Binary Classification of Metastatic Lymph Nodes using Deep Learning

Minhaj Uddin Ansari
Queens University
Kingston, Canada
20mua@queensu.ca

Abstract

Lymph nodes can get swollen due to injury, infection, or the spread of cancer from other parts of the body. Cancer spreading from other parts of the body into the lymph nodes is called metastasis. Detecting metastasis early can help locate the source of cancer, mitigating the risk of it spreading throughout the body. In this project, we will use CT images of healthy and metastatic lymph nodes to train 2D CNN deep learning binary classifiers. The feasibility of our classifiers will be evaluated using test accuracy and AUC as performance metrics.

1 Introduction

Lymph nodes are small bean-like structures distributed throughout the body. They are responsible for filtering the lymph. Lymph is a clear yellowish-white fluid, originally plasma, that has escaped the capillaries into the tissues. Lymph vessels route the lymph through the lymph nodes back into the bloodstream. All the lymph nodes and lymph vessels are part of the lymphatic system.

Sometimes lymph nodes can get swollen due to injury, infection or spread of cancer cells from other parts of the body [6]. For example, early-stage breast cancer is indicated by the spread of breast cancer cells to no more than three nodes, and late-stage breast cancer happens when the cells spread to more than three nodes [1]. Doctors usually check the lymph nodes in the axillary region to understand the severity of breast cancer [9].

Cancer is the leading cause of death responsible for about 9.6 million deaths worldwide in 2018 [10], according to the World Health Organization. Researchers are busy finding solutions for detecting cancer early so that it can be stopped before it spreads throughout the body. Lymph nodes closer to the cancer region are infected first, and as cancer spreads lymph nodes further away from start becoming infected. Predicting whether a lymph node in a CT scan is infected can help locate and stop cancer, which is significant because both powerful computing and CT scan imagery is affordable and quick in giving instant and accurate results.

In this work, we will use a novel data set containing CT images of the abdomen provided by a radiologist, and our objective would be to detect colon cancer. Each CT image contains metastatic and non-metastatic lymph nodes. Lymph node images are typically stored as three-dimensional format CT scan in which a group of two-dimensional images stacked together where each image shows a view of the tissue at that location. These two-dimensional images are also known as slices or frames.

Classifying lymph nodes has been done by many researchers in the past, but there have been intrinsic challenges associated with it. Particularly, metastatic lymph nodes aren't very visible to the naked eye, unlike other types of tumors. Some authors claim that metastatic lymph nodes can be effectively characterized through radiomics features such as size and intensity, while others have used deep learning to automate the feature extraction and classification process. Since this paper primarily

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

focuses on deep CNNs for classification, we will mostly explore literature in that area but will look into a few non-CNN approaches.

Another challenge lies in the fact that lymph nodes vary by size depending on their location in the human body. This can lead to different lymph node sizes at which metastases can occur. For example, in rectal cancer, lymph node metastases occurs in nodes less than 1 cm [13] but in breast cancer, micro metastases starts off when lymph nodes near the breast have sizes between 0.2mm and 2mm [3]. In contrast, lymph node size doesn't matter for lymph node metastases in colon cancer. This makes it challenging to create a universal classifier to classify lymph node metastases from any part of the human body.

This paper aims to classify whether a lymph node CT image is metastatic or non-metastatic using 2D-CNN architectures. We were initially focusing on using both 2D-CNN and 3D-CNN, but decided to lower the scope because most of our literature review has 2D-CNN approaches. This problem is challenging because we have a new dataset that hasn't been used by anyone before, so we don't what to expect from our data and in our results. Another challenge we realized after a meeting with a radiologist is that some of the images weren't labeled properly so we had to manually look through the entire dataset and remove images we thought didn't represent lymph nodes. We still have to verify the images we removed from the dataset, and the images we've kept with the radiologist. Currently, this proposal does not consider other methods found in literature such as applying deep neural networks on radiomic features, but in the future, we may explore that space.

