An Integrated Deep Learning Approach towards Automatic Evaluation of Ki-67 Labeling Index

Lakshmi S.[a], Deepu Vijayasenan[a], Sumam David S.[a], Saraswathy Sreeram[b] and Pooja K Suresh[b]
[a]Department of Electronics and Communication Engineering, National Institute of Technology Karnataka, Surathkal, Karnataka, India
[b]Department of Pathology, Kasturba Medical College Mangalore, Manipal Academy of Higher Education, Manipal, Karnataka, India
lakshmi1510@gmail.com, deepu.senan@gmail.com, sumam@nitk.edu.in

***Abstract*—Ki-67 labeling index is a widely used biomarker for the diagnosis and monitoring of cancer. Many automated techniques have been proposed for evaluating Ki-67 index. In this paper, we introduce an integrated deep learning based approach. We use MobileUnet model for segmentation and classification and connected component based algorithm for the estimation of Ki-67 index in bladder cancer cases. The average F1 score is 0.92 and dice score is 0.96. The mean absolute error in the evaluated Ki-67 index is 2.1.**

We also explore possible pre-processing steps to generalize the segmentation model to at least one another type of cancer. Histogram matching and re-sizing improve the performance in breast cancer data by 12 % in F1 score and 8 % in dice score.

***Index Terms*—Ki-67 index, Carcinoma bladder, Deep Neural Network**

I. INTRODUCTION

The Ki-67 labeling index is an accepted biomarker to assess tumor cell advance and thus predict therapy responses. It is the ratio between the number of the immunopositive tumor cell to all tumor cells [1]. In Ki-67 staining, the color of the cells changes from blue to brown as the cell varies from immunonegative to immunopositive. Manual Ki-67 assessment is a slow process and might include personal biases in the assessment. Objectivity and reproducibility can be improved using automatic image analysis techniques. Such automated methods should be reliable and robust across different image variabilities such as brightness and background color.

In the past few years, several automated methods for Ki-67 index evaluation have been introduced [2], [3]. ImmunoRatio, a free web-based application employs background correction, color deconvolution, thresholding, segmentation, identification of nucleus and classifying them into positive and negative [4], [5]. The Aperio Image analysis software uses color deconvolution and watershed operations for the estimation of Ki-67 index [6]. The non-tumor cells, such as lymphocytes and stromal cells, are to be removed manually. Hence it is not completely automatic.

Recently machine learning based algorithms were used for Ki-67 assessment [7]–[9]. These algorithms identify the nuclei based on seed point detection [7], [8] or by using morphological operations [9]. Textural and cellular features are used as an input for a support vector machine (SVM) based machine learning classifier. The classified cells are then counted for the estimation of the Ki-67 index.

An improved deep learning approach is based on Gamma Mixture Model with Expectation-Maximization for seed point detection and patch-selection and a deep learning model for classification. The deep learning model used consists of five convolutional layers, two fully connected layers, and one decision layer [10]. These algorithms perform segmentation and classification in independent steps due to which errors in each stage gets accumulated. Besides, all the algorithms target only one type of cancer.

In this work, we propose a deep learning model with integrated segmentation and classification. We use MobileUnet for the semantic segmentation and a blob extraction based algorithm for the Ki-67 index estimation. This algorithm is evaluated mainly on bladder carcinoma images. We also explore the pre-processing of input images to generalize the segmentation model to breast cancer data.

II. METHODOLOGY

To compute the Ki-67 index, we need to identify the immunopositive and immunonegative nuclei from the background. Thus every pixel in the input image has to be classified into three classes (background, immunonegative/ positive). This is a semantic segmentation problem to generate pixel-wise labels. One of the most commonly used approaches for this task is U-Net.

A. U-Net Model

U-Net is a fully convolutional neural network, which gives better segmentation results, especially in medical imaging [11]. It was named from its U-shaped architecture, as shown in Fig. 1. It comprises of a contracting path, bottleneck, and an expanding path. The contracting path, also known as an encoder path reduces the spatial dimension, whereas the expanding part (decoder path) gradually recovers the object details and feature dimensions. There are skip connections between the encoder-decoder path which provide the local information to the global information while upsampling. The bottleneck region, which comprises of two convolution layers with batch normalization and drop out, connects both contracting and expanding path.

