A Convolutional Neural Network to Assess Malignancy in Breast Cancer Histology

Jesse Knight
Image Analysis In Medicine Lab
University of Guelph
Guelph, ON, N1G 2W1
jesse.x.knight@gmail.com

Abstract

Histological grade provides important prognostic information for breast cancer, yet manual assignment has low inter-observer agreement. Methods for automated analysis have yet to consider a convolutional neural network (CNN) approach to this task. We therefore present a basic CNN model for differentiating between malignant and benign breast cancer histology images. Our model achieves a humble accuracy of 69 ± 5%. However, we demonstrate the sensitivity of this approach to quantity of training data, filter initialization methods, and image normalization steps. We also characterize the types of features used for classification in this approach, which include nucleus-like objects and structured combinations of these.

1 Introduction

1.1 Histopathology in Breast Cancer

Breast cancer is the second leading cause of cancer death in U.S. women [1]. However, survival rate is highly dependent on the prognosis of tumour(s); while the 5 year rate for localized tumours is 98.6%, it is only 23.3% for metastatic cancer [1]. Histopathology is the analysis of stained, biopsy-extracted tissue samples under a microscope. Despite advances in genomic tumor profiling, histological grading continues to be a reliable and inexpensive indicator for breast cancer prognosis [2, 3]. High grade tumours are more aggressive and associated with increased likeliness of metastasis [2]. Tumour grade is determined using semi-quantitative morphological features in stained tissue sections, including number of mitotic (dividing) cells, amount of structure in the tissues, and nuclei size, shape and texture [2, 3]. These key malignancy features are shown in Figure 1.

Unfortunately, manual tumour grading is highly subjective, and inter-observer agreement for the popular Nottingham Grading System is surprisingly low, with kappa statistics ranging from 0.23 – 0.84 [3, 4]. This challenge has stimulated research into the development of automated methods for histological grading [5, 6, 7]. An ideal automated models would generate identicle results, regardless of institution, scan conditions, user, and stain parameters. Stain parameters, in fact, are a significant obstacle to conventional image processing approaches, as slide digitization protocols are not yet standardized, resulting in large variability of hue and saturation characteristics in images.

In this paper, we present a basic convolutional neural network (CNN) model for classifying breast cancer histology images as malignant or benign. This method has the capacity to overcome many of the challenges of other methods. We also examine a selection of the important parameters for the success of this approach.

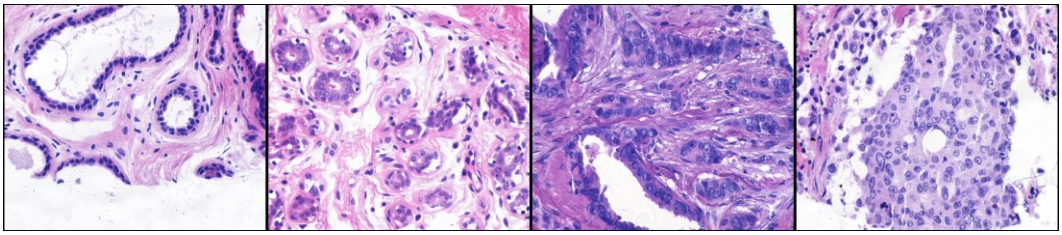

Figure 1: (Left to right) increasing histological features of malignancy: diminishing tubule structure, increasing nucleus texture, size and transparency.

1.2 Convolutional Neural Networks

Convolutional neural networks (CNN) have recently emerged as a powerful tool for image analysis and classification, having achieved landmark accuracies on challenging datasets like ImageNet [8, 9]. Similar to standard neural networks, CNNs are feedforward, often supervised, and usually 'deep'. That is, they have many layers which compute weighted combinations of the previous layer's output, then perform some nonlinear operation on the combination and output the result to the next layer. In CNNs, the weights are simply 2 dimensional, and are applied using the convolution operator. The nonlinearities may vary, but usually some pooling function is used.

An important concept for understanding CNN models is the duality between convolution and correlation; they are, in fact, the same operation with the kernel reflected. This allows us to observe that filters (weight matrices) learned in CNN models are directly representative of discriminatory features in the image, provided they are shift and reflection-invariant. Moreover, using cascaded layers of convolution, combinations of the lower features can construct complex representations of the input which are pertinent to the training task [8].