2 Literature Review

In our literature review, we go through some of the methods used in classifying lymph node metastasis and highlight the core contributions and results obtained in each paper. Hongkai Wang et al [14] used classical machine learning and AlexNet CNN to classify mediastinal (part of the chest) lymph node metastasis in 1397 lymph nodes. The results of AlexNet and classical methods were similar in AUC, 0.91, which meant that features extracted by the CNN were just as important as 13 diagnostic features and 82 texture features used by random forest and SVM for classification. Jeong Hoon Lee et al [8] used deep learning models to diagnose cervical (neck area) lymph node metastasis in 3838 axial CT images from thyroid cancer patients with a best AUC of 0.884 obtained from the Xception model. Li-Qiang Zhou et al [15] used RESNET-101, Inception-RESNET V2 and Inception V3 and discovered that the best CNN was the Inception V3 achieving an AUC of 0.890. Babak et al [5] assessed the diagnostic capability of 32 deep learning algorithms on detection of lymph node metastases in women with breast cancer from the research challenge competition CAMELYON16 held in 2016. The top 5 algorithms had mean AUC 0.960. The top algorithm submitted by a team from Harvard and MIT, GoogleNet, achieved an AUC of 0.994. Eleven pathologists also participated in the experiment and, combined, their mean AUC was 0.966. MuthuSubash et al [7], in a report published in Nature, used multi-layer fully connected deep networks (MFDN) for detection of metastatic lymph node from thyroid tissue resulting in a mean precision of 84.7%-85.3% and mean F1-score of 82.3%-84.4%. They also used CNNs but could only achieve moderate performance in lymph node detection. Hoa Hoang et al [11] used a two-step deep learning approach to detect lung cancer in lymph nodes. The first step eliminated frequently misclassified non-cancerous regions in the images using LFCNN (Lymphoid Follicle Convolutional Neural Network) and the second step detected cancer cells using a TDCNN (Tumor Detection Convolutional Neural Network) classifier. They achieved sensitivity of 79.6% and 96.5%, specificity of 75.5% and 98.2%, and AUC of 0.922.

Steiner et al [4] used a slightly different approach in classifying breast cancer metastases in lymph nodes. Instead of letting the algorithm or the pathologist diagnose independently, they developed an algorithm, the LYmph Node Assistant (LYNA), that helped pathologist make decisions and demonstrated that the algorithm-assisted diagnosis gave better sensitivity (91% vs 83%) and achieved an AUC of 0.99. A similar approach involving pathologist feedback was used by M. Sadeghi et al [12]. In their research, they first trained a CNN model to classify lymph node metastases in breast cancer. The CNN itself achieved a 97.8% on the validation set. Then two pathologist used a GUI software to manually classify lymph node metastases. The images that were misclassified by the algorithm as either false positive or false negative were augmented, added into the training set and used to retrain the CNN. This feedback technique improved the 25% quantile of the probability score of prediction from 0.8 to 0.89.

Paper	Best Model	Metric	Results
Hongkai Wang et al	AlexNet	AUC	0.910
Jeong Hoon Lee et al	Xception	AUC	0.884
Li-Qiang Zhou et al	Inception V3	AUC	0.890
Babak et al	GoogleNet	AUC	0.994
MuthuSubash et al	MFDN	F1-Score	82.3-84.4
Hoa Hoang et al et al	LFCNN + TDCNN	AUC	0.922
Steiner et al	LYNA	AUC	0.99
M. Sadeghi et al	Custom CNN	Accuracy	0.978

Table 1: Comparison of Literature

We will use some of the deep learning algorithms used by authors in our literature review such as pre-trained off-the-shelf models. We will also test our own custom CNN model on the dataset. Our work is different from previous works because we are using a new dataset that has been provided by a radiologist. The dataset is unfiltered and contains many bad labels that we need to investigate and remove before training our model. The CNN models we use will be very similar to the models used by authors in our literature review.

3 Approach

Our methodology involves four major steps highlighted in points below:

1. Data exploration.
2. Data preprocessing.
3. Training a 2D-CNN model
4. Results and Fine tuning

3.1 Data exploration

The dataset we used is a private dataset provided by the Simpsons Lab at Queen's University, Canada. It contains 567 3D CT scans of the lymph node area in the abdomen region with 217 positive cases and 255 negative cases. The dataset also contains binary masks that correspond to the region of the lymph nodes in each CT image.

We explore the data in two ways: qualitatively and statistically. In qualitative analysis, the goal was to understand the format and visual features of the images in the dataset. In our case, the dataset was in CT format. CT imagery is usually done in Hounsfield units (HU). Water has a HU value of zero. Tissues denser than water have positive HU values and tissues lighter than water have negative HU values. Denser tissues appear brighter than lighter tissues. Since there is a huge range of density values, and the human eye can only discern a small number of gray levels [2], it becomes necessary to map a small window taken from the entire range of HU values onto the visible gray levels. Different parts of the human body have different window ranges. Since we are dealing with CT images of the abdomen, our window size was between -135 HU and 215 HU. This ensured the abdomen image in the CT scan was displayed and processed as a proper gray image on the computer.

