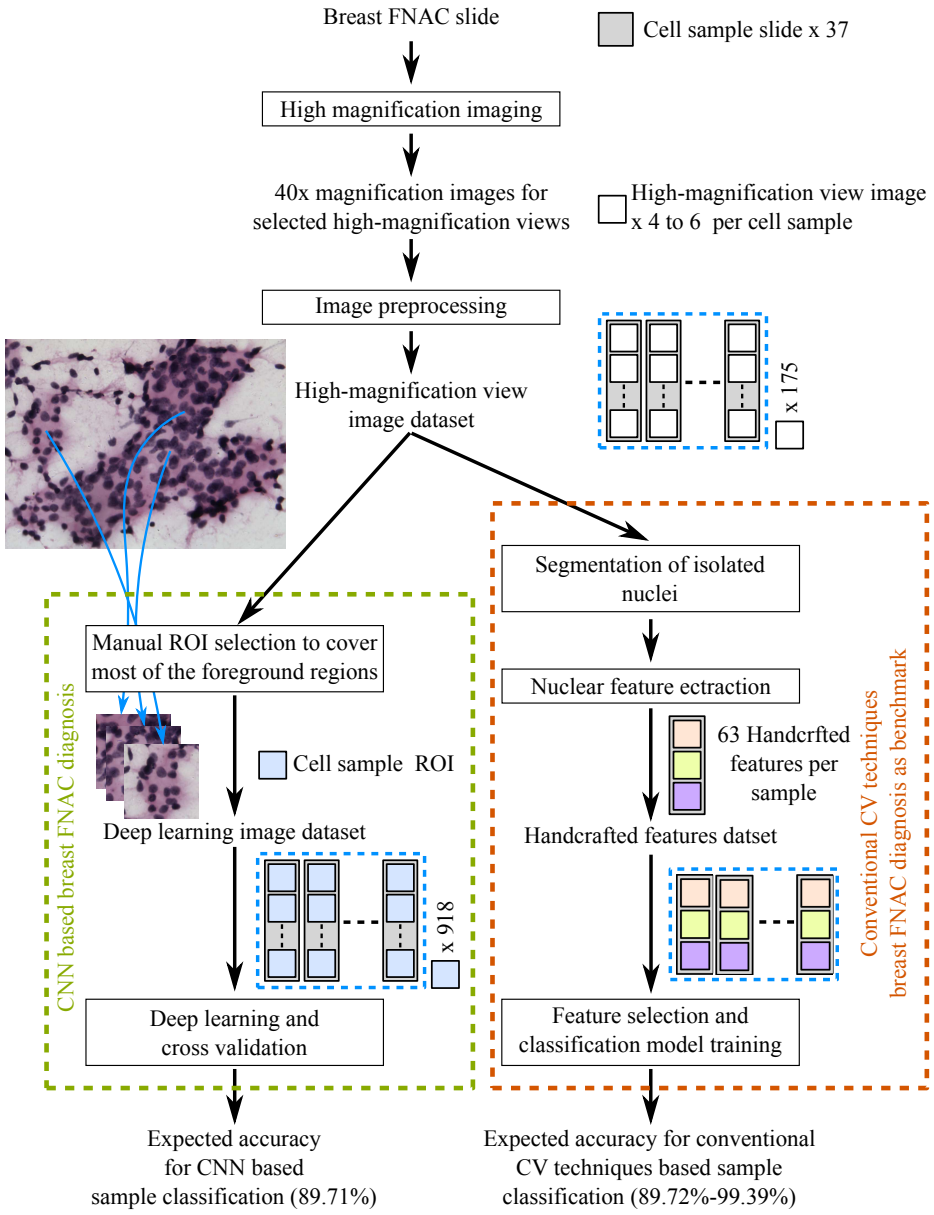

Figure 1. Overview of our proposed approach for benign vs. malignant classification of breast FNAC cell samples using multiple views of the sample at high magnification and consequent benchmarking with conventional computer vision approaches to breast FNAC diagnosis.

models parameters of learning rate, step size, epochs and batch size were set to 0.001, 20, 200, and 64 respectively. The weights initialization method for the convolution layers was done by the method proposed by [14] and optimization technique used was ADAM [17].