2 Related Works

2.1 Pathology analysis

The majority of existing automated pathology analysis methods minimally employ some machine learning decision tool, be it regression, Support Vector Machines, Naïve Bayes, or Random Forest. Most also detect features partially automatically, using image processing methods like segmentation, morphological feature extraction, and texture analysis [7].

For instance, as early as 20 years ago [10], semi-automated feature detection and simple linear regression for breast cancer histology demonstrated accuracy as high as 90%, however, this method required manual identification of individual nuclei, and only considered single-centre data. Similarly, in [6], the authors extract 50 features based on automatically estimated lymphocyte locations and employ graph embedding to arrive at a crisp manifold depicting a continuum of malignancy. In [5], the authors employ Gabor filter texture features and graph embedding with a SVM kernel to achieve 95.8% accuracy in identifying cancerous images.

Yet, the features employed in such methods are always partly hand-selected, introducing designer-bias and potentially leaving gaps in the translation of image information to the classification decision. Deep CNN models overcome this limitation by inherently considering an immense gamut of potential features as candidates for relevancy, so have the potential to find more optimal solutions. Perhaps as importantly, all of the above works for breast cancer analysis only consider single-centre images; this is not representative of the aforementioned variability in hue and saturation for histology slides, and model robustness to these parameters is critical for clinical utility and institutional uptake.

2.2 Mitotic Identification

One of the most tedious and challenging tasks in histological analysis is the identification of rare mitotic (dividing) cells. Yet this feature is very important, as the number of dividing cells is highly predictive of malignancy [4]. A large number of machine learning models, including

several CNNs have been applied to this specific task of counting mitotic cells [11, 7, 12, 13]. For instance, in [11], a combination of image processing methods and a 'light' CNN model are used to reduce the training expense versus a standard CNN model to identify candidates for and then assign mitotic class membership. However, these methods aim only to identify mitotic cells, and do not consider other malignant features in the image or attempt to classify the entire image as benign or malignant.

3 Approach

We present a basic CNN for predicting the malignancy of a breast cancer histology image. To our knowledge, CNNs have not yet been applied for prognostic classification in breast cancer pathology – only to detect individual features like mitosis.

3.1 Architecture

We designed our CNN to recognize features on the scale of super-cellular structures (e.g. the cross sectioned tubules shown in the input image, Figure 2). This is because the range in scale needed to simultaneously capture individual nucleus features and structural features is approximately 1:100, necessitating sophisticated architectures. In fact, this multi-scale problem is common in CNN applications, and has motivated the development of very deep pyramid models like GoogLeNet and AlexNet. This approach was beyond our time constraints.

We use 3 convolutional layers, divided by pooling layers to ensure some local shift invariance of the filter bank below. Our filter bank sizes are 64 for the first two layers, and a single filter in the final layer. Similarly, the first two layers employ max pooling, while the final layer averages over the final feature map. Average pooling in the final layer ensures that activations over the entire image are considered, preventing decisions based on a single first layer receptive field.

Unlike other approaches, we do not append any fully connected non-convolutional layers before the output. This was due to time constraints of model development, but the restriction emphasizes direct correlations between learned features and the malignancy label. The full path of our network is shown in Figure 2, and parameters are described in Table 1.

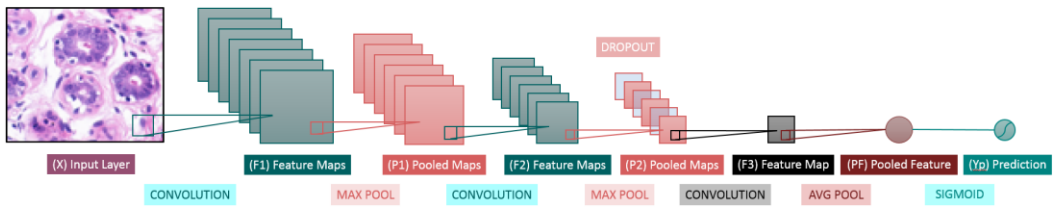

Figure 2: CNN architecture: data structures are labeled in dark boxes, light text while operations are labeled with light boxes, dark text.