Observing the images in the dataset involved locating features that may be apparent either directly or through extra processes such as intensity manipulation, intensity thresholding, or edge outlining. The goal was to understand the data as much as possible and to make sure that mislabeled data was removed. With this process, we identified a few images that were empty, and a few others that didn't have proper binary masks. In removing them, our dataset reduced from 567 to 481 CT images. We also had a meeting with the radiologist back in March and he pointed out by looking at only a few examples that our dataset contained mislabeled data, so we manually removed some images we believed didn't represent lymph nodes further reducing the dataset size from 481 to 385 CT images. We also noticed that the CT of the abdomen was very large and the binary masks of the lymph nodes were a very small part of the abdomen.

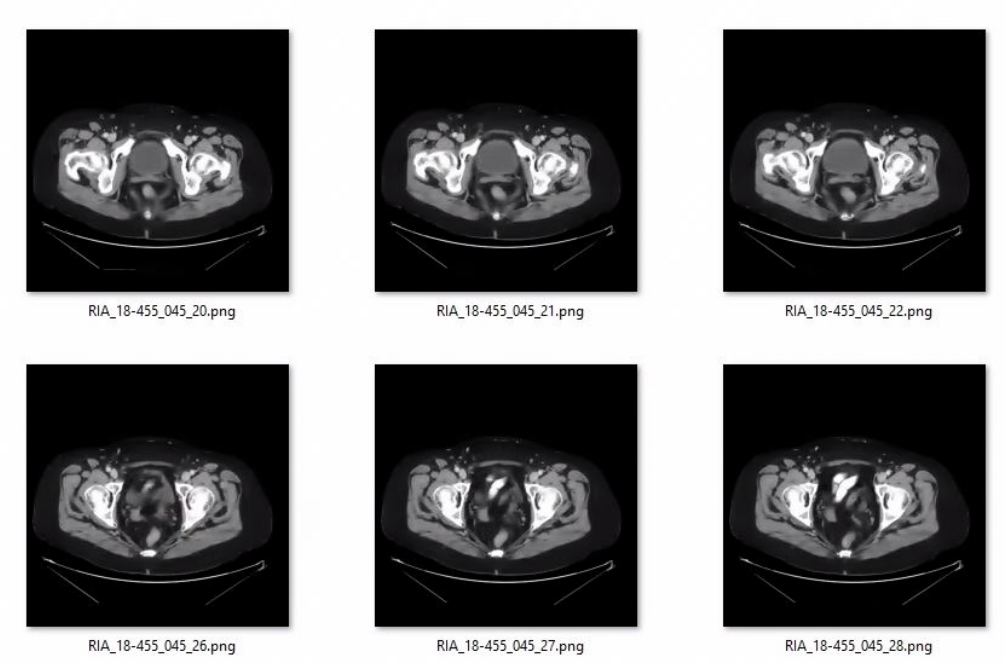

Figure 1: Slides in a CT image

In statistical analysis, the goal was to understand the statistical aspects of the dataset. This involved finding the dimensions of the slices; the spacing in between pixels; mean, minimum, and maximum pixel intensities; and the number of slices in each image. Histograms were plotted to graphically display the frequency distribution of the spacing and number of slices. We'll explore in detail the role of each statistical value in the data preprocessing subsection.

3.2 Data preprocessing

Before the data was sent to a model for training, it had to be preprocessed. In the previous step, we plotted frequency histograms of the spacing and number of slices and found the intensity values and dimensions of each image. We used some of this information in three preprocessing techniques: resampling, cropping the lymph nodes from the CT scans, and augmenting the cropped images.

3.2.1 Resampling

Medical images are usually inconsistent in terms of pixel-wise distance in the x,y, and z-axis. This depends on the type and settings of the machine used to generate the CT scan. A resampling function fixes the spacing to specific values, in our case the most frequent spacing (0.98 mm, 0.98 mm, 5.0 mm), passed to it as an argument. In other words, it levels the resolution while preserving the contours of all the images in the dataset. In the resampler, linear interpolation was used for raw images and B-Spline interpolation was be used for the binary masks. We later discovered that (0.98 mm, 0.98 mm, 5.0 mm) resulted in losing slice information since the 5.0 mm in the z direction was very large, so we lowered the spacing to (0.79 mm, 0.79 mm, 0.79 mm).