Since clinicians report histopathology using multiple views, we use a majority voting scheme based on such images for classifying a cell sample. For the cell sample classification by voting scheme, a cell sample was classified as malignant if more than 30% of its ROI images were classified as malignant otherwise it was classified as benign.

5. Experimental Setup

5.1. Multi-experiment validation

To estimate the expected classification efficacy of the CNN, a sample level, randomized train-test trial based validation (*eight* experiments) was performed. During each experiment, the ROI image dataset was split into two groups at cell sample level, where, training data used to train the CNN comprised of ROIs corresponding to 20 randomly selected samples (12 benign and 8 malignant). The testing data comprised of remaining ROIs.

5.1.1 A Train-test experiment

A train-test trial in the context of this paper consisted of an experiment to train and test the classification accuracy for a training data-classifier combination. In this regard, the features data was randomly split into non-overlapping sets of training and test data, and CNN was then trained and tested with the sets respectively. The measures of accuracy, positive predictive value and negative predictive value in two class classification problem have been used as objective measures in these studies. Where a positive sample is from a malignant breast lesion and a negative sample from a benign breast lesion. The process of a randomized train-test trial is depicted in Figure 2.

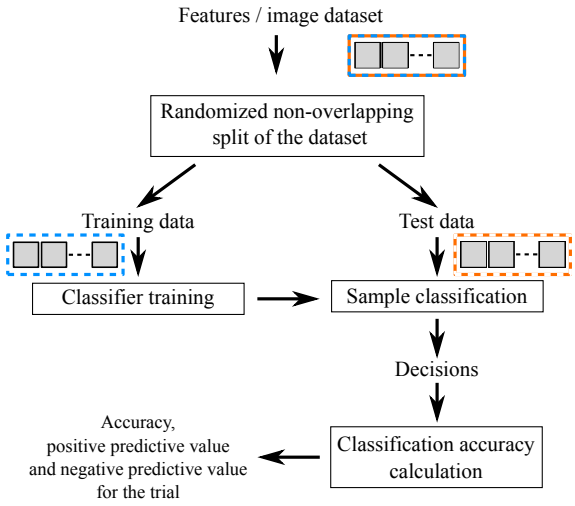

Figure 2. A train-test trial for estimation of classification accuracy, positive predictive value and negative predictive value

5.2. Comparative methods

In this paper, the classification performance of the CNN technique has been benchmarked against our custom implementation of the conventional bottom-up approach of diagnostic reasoning from evidence to hypothesis [23]. Where, *five* cytological criteria of- nuclear size, nuclear shape, nuclear membrane, nuclear chromatin and cellular/nuclear pleomorphism [13], quantified using handcrafted features extracted using the same 175 high magnification views, are applied as evidence for classification by multiple statistical classifiers.

This approach of using nuclear cytological criteria as evidence for cell sample diagnosis, involved (i) segmentation of isolated nuclei in the high magnification images of a sample using a multistage segmentation system involving various image processing [11, 12] and segmentation techniques [10]; (ii) quantification of nuclear features and derivation of features to represent the cell samples in the classification problem (selection of optimal set of features by analysis of a master dataset of handcrafted features using Feature Usability Index technique [25] (feature ranking + forward addition)); followed by classification by naive Bayes [8], support vector machine [4], and random forest classifiers [3] for prediction of abnormalities and anomalies.

The master feature dataset involves 63 features representing each of the cell samples. To determine the 63 features, mean, variance and worst values of 21 nuclear features (11-morphometric, 7-textural, and 3-densitometric) computed over isolated nuclei in the high magnification images of the cell sample were used. Where mean of largest 10% feature values for the sample were used as 'worst' value of the nuclear features computed over isolated nuclei. *Eleven* morphometric features extracted from each of the segmented nuclei include- area, mean radius, equivalent diameter, major axis length and minor axis length quantifying nuclear size; perimeter quantifying nuclear size and shape; dynamic range of radius, eccentricity, and solidity quantifying nuclear shape; and variance of distance of peripheral points from centroid and number of concave points that quantify the state of nuclear membrane. *seven* textural features extracted for each of the segmented nuclear regions include- contrast, homogeneity, correlation, and energy measures computed using the gray level co-occurrence matrix [15], and gray level and run-length non-uniformity measures calculated using gray level run length matrix [9]. The *three* densitometric features are (i) mean hematoxylin stain quantity computed over a nuclear region, (ii) mean hematoxylin stain quantity combined with the area, and (iii) mean hematoxylin stain quantity combined with the variance of the nuclear area over the sample.