Separable convolutions are often used to reduce the complexity of convolutional neural networks (CNN). A standard convolution is factorized into a depth wise convolution and a point wise convolution [12]. In the depth wise convolution,

978-1-7281-1895-6/19/$31.00

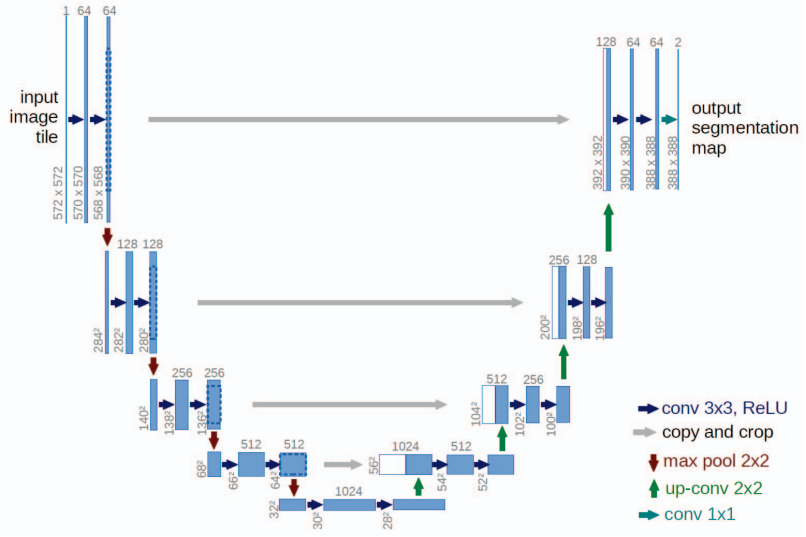


Fig. 1: U-Net Architecture [11].

each input channel is convolved separately with a single filter. In the next stage, a 1x1 convolution is performed to the output from the depthwise convolution, as shown in Fig. 2. The separable convolutions reduce the number of parameters as well as computations.

MobileUnet uses separable convolutions in the U-Net architecture. It combines the advantage of UNet for accuracy in segmentation and that of separable convolutions for its high speed and less number of parameters.

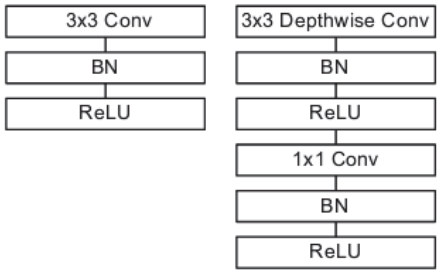


Fig. 2: Standard convolutional layer and Depth wise seperable convolutional layer [12].

B. Post processing

We obtain a three class segmented output from the MobileUnet Model. The segmented image often has a few isolated pixels. To eliminate these, we use a sequence of morphological opening and closing operations. In order to obtain the count, we need to group the pixels based on pixel connectivity. We used connected component analysis to serve this. Binary masks corresponding to immunopositive and immunonegative class are generated. Blob based extraction is applied to each of the generated masks and the count for the corresponding classes are thus obtained.

III. DATASET

This study involves 80 Ki-67 stained histology images of bladder cancer. The whole slide images were obtained using Olympus BX53C, a decahead microscope with attached camera at 40x optical magnification. Ten non-overlapping images of resolution 1920 $\times$ 1440 pixels were cropped from the hotspot regions of the whole slide image corresponding to 8 patients. All the data are collected from the Department of Pathology at Kasturba Medical College (KMC), Mangalore. Ethical approval has been taken from KMC, Mangalore (ref. no. IEC KMC MLR 01-19/35; dated January 16, 2019). All the procedures were performed according to the institutional policies.