3.2.2 Cropping Lymph Nodes

We cropped lymph nodes from the binary masks by placing a bounding box with the lymph node at the center, and then use the bounding box coordinates to crop lymph nodes from the raw image. The size of the bounding box was picked such that it encloses the most number of lymph nodes. Since the bounding box size needs to be consistent, we disregarded any lymph node that was at the edge of the image. The data obtained contained 307 slices of negative cases and 301 slices of positive cases. From this, 10% or 62 images were used for validation, 10% or 62 images for testing and the remaining 80% or 484 images for training.

3.2.3 Data Augmentation

The dataset was augmented by vertically and horizontally flipping the images, randomly rotating in the range 0-180 degrees and horizontally and vertically translating by 10 pixels in both directions. Only the data reserved for training was augmented. This increased the number of images from 484 to 947 (466 positive images and 481 negative images).

3.3 2D-CNN Model

A CNN is constructed using a combination of convolution and activation, pooling, and fully connected layers. These weights are initialized randomly.

Layer (type)	Output Shape
(Conv2D)	(None, 224, 224, 32)
(MaxPooling)	(None, 112, 112, 32)
(Dropout)	(None, 112, 112, 32)
(Conv2D)	(None, 112, 112, 64)
(MaxPooling)	(None, 56, 56, 64)
(Dropout)	(None, 56, 56, 64)
(Conv2D)	(None, 56, 56, 128)
(MaxPooling)	(None, 28, 28, 128)
(Dropout)	(None, 28, 28, 128)
(Flatten)	(None, 100352)
(Dense)	(None, 128)
(Dropout)	(None, 128)
(Dense)	(None, 1)

Table 2: Custom 2D CNN Architecture

We used ReLU activation in our convolution layers and sigmoid activation in our final layer since this was a single class binary classification problem. Max pooling was used to reduce the dimensions of the feature maps.

4 Experimental Setup

Three experiments were performed. The first experiment used a custom CNN on images that weren't augmented, the second experiment used the same CNN on augmented data and the third experiment used a pre-trained RESNET-50 on images that weren't augmented since augmenting the images resulted in poorer performance which indicated that these type of medical images were highly sensitive to any type of artificial variation.

We used Binary Cross-Entropy Loss because this loss function is the most widely used for classification tasks and Adam Optimizer for optimization because it has adaptive learning rates, is fast, and is also very popular in deep learning. The batch size was set to 4, epochs to 30, and an early stopping was set to stop training if the validation loss did not decrease after 30 epochs. We kept the stopping criteria high for experimental purposed just to see how the model performed in the long run. As mentioned before, the datatset was be divided into 80% training, 10% testing, and 10% validation sets. NVIDIA GTX 1660 Ti was used for training the model. Training time for non-augmented data was around 1 minute and for augmented data around 2 minutes.

5 Results

The performance metrics that were evaluated were the training loss, training accuracy, validation loss, validation accuracy, test accuracy and test AUC. Ideally, we wanted to see a decrease in both training and validation losses, and an increase in training and validation accuracy's, with the validation accuracy being slightly lower than the training accuracy. However, that wasn't the case. Our train loss was decreasing but the validation loss was increasing. Our train accuracy was increasing, but the validation accuracy was stagnant in the range 0.5-0.75. This indicated that the model wasn't

learning anything from the data. It was only memorizing the training data and overfitting. The test accuracy and test AUC in all three experiments was also below 0.7. We took all measures of preventing overfitting such as regularization, adding dropout layers and augmentation. Surprisingly, augmentation increased the dataset size but decreased the validation accuracy which indicated that our dataset was highly sensitive to artificial adjustments. We also tried different values for the hyper-parameters but the problem remained.

Few reasons we believe were the cause of the results:

- When we had a meeting with the radiologist in March, we showed him some pictures from the dataset we thought were odd and he admitted those images didn't represent lymph nodes. Having bad data can hinder the learning of a model. We then examined the entire dataset and manually separated images we thought didn't represent lymph nodes. Since we weren't experts in radiology, and we couldn't take a radiologists time to examine the entire dataset with us, we couldn't confidently assume that the images we filtered out were correct.
- Point 1 was also reinforced from the fact that we trained our CNN model on images of cats and dogs. The model was able to learn on those new images, with a slowly increasing trend in the train and validation accuracy. This also indicated that our data might be bad.
- From the literature we've reviewed, the authors used large datasets to train their models. Unfortunately for us, since we manually filtered out so many images, we were left with only 385 CT images.
- Pretrained models failed to work on our dataset because they were trained on images from Imagenet. The images in Imagenet belong to a different domain, and transfer learning from that domain to CT medical imaging domain isn't that effective. Adding points 1, 2, and 3 further made training on pretrained models difficult.