Table 1: Model parameters

	Convolution Kernels			Pooling			Receptive Field
	Size	Count	Stride	Size	Type	Stride	
Layer 1	7 × 7	64	2	5 × 5	MAX	2	15 × 15
Layer 2	5 × 5	64	1	5 × 5	MAX	2	47 × 47
Layer 3	5 × 5	1	1	6 × 7	AVG	N/A	119 × 127

Kernel and pool sizes were selected such that the intermediate layer maintains a wide receptive field – large enough to capture tubular features – while permitting variability in the exact organization of the structure. For this, we maintain small convolution operation strides, large filter sizes, and employ slightly larger pool sizes (versus other models).

3.2 Implementation

We construct the model in MATLAB using the recently developed MatConvNet framework [14]. While MatConvNet does not employ symbolic differentiation, large matrix operations are computed efficiently using precompiled MEX files (MATLAB-callable C/C++ source code), and parallel processing options are available for CUDA GPUs.

3.3 Training

We train our model using stochastic gradient descent (SGD) with basic momentum (i.e. not Nesterov) and full-batch learning. Full batch learning encourages reliable weight updates, which combats instability for such a high dimensional model. For momentum, we use $\delta = 0.9$. Higher values were liable to introduce instability. Our error function for a given input X is simply the difference between the sigmoid-predicted label, $Y_{P|\theta} \in (0,1)$, and the true label Y,

$$Error(X, Y \mid \theta) = \left(Y_{P|\theta} - Y\right)$$

For regularization, we employ L_1 weight decay and dropout after layer 2, yielding the objective function,

$$J(X, Y \mid \theta) = Error(X, Y|\theta) + \lambda \sum \left\| \mathrm{W}_{l,k} \right\|_1$$

As noted in [15], dropout is critical to improving generalization performance in CNNs; while the high capacity of these networks makes them extremely vulnerable to overfitting the training set, dropout reduces the potential for codependence of features. We apply 50% dropout to the second layer only, permitting codependence of small 1st layer features.

Finally, we keep learning rate (LR) small, and use layer-specific LR modifiers encourage fast bias learning and very slow output predictor learning, such that high quality low and mid-level features will be constructed.

Table 2: Model training parameters

	Dropout	Weight LR Mod	Bias LR Mod	Base LR	L_1 λ	Momentum
Layer 1	–	0.1	2.0	0.005	0.0005	0.90
Layer 2	$\alpha = 0.5$	0.1	2.0			
Layer 3	–	0.01	2.0			

4 Experiments

4.1 Database

We train and test our CNN on a database of 58 digitized pathology slides, of which 32 are from benign tumours and 26 are from malignant. The database is a standard validation set for local developers of similar algorithms, and, critically, is representative of the wide range in stain hues and saturation levels (Figure 3 (A)) as described above.

We downsample the images, originally size 896×768, by a factor of 5, to size 180×154, Figure 3 (B). This reduces the computational expense in the input layer by 5^2, but also permits smaller 1st layer kernels (e.g. 7×7, Table 2) capable of capturing individual nuclei, therefore accelerating training twofold.

Noting the high chance of overfitting on such a small data set, we also increase the number of training examples by subdividing each image into 4 quadrants, and then flipping each quadrant left-right, up-down, and also in both directions, yielding 16 training frames, size 90×77, per input image (928 total), Figure 3 (C). The potential drawback to this expansion method is the generation of images with mostly background, and very little semantic information, as in Figure 3 (C), lower right quadrant. This makes training and testing equally more challenging.

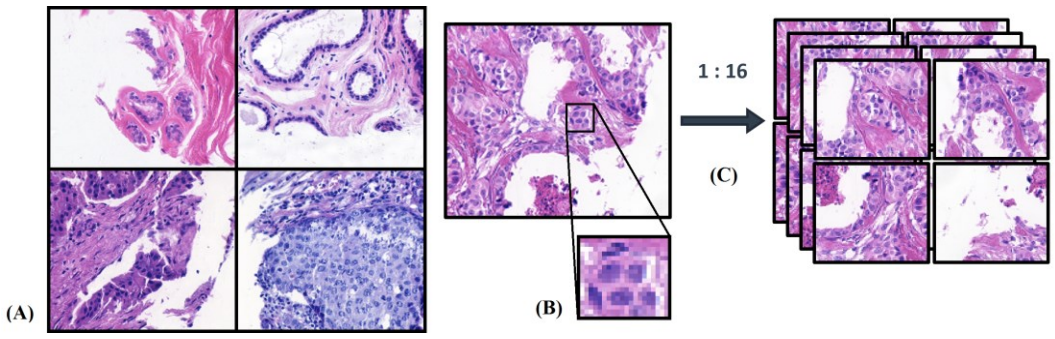

Figure 3: Database characteristics: (A) variability in stain hue and saturation, (B) resolution of nuclei after downsampling by 5, (B) increasing the number of training examples.