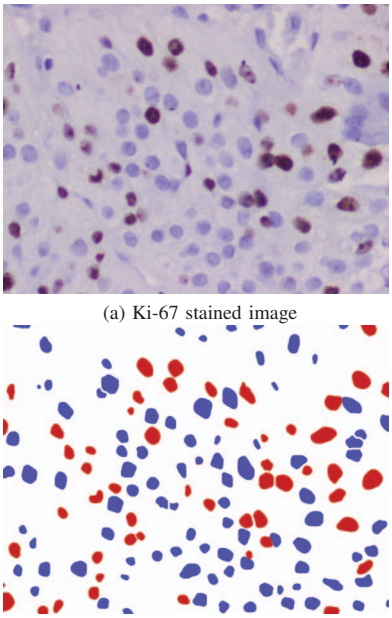
The ground truth images were prepared by manual annotations with the guidance of an expert pathologist. Ki-67 positive nuclei are marked red while negative tumor nuclei are marked blue and the background are marked white. All the images contained almost 12300 nuclei, out of which 1613 nuclei were immunopositive. Fig. 3 shows a sample Ki-67 stained image and an annotated mask corresponding to it.

IV. EXPERIMENTAL SET UP

We had 80 images of size 1920 $\times$ 1440 pixels. The entire data was divided into 20 training images and 60 testing images. Since the model that we are using is a fully convolutional neural network, it supports variable length inputs. However, for batch training, we divided the input images into patches of 512 $\times$ 512 pixels. We had kept a horizontal shift of 128 pixels and a vertical shift of 232 pixels so that each image is divided into 60 input patches. Thus we had 1200 image patches for training. For validation, 5 % of the training images were used. The test images are kept at the original size.

A. Evaluation metric

We evaluated both the segmentation performance as well as the predicted Ki-67 index. F1 score and dice score were



(a) Ki-67 stained image

(b) Annotated ground truth

Fig. 3: Sample Images

used to evaluate the segmentation performance. F1 score is the harmonic mean of precision and recall.

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{1}$$

where precision (*P*) is the ratio of true positives over the sum of true positives, and false positives and recall (*R*) is the ratio of true positives over the sum of true positives and false negatives.

Another common evaluation metric for segmentation is the dice score. Given two sets A and B, dice score is defined as the ratio of two times the number of elements common for both sets to the sum of elements in each set. It is also called as dice similarity coefficient (DSC).

$$DSC = \frac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid} \tag{2}$$

Ki-67 indices were computed by an expert pathologist as well as from the predicted images. The mean absolute error between these two values was used as the metric for evaluation.

B. Training

The MobileUnet Model was trained with cross-entropy as the loss function. The Model had 8.87M trainable parameters. The model was trained with pixel-wise one hot encoded targets corresponding to three classes. As the number of images in the training set was less, data augmentation was performed. A random horizontal flip, vertical flip, brightness alteration, and rotation were performed as part of data augmentation. We obtained pixel-wise classified images from the network. Fig. 4 shows the validation accuracy vs. epoch for training. The accuracy value reached an optimum at 72^{nd} epoch. We chose this model for testing. Maximum validation accuracy at this epoch was 0.94.

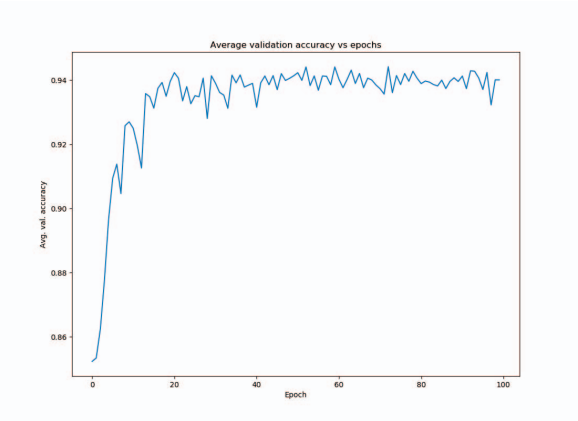


Fig. 4: Validation accuracy vs epochs

V. Results and Discussions

Table. I shows the evaluation measures of the segmentation output in the test data. The precision and recall are consistently above 0.90 for all cases except case 6. A similar trend can be seen in dice score. Fig. 5 shows the plot of F1 and Dice score for all the 60 test images. The average F1 Score is 0.92, and dice score is 0.96. Out of 60 test images, 47 images has an average dice score above 0.95.