To determine expected classification for the classifiers 1000 randomized train-test experiments were conducted. During experimentation, the computed features data for 37 cell samples was split into two groups as training data corresponding to 24 randomly selected samples (16 benign and 8 malignant) and testing data comprising of remaining 13 samples (8 benign and 5 malignant). Linear dot product kernel was used for the support vector machine classifier, where margin selection was performed by sequential minimal optimization over 2000 iterations with the tolerance level of 1×10^{-3}. The random forest was defined as a classifier consisting of 32 tree-structured decision makers where the minimum number of samples terminating at a leaf was set to *three*, and each tree casts a unit vote.

6. Results

Results for the multi-experiment cross-validation as described in Section 5.1 are presented in Table 1. The table presents classification accuracy performance across each experiment along with the details of training and test data in each experiment. From the results, it can be observed

that in general GoogLeNet [29] can learn visual features and classification margin from different training samples during transfer learning and achieves the ROI level classification accuracy of 80.76%. The voting scheme based cell sample classification achieves the accuracy of 89.71% at 85.48% positive predictive value and 92.96% negative predictive value. This accuracy is comparable with the accuracy of $84 - 90\%$ that [27] achieved on the breast histopathology dataset. However, it is less than the classification accuracies achieved with conventional approach of breast cancer diagnosis summarized in Table 2, which achieves up to 99.39% accuracy using just as many as 14 handcrafted features.

7. Conclusion

In this paper, we presented an image dataset of high-magnification multi-views of breast cytopathology samples, along with the performance of GoogLeNet CNN architecture in breast FNAC cell sample diagnosis in malignant or benign categories. It was observed that in general GoogLeNet can learn visual features and classification margin from different training samples included in the dataset and achieves the ROI level classification accuracy of 80.76%. The voting scheme based cell sample classification achieves the accuracy of 89.71%. which is less than the classification accuracies achieved with conventionally defined 'Computed Feature Dataset' and statistical classifiers. The proposed scheme can be considered as a baseline for future research. Data augmentation by adding more samples and data replication, transfer learning can be used to improve classification.

References

[1] Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18 – 31, 2017.

[2] N. Al-Kaisi. The spectrum of the "gray zone" in breast cytology. a review of 186 cases of atypical and suspicious cytology. *Acta cytologica*, 38(6):898–08, 1994.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[5] K. Das, S. P. K. Karri, A. Guha Roy, J. Chatterjee, and D. Sheet. Classifying histopathology whole-slides using fusion of decisions from deep convolutional network on a collection of random multi-views at multi-magnification. In *2017 IEEE International Symposium on Biomedical Imaging*.

[6] C. Demir and B. Yener. Automated cancer diagnosis based on histopathological images: a systematic survey. *Rensselaer Polytechnic Institute, Tech. Rep*, 2005.

[7] B. S. Ducatman and H. H. Wang. Chapter 8 - breast. In E. S. Cibas and B. S. Ducatman, editors, *Cytology (Third Edition)*, pages 221 – 254. W.B. Saunders, Philadelphia, third edition edition, 2009.

[8] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[9] M. M. Galloway. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*, 4(2):172 – 179, 1975.

[10] H. Garud, S. P. K. Karri, D. Sheet, A. K. Maity, J. Chatterjee, M. Mahadevappa, and A. K. Ray. Methods and system for segmentation of isolated nuclei in microscopic images of breast fine needle aspiration cytology images. In *MedImage-2016*, 2016.

[11] H. Garud, A. K. Ray, M. Manjunatha, J. Chatterjee, and S. Mandal. A fast auto white balance scheme for digital pathology. In *Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference on*, pages 153–156, June 2014.

[12] H. Garud, D. Sheet, A. Ray, M. Mahadevappa, and J. Chatterjee. Adaptive weighted-local-difference order statistics filters, 2015. US Patent 9,208,545.

[13] H. T. Garud, D. Sheet, M. Mahadevappa, J. Chatterjee, A. K. Ray, and A. Ghosh. Breast fine needle aspiration cytology practices and commonly perceived diagnostic significance of cytological features: A pan-india survey. *Journal of Cytology*, 29(3):183, 2012.

[14] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *13^{th}, International Conference on Artificial Intelligence and Statistics (AISTATS) 2010*, 2010.