4.2 Training & Validation

Our training and test sets were selected randomly based on the original downsampled image (180×154) so that all 16 frames were assigned to the same set. This was done to minimize any similarities between the training images and the test set, as quadrants from the same image may have characteristics which allow the model to 'memorize' the correct label. We randomly selected 30 images (52%, 480 frames) for the training set, and 14 (24%, 224 frames) for the training and validation sets. This approach very conservatively demonstrates performance.

To more fully consider model generalization, we also checkpoint the network parameters every time the validation set error reaches a new minima. While this likely represents chance overfitting on the validation set, it also provides a temporal description of model performance and allows a controlled comparison between validation and training sets to show variability.

Finally, we rotate the four quarters of the database (14 or 15 images) through assignments to training, validation, and test sets for cross validation. We train for 1000 epochs, twice for each assignment (8 total instances) to observe differences associated with random initializations.

4.3 Hyperparameter Selection

We also experimented with different hyperparameters, including image normalization, non-random first layer filter initializations, and learning rate adjustments. We employed a user-guided search, in contrast to grid search, as training the model is very computationally expensive and retraining the model more than tens of times was not feasible given time and resource constraints. Where possible, we make comparisons while varying a given parameter, though unstable solutions prevent this in a few cases.

5 Results & Discussion

5.1 Learning

Due to the number of parameters, CNN models are inherently high dimensional; this makes them sensitive to hyperparameter selection and generally difficult to train. We also observed large variability in the steady-state validation error within and between training instances (Figure 5), indicating that the parameter space unfortunately has many local minima.

Using the MEX-based MatConvNet, training took approximately 13 ms per image-pass, or 6.25 s per 480-image training epoch, requiring 1 h 45 min for a typical 1000 epoch descent.

5.1.1 Input Normalization

We trained the network with and without local normalization of the input images, using local neighbourhood $N_{x,y}$: [7 × 7] mean subtraction,

$$I_N(x,y) = I(x,y) - \mu(N_{x,y})$$

We also considered division by the neighbourhood variance, as shown in Figure 4 (A), but this produced artificially noisy regions and increased non-object contrast in the background; preliminary training (500 epochs) showed very poor performance. Image normalization, in fact, decreased performance overall (Figure 5). We hypothesize that this is due to the reduced contrast for large, structural objects like tubules, which are important discriminatory features.

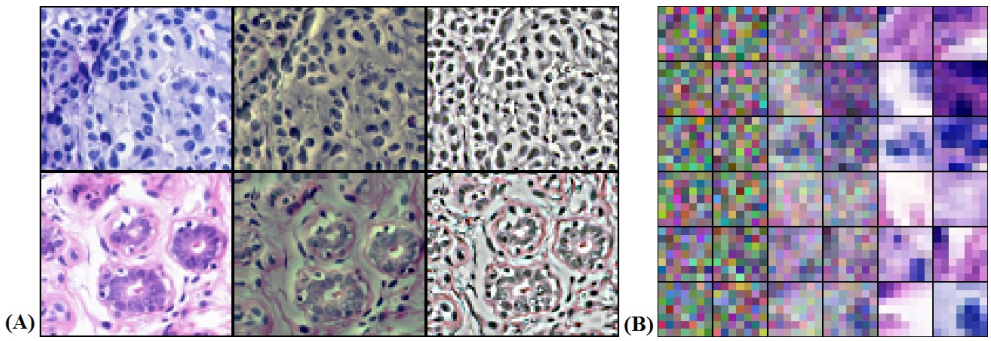

Figure 4: (A) Image normalization impacts: highlighted individual nuclei, but reduced structural object contrast. Left to right: no normalization, mean subtraction, mean and variance normalization. (B) Filter initialization schema (column pairs): random, random-image section mixture, and pure image section.