TABLE I: Segmentation Evaluation

Cancer case	**Evaluation Measures**			
	Precision	***Recall***	***F1 score***	***Dice Score***
1	0.93	0.93	0.93	0.96
2	0.93	0.93	0.93	0.96
3	0.94	0.94	0.94	0.97
4	0.94	0.94	0.94	0.97
5	0.92	0.91	0.91	0.95
6	0.89	0.87	0.88	0.93
Average	0.93	0.92	0.92	0.96

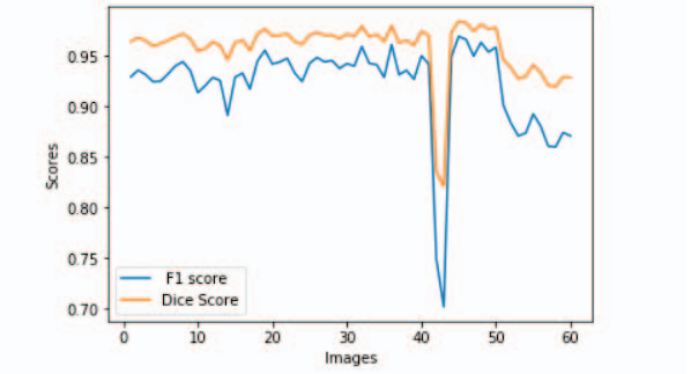
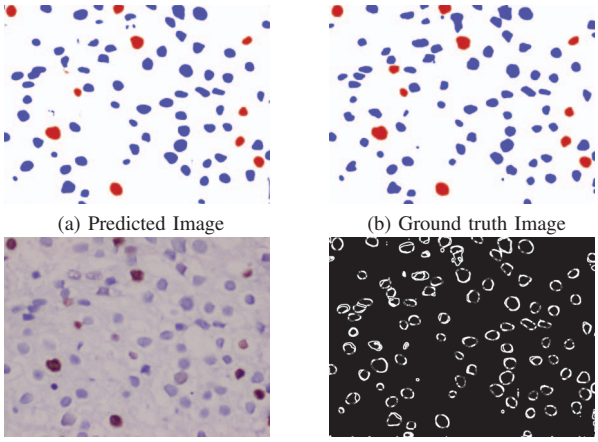


Fig. 5: Scores Vs Images

Fig. 6 shows an example input image, its ground truth and prediction, and the segmentation error. It can be observed that the segmentation errors occur mostly in the cell boundaries. This may not affect the count of the nuclei. The error image also shows that the predicted nuclei boundaries are typically smaller than the ground truth. This is because the background is the largest segment in all the images. Thus the model has a bias towards background class. This could be offset by an appropriate weighted loss function such as Tversky loss function [13].



(a) Predicted Image (b) Ground truth Image

(c) Original Image (d) Error Image

Fig. 6: An example of Prediction

From the segmentation image, we count the immunopositive and immunonegative nuclei and compute the Ki-67 index. Fig. 7 shows the plot of Ki-67 computed from the predicted image and the count provided by the pathologist.

The mean absolute error rate was found to be 2.1. Case 4 has the highest absolute error of 5.1. These images contain a large number of overlapping nuclei which the segmentation merge. This resulted in a higher error in the counts.

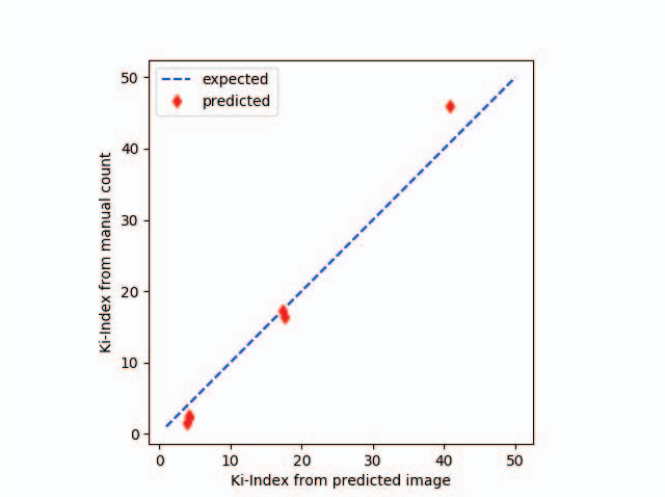


Fig. 7: Ki-67 index evaluation

A. Testing Generalization

Ki-67 index is widely used across different types of cancers for the prognostic evaluation. However, labeled image data may not be available for different cancer types. Hence it is important to use the same model for other types of cancer. In this effort, we explore a set of pre-processing steps to reuse the model trained on bladder carcinoma images to breast cancer. One whole slide data corresponding to breast cancer was available from [14]. We took 40 non-overlapping images of size 512 $\times$ 512 pixels. Manual annotations were performed to prepare the ground truth. These images were acquired using a different model of equipment. This had resulted in breast cancer images to have a different color spectrum and different nuclei size. Thus blindly using the same network resulted in poor performance.