We also used GRAD CAMs to understand where our CNN model was focusing on when making predictions on an image. The results of the GRAD CAM showed that the model wasn't focusing on the correct part of the image i.e. the lymph node, rather it was focusing on some other aspect of the image when making predictions. This demonstrated that our model hadn't learnt the data properly.

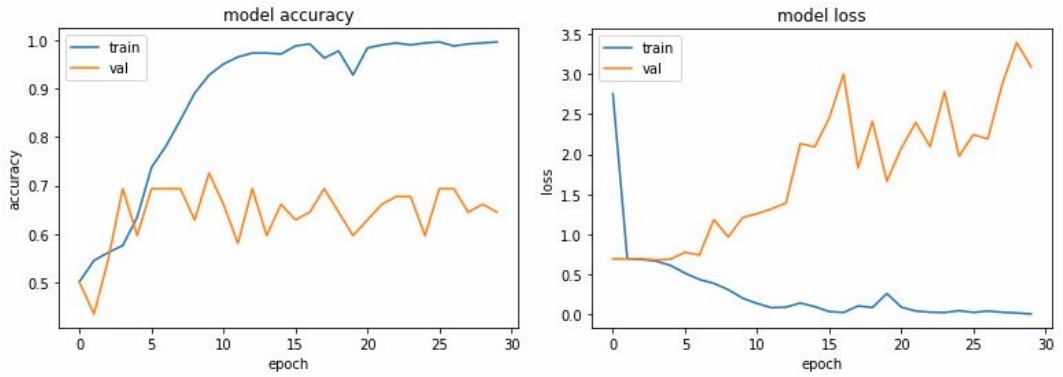

Figure 2: Accuracy, Loss and AUC of custom 2D CNN trained on original data

Algorithm	Train Acc	Test Acc	Test AUC
Custom CNN (No Aug)	0.99	0.69	0.42
Custom CNN (Aug)	0.99	0.58	0.63
Resnet-50 (No Aug)	0.99	0.61	0.53

Table 3: Results: Train Accuracy, Test Accuracy and Test AUC

6 Summary and Conclusion

After extensive work on this project, the final results weren't as great as we'd hoped. The performance metrics on the test data were below average thereby not suitable for any clinical purpose. Here we