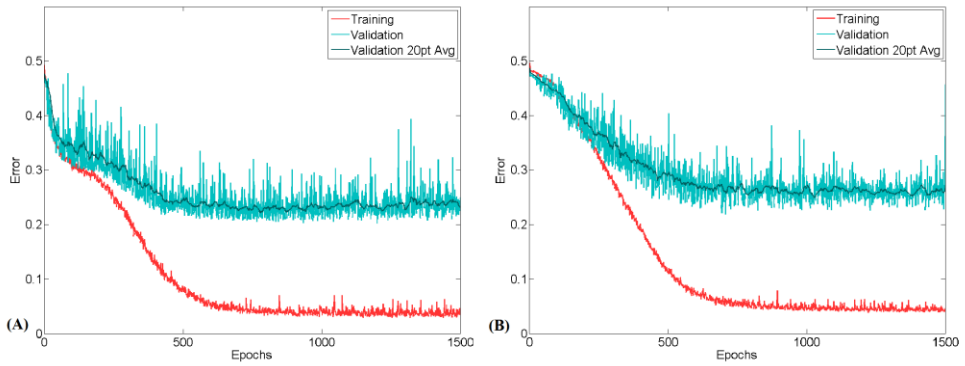

Figure 5: Network convergence (A) without normalization and (B) with mean subtraction.

5.1.2 Weight Initialization

We experimented with different methods for initializing the parameters of the first layer – the input convolutional kernels – including weighted combinations of randomly sampled image sections and normally distributed small random numbers, as shown in Figure 4 (B). We found that the addition of any sampled sections during initialization did not noticeably improve convergence. In fact, in ill-posed scenarios, the model typically became unstable and plateaued for many epochs before we stopped training manually. For instance, the image data is strictly positive, so initializing with un-normalized image sections places the model on the positive half of all dimensions of the vast feature space – likely far from any local optima.

Interestingly, there also exists an extreme case, where, if the image sections are used without any distortion (additive noise), the exact reflection of the initialized kernel is present in the training set, due to our database expansion method and full batch learning. Convolution of this exact section during the first pass, therefore, will simplify to autocorrelation, and the resulting activation will be exceptionally high relative to other filters. We hypothesize that this would induce the 'exploding gradients' problem, resulting in the observed plateau(s).

5.2 Learned Features

Many of the filters learned by the model (Figure 6) contain dark, violet, nucleus-sized features, suggesting that the network employed nuclei for the classification task. We also note that there is a reasonable variety in these features in terms of size and intensity, likely reflective of the differences in actual nuclei. In fact, nuclei from malignant tissue samples are often enlarged and more transparent versus their benign counterparts [2].

Relative to natural image (e.g. ImageNet) trained CNNs, we observe that our learned features are missing typical strong edge filters and Gabor-style wavelet features. While histology slides contain only small amounts of these elements, they are also not likely to be relevant to our classification task. Additionally, many of the kernels from the model trained on un-processed images contain more hue information than those from the normalized image trained model, reflective of the underlying image characteristics in each training set.

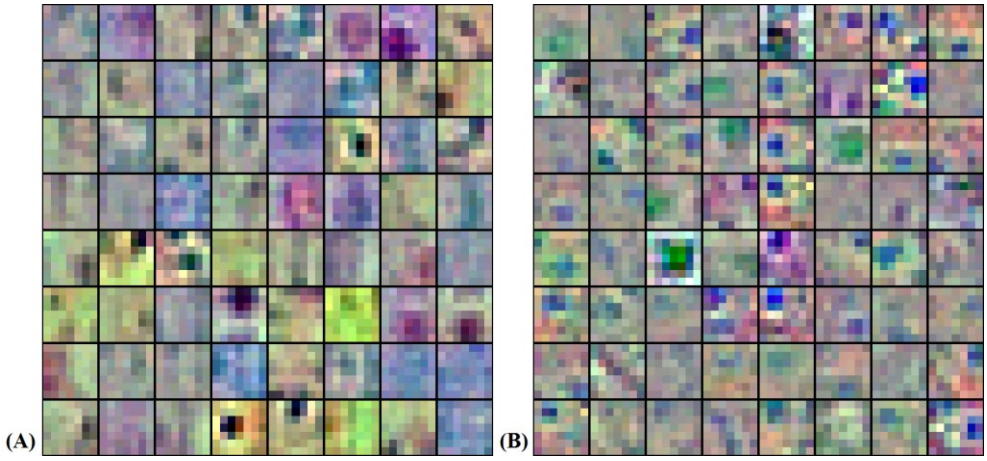

Figure 6: All leaned layer 1 convolutional kernels after 1500 epochs (A) without normalization (B) with input image local mean subtraction.