We performed a histogram matching to compensate for the color space differences. The color histograms of each breast cancer images were mapped to the average color histogram of our training dataset [15]. The histograms corresponding to R, G, B channels for source, reference, and matched images are shown in Fig. 8. Also, the images were resized to match the average nuclei size of our data set using cubic interpolation.

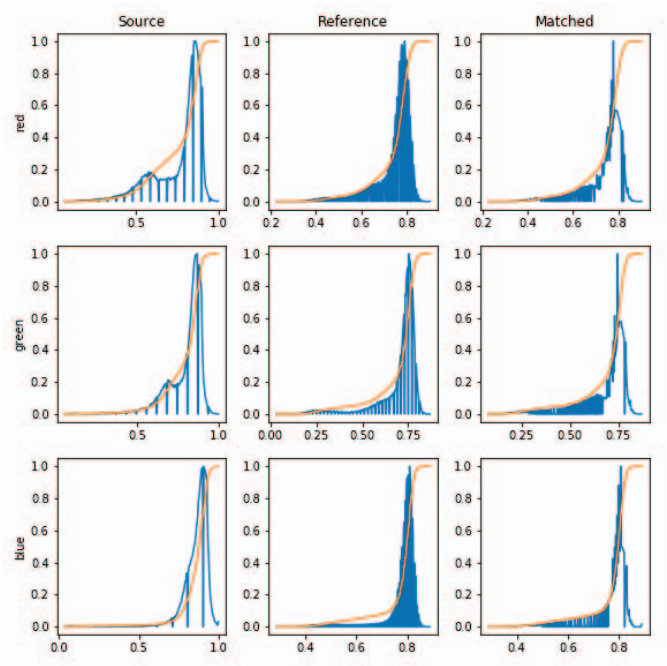
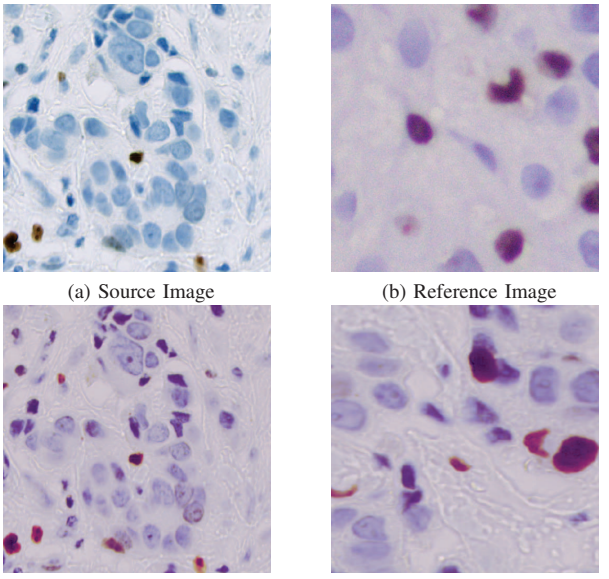


Fig. 8: Histogram plot of the source reference and matched images

Fig. 9 shows an example image from the breast cancer data set and effect of histogram matching and resizing. Table. II shows the corresponding results. It can be seen that with the histogram and size matching the F1 score has improved by 12 % and dice score has improved by 8 %. Thus the model is generalizable with minimum pre-processing for a new data set. The mean absolute error in Ki-67 index was not computed for this dataset as it was not validated by pathologists.