[15] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, Nov 1973.

[16] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu. Methods for nuclei detection, segmentation, and classification in digital histopathology: A review– current status and future potential. *IEEE Reviews in Biomedical Engineering*, 7:97–114, 2014.

[17] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] G. Kocjan. *Diagnostic Dilemmas in FNAC Cytology: Difficult Breast Lesions*, pages 181–211. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[20] L. Langer, Y. Binenbaum, L. Gugel, M. Amit, Z. Gil, and S. Dekel. Computer-aided diagnostics in digital pathology: automated evaluation of early-phase pancreatic cancer in mice. *International Journal of Computer Assisted Radiology and Surgery*, 10(7):1043–1054, 2015.

[21] D.-Y. Liu, T. Gan, N.-N. Rao, Y.-W. Xing, J. Zheng, S. Li, C.-S. Luo, Z.-J. Zhou, and Y.-L. Wan. Identification of lesion images from gastrointestinal endoscope based on feature extraction of combinational methods with and without learning process. *Med. Image Anal.*, 32:281–294, 2016.

[22] C. Mathers, D. M. Fat, and J. T. Boerma. *The global burden of disease: 2004 update*. World Health Organization, 2008.

Table 1. Accuracy of classifying cell samples in the test dataset using deep learning approach. Each sample level classification performance entry in the table presents accuracy, positive predictive value, and negative predictive values respectively

	Training ROIs	**Test ROIs**	**Sample level calssification performance**			**ROI level classification accuracy**
	Benign, Malignant		Accuracy	Positive Predictive Value	Negative Predictive Value	
Experiment 1	280, 157	345, 136	94.12%	83.33%	100.00%	83.31%
Experiment 2	288, 199	337, 94	88.24%	80.00%	91.67%	81.20%
Experiment 3	317, 176	308, 117	94.12%	100.00%	92.31%	84.26%
Experiment 4	389, 191	236, 102	94.12%	83.33%	100.00%	82.62%
Experiment 5	330, 173	295, 120	88.24%	100.00%	85.71%	81.06%
Experiment 6	312, 203	313, 90	94.12%	100.00%	92.31%	82.29%
Experiment 7	322, 227	303, 66	88.24%	80.00%	91.67%	80.19%
Experiment 8	305, 180	320, 113	76.47%	57.14%	90.00%	71.15%
Average			**89.71%**	**85.48%**	**92.96%**	80.76%

Table 2. Classification performance of the naive Bayes, support vector machine and random forest classifiers using handcrafted nuclear morphometric, densitometric and textural features.

Classifier	**Accuracy**	**Positive predictive value**	**Negative predictive value**	**Number of features in the optimal subset**
Naive Bayes classifier	89.72%	74.68%	99.13%	45
Support vector machine	97.05%	94.94%	98.36%	13
Random forest	99.39%	98.42%	100.00%	14

[23] V. L. Patel and G. J. Groen. Knowledge based solution strategies in medical reasoning. *Cognitive Science*, 10(1):91 – 116, 1986.

[24] K. Rodenacker and E. Bengtsson. A feature set for cytometry on digitized microscopic images. *Analytical Cellular Pathology*, 25(1):1–36, 2003.

[25] D. Sheet, J. Chatterjee, and H. Garud. Feature usability index and optimal feature subset selection. *International Journal of Computer Applications*, 12(2):29–36, 2010.

[26] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. J. Snead, I. A. Cree, and N. M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imaging*, 35(5):1196–1206, 2016.

[27] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte. Breast cancer histopathological image classification using convolutional neural networks. In *Proc. Int. Jt. Conf. Neural Net.*, 2016.

[28] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, July 2016.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. IEEE Conf. Comp. Vis., Patt. Recog.*, pages 1–9, 2015.

[30] D. West and V. West. Model selection for a medical diagnostic decision support system: a breast cancer detection case. *Artificial Intelligence in Medicine*, 20(3):183 – 204, 2000.

[31] W. H. Wolberg and O. L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the national academy of sciences*, 87(23):9193–9196, 1990.

[32] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE Trans. Med. Imaging*, 35(1):119–130, 2016.

[33] I. T. Young. Shading correction: compensation for illumination and sensor inhomogeneities. *Current Protocols in Cytometry*, pages 2–11, 2000.