5.3 Classification Performance

The median test set accuracy over all cross validation runs was 69% (IQR: 5%, Table 3), while the median training set accuracy was 99% (5%). This indicates that significant overfitting occurred, but that some generalization performance was still maintained. Results presented for the validation set are the 'best case scenario', as the model was saved at each new 'best error' checkpoint using this set. All results pertain to image sets without normalizations.

Interestingly, our model is more specific than sensitive, meaning it is more likely to miss a malignant case than wrongly assign a benign image as malignant. Clinically, this would make it useful as a diagnostic aid to confirm pathologists' assessments, versus a screening tool. As is evident in the table and also Figure 5, the model unfortunately varies widely from epoch to epoch; however this is more reflective of our small test database than model performance.

Table 3: 1000-Epoch† model performance on each image set from 4×2 fold validation. Results are shown as median (IQR). †Last 'new best validation error' checkpoint: reason for better validation set performance vs. test set.

Data Set	Error (before rounding)	Accuracy (after rounding)	Sensitivity	Specificity
Training	0.07 (0.06)	0.99 (0.05)	0.98 (0.02)	1.00 (0.04)
Validation	0.27 (0.08)	0.78 (0.10)	0.75 (0.16)	0.81 (0.23)
Test	0.33 (0.05)	0.69 (0.05)	0.62 (0.20)	0.80 (0.28)

Recalling some of the limitations of our database expansion method, we note that many of the false negatives (missed malignant images) actually contain mostly background, as shown in Figure 7 (A). Moreover, we observed that reflected image frames (quadrants) were almost always assigned the same label, and that frames from the same larger source image, were usually also assigned the same label also (except for mostly background frames). This validates our decision to keep image quadrants in the same set, else the model would falsely appear to perform much better than our results indicate.

Finally, Figure 7 demonstrates typical errors (left) and truths (right) predicted by the model, showing that nuclei distribution appears to be the major discriminatory feature learned.

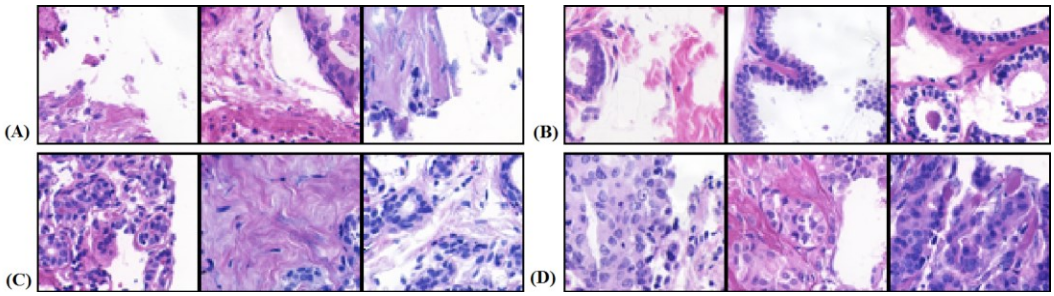

Figure 7: Typical successes and failures of the model: (A) False Negatives with few nuclei (B) True Positives with large structural objects (C) False Positives with minimal structure and (D) True Negatives with large, distributed nuclei.

6 Conclusion

We have demonstrated the potential application of a convolutional neural network for differentiating between malignant and benign pathology images in breast cancer. We designed our network with the aim of extracting structural information based on the arrangement of nuclei in the image. The learned first layer filters suggest that these features are, in fact, important to the classification task, and that the model may have additionally learned simple differences between nuclei shape and opacity.

While our model achieves only moderate accuracy (69%, IQR: 5%), the testing conditions we employed were more rigorous than other proposed methods, including multi-centre images. Moreover, CNN models have not yet been applied to breast cancer grading tasks specifically.