TABLE II: Evaluation of segmentation in breast Cancer data

Data Class	Evaluation Measures %	
	F1 Score	*Dice Score*
Breast cancer data	0.74	0.85
Histogram Matched	0.77	0.86
Histogram and size Matched	0.86	0.93



(a) Source Image (b) Reference Image

(c) Histogram Matched Image (d) Histogram & Size Matched Image

Fig. 9: Histogram Matching and resizing

VI. CONCLUSION

In this work, we proposed an integrated approach for Ki-67 index computation on carcinoma bladder data set. We used MobileUnet model for integrated segmentation and classification and connected component based algorithm to compute the Ki-67 index. The average F1 score obtained is 0.92, and the average dice score is 0.96. The mean absolute error between the predicted and the pathologist computed Ki-67 indices is 2.1. To the best of our knowledge, this is the first integrated approach based on deep learning for the computation of Ki-67 index.

This work also proposes a set of pre-processing steps such that the same model can be re-used for other types of cancer. Histogram matching and re-sizing were performed on the input breast cancer data. This resulted in an increase of 12 % in F1 score and 8 % in dice score as compared to blindly using the same model on raw data.

The error between predicted and ground truth Ki-67 indices is high in the presence of a large number of overlapping nuclei. Morphological post-processing algorithms could help to resolve the overlapping nuclei and would be addressed in future work.

ACKNOWLEDGMENT

We would like to thank Gautham Sambath, Khalender Mahfooz, Abdul Hadi and Abhijith S, B.Tech students (ECE), NITK Surathkal, India for their help and support extended to this work.

REFERENCES

[1] W. Jonat and N. Arnold, "Is the ki-67 labelling index ready for clinical use?" 2011.

[2] P. Shi, J. Zhong, J. Hong, R. Huang, K. Wang, and Y. Chen, "Automated ki-67 quantification of immunohistochemical staining image of human nasopharyngeal carcinoma xenografts," *Scientific reports*, vol. 6, p. 32127, 2016.

[3] M. A. Samols, N. E. Smith, J. M. Gerber, M. Vuica-Ross, C. D. Gocke, K. H. Burns, M. J. Borowitz, T. C. Cornish, and A. S. Duffield, "Software-automated counting of ki-67 proliferation index correlates with pathologic grade and disease progression of follicular lymphomas," *American journal of clinical pathology*, vol. 140, no. 4, pp. 579–587, 2013.

[4] V. J. Tuominen, S. Ruotoistenmäki, A. Viitanen, M. Jumppanen, and J. Isola, "Immunoratio: a publicly available web application for quantitative image analysis of estrogen receptor (er), progesterone receptor (pr), and ki-67," *Breast cancer research*, vol. 12, no. 4, p. R56, 2010.

[5] M.-K. Yeo, H. E. Kim, S. H. Kim, B. J. Chae, B. J. Song, and A. Lee, "Clinical usefulness of the free web-based image analysis application immunoratio for assessment of ki-67 labelling index in breast cancer," *Journal of clinical pathology*, vol. 70, no. 8, pp. 715–719, 2017.

[6] L. H. Tang, M. Gonen, C. Hedvat, I. M. Modlin, and D. S. Klimstra, "Objective quantification of the ki67 proliferative index in neuroendocrine tumors of the gastroenteropancreatic system: a comparison of digital image analysis with manual methods," *The American journal of surgical pathology*, vol. 36, no. 12, pp. 1761–1770, 2012.

[7] F. Xing, H. Su, and L. Yang, "An integrated framework for automatic ki-67 scoring in pancreatic neuroendocrine tumor," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2013, pp. 436–443.

[8] F. Xing, H. Su, J. Neltner, and L. Yang, "Automatic ki-67 counting using robust cell detection and online dictionary learning," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 859–870, 2014.

[9] B. Grala, T. Markiewicz, W. Kozłowski, S. Osowski, J. Słodkowska, and W. Papierz, "New automated image analysis method for the assessment of ki-67 labeling index in meningiomas." *Folia Histochemica et Cytobiologica*, vol. 47, no. 4, pp. 587–592, 2009.

[10] M. Saha, C. Chakraborty, I. Arun, R. Ahmed, and S. Chatterjee, "An advanced deep learning approach for ki-67 stained hotspot detection and proliferation rate scoring for prognostic evaluation of breast cancer," *Scientific reports*, vol. 7, no. 1, p. 3213, 2017.

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[13] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2017, pp. 379–387.

[14] "Hamamatsu ndpi." [Online]. Available: http://openslide.cs.cmu.edu/download/openslide-testdata/Hamamatsu/

[15] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006.