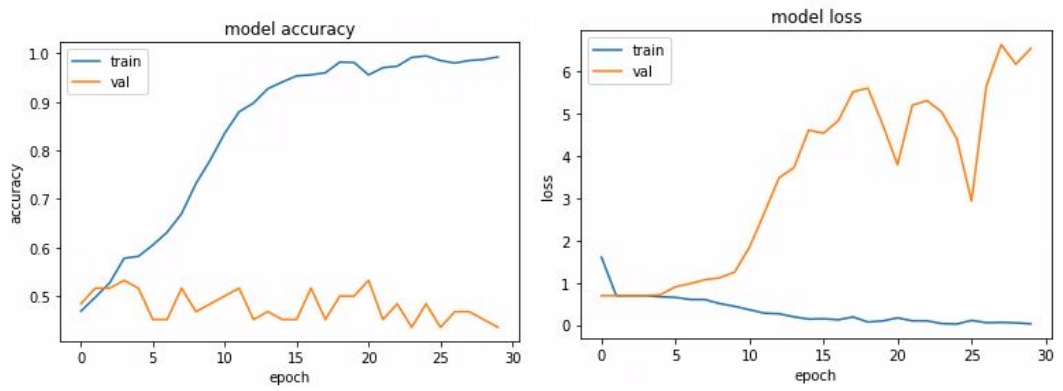

Figure 3: Accuracy, Loss and AUC of custom 2D CNN trained on augmented data

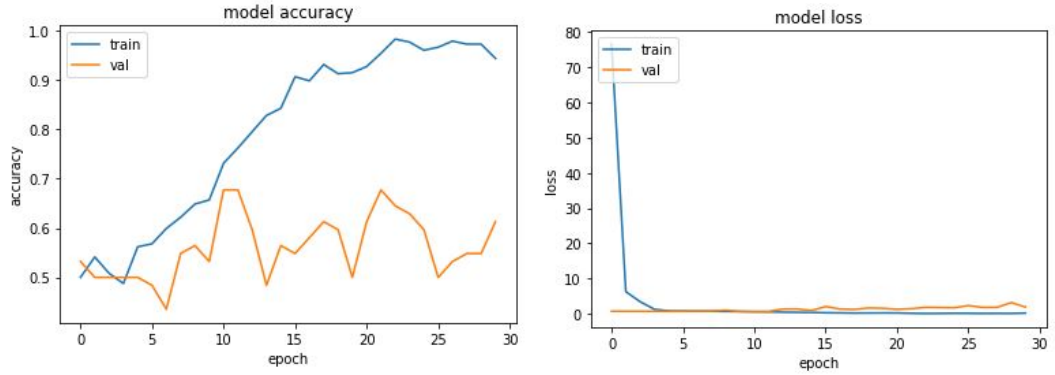

Figure 4: Accuracy, Loss and AUC of RESNET-50 trained on original data

summarize our entire work. Our task was to classify metastatic and non-metastatic lymph nodes in CT images of the abdomen using Convolution Neural Networks. The data provided to us by a radiologist was not clean, so we had to explore and preprocess it before we could train a model. Preprocessing contained standard image processing techniques we learnt in our medical imaging course last term, and all steps were discussed with people from the Simpsons Lab that had a background in medical image analysis prior to execution. The training and testing of the CNN was similar to the approaches taken by authors in the literature we cited, i.e. take a custom or pre-trained CNN model and train on the data. However our results failed to meet the success criteria. The accuracy's were below 0.75 for all the algorithms we tried, which was no where near the accuracy's achieved by authors in our cited literature.

We believe the problem pertains to our dataset with reasons detailed in the results section of this report. This was a side project given to us by our supervisor, Dr. Amber Simpson, and in the duration of the deep learning course, we were able to pinpoint the probable causes of our unsuccessful results, which was an accomplishment for us. This deep learning course also helped us dive deeper into the workings of CNN, implement custom CNNs using Keras and Pytorch, and tweak the hyperparameters to see how they affected the learning and results. Future work will include studying the dataset in detail, and finding a radiologist who's willing to dedicate their time to help us sort out the dataset.

References

[1] "Stages of breast cancer," https://www.cancer.ca/.

[2] J. Broder and R. Preston, "Chapter 1 - imaging the head and brain," in *Diagnostic Imaging for the Emergency Physician*, J. Broder, Ed. Saint Louis: W.B. Saunders, 2011, pp. 1–45.

[3] D. Dabbs, M. Fung, D. Landsittel, K. McManus, and R. Johnson, "Sentinel lymph node micrometastasis as a predictor of axillary tumor burden," *Breast J.*, pp. 101–5, 2004.

[4] S. DF, R. MacDonald, Y. Liu, P. Truszkowski, J. Hipp, C. Gammage, F. Thng, L. Peng, and M. Stumpe, "Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer," *The American Journal of Surgery Pathology*, pp. 1636–1646, 2018.

[5] B. Ehteshami Bejnordi, M. Veta, P. Johannes van Diest, B. van Ginneken, N. Karssemeijer, G. Litjens, J. A. W. M. van der Laak, , and the CAMELYON16 Consortium, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *JAMA*, vol. 318, pp. 2199–2210, 2017.

[6] R.-C. Ji, "Lymph nodes and cancer metastasis: New perspectives on the role of intranodal lymphatic sinuses," *International Journal of Molecular Sciences*, vol. 18, p. 51, 12 2016.

[7] M. S. Kavitha, C.-H. Lee, S. Kattakkalil Subhashdas, T. Kurita, and B.-C. Ahn, "Deep learning enables automated localization of the metastatic lymph node for thyroid cancer on 131i post-ablation whole-body planar scans," *Scientific Reports*, vol. 10, p. 7738, 05 2020.

[8] J. H. Lee, E. Ha, D. Kim, Y. Jung, S. Heo, Y.-H. Jang, S. An, and K. Lee, "Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with ct: external validation and clinical utility for resident training," *European Radiology*, vol. 30, 02 2020.

[9] M. Nouh, N. E.-D. H Ismail, and M. El-Bolkainy, "Lymph node metastasis in breast carcinoma: clinicopathological correlations in 3747 patients," *J Egypt Natl Canc Inst*, pp. 50–6, 2004.

[10] W. H. Organization *et al.*, "Global health observatory. geneva: World health organization; 2018," *Available: who. int/gho/database/en/.[Accessed 10 July 2019]*, 2018.

[11] H. H