In the future, we recommend investigating unpool-deconvolution methods for visualizing the activations of deeper layers, as this would provide critical insights into discrimination features of the model. Additionally, histology grading CNNs would greatly benefit from the addition of multi-scale feature extraction, as mitotic cell identification and fine nuclear texture are often used by pathologists during manual assessment.

Finally, we express the need for large standardized image databases for training and validating machine learning approaches to histology analysis. For instance, in [6], despite impressive model complexity (50 features), only 41 images from 12 patients and a single centre (i.e. consistent stain parameters) are used for training and validation. The various regularization and validation methods employed in this and related works are only Band-Aid solutions to a lack of training and testing data.

Acknowledgments

I am grateful to Dr. Graham Taylor for his guidance on this project and excellent course design for an introduction to machine learning. I am also grateful to Dr. April Khademi for providing the dataset and facilitating this and other research in image processing. I would also like to acknowledge the NIPS submission style file for helping me overcome my fear of dreadful typesetting – can it get any worse?

References

[1] J. Ma and A. Jemal, "Breast Cancer Statistics," in Breast Cancer Metastasis and Drug Resistance, New York, Springer Science, 2013, pp. 1-18.

[2] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow up," Histopathology, vol. 13, pp. 403-410, 1991.

[3] E. A. Rakha, J. S. Reis-Filho, F. Baehner, D. J. Dabbs, T. Decker, V. Eusebi, S. B. Fox, S. Ichihara, J. Jacquemier, S. R. Lakhani, J. Palacios, A. L. Richardson, S. J. Schnitt, F. C. Schmitt, P.-H. Tan, G. M. Tse, S. Badve and I. O. Ellis, "Breast cancer prognostic classification in the molecular era: the role of histological grade," Breast Cancer Research, vol. 12, no. 207, 2010.

[4] A.-M. O'Shea, E. A. Rakha, A. Hodi, I. O. Ellis and A. H. S. Lee, "Histological grade of invasive carcinoma of the breast assessed on needle core biopsy - modifications to mitotic count assessment to improve agreement with surgical specimens," Histopathology, vol. 59, pp. 543-548, 2011.

[5] S. Doyle, S. Agner, M. Feldman, J. Tomaszewski and M. Anant, "Automated grading of breast cancer histopathology using spectral clustering with textural and architextural image features," ISBI, pp. 496-499, 2008.

[6] A. N. Basavanhally, S. Ganesan, S. Agner, J. P. Monaco, M. D. Feldman, J. E. Tomaszewski, G. Bhanot and A. Madabhushi, "Computerized Image-Based Detection and Grading of Lymphocytic Infiltration in HER2+ Breast Cancer Histopathology," IEEE Transactions on Biomedical Engineering, vol. 57, no. 3, pp. 642-652, 2012.

[7] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot and B. Yener, "Histopathological Image Analysis: A Review," IEEE Reviews in Biomedical Engineering, vol. 2, pp. 147-171, 2009.

[8] Y. Bengio, Learning Deep Architectures for AI, Montreal, 2009.

[9] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems 25, 2012.

[10] W. H. Wolberg, W. N. Street, D. M. Heisey and O. L. Mangasarian, "Computer-Derived Nuclear Features Distinguisg Malignant from Benign Breast Cytology," Human Pathology, vol. 26, no. 7, pp. 792-796, 1995.

[11] H. Wang, A. Cruz-Roa, A. Basavanhally, H. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez and A. Madabhushi, "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features," Journal of Medical Imaging, vol. 1, no. 3, 2014.

[12] C. Malon, M. Miller, H. C. Burger, E. Cosatto and H. P. Graf, "Identifying histological elements with convolutional neural networks," in Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology, Cergy-Pontoise, 2008.

[13] A. Tashk, S. M. Helfroush, M. Akbarzadeh and H. Danyali, "An Automatic Mitosis Detection Method for Breast Cancer Histopathology Slide Images based on Objective and Pixel-wise Textural Features Classification".

[14] A. Vedaldi and K. Lenc, "MatConvNet -- Convolutional Neural Networks for MATLAB," Proceeding of the ACM Int. Conf. on Multimedia, 2015.

[15] N. Srivastava, "Improving Neural Networks with Dropout, MSc Thesis," Graduate Department of Computer Science, University of Toronto, Toronto, 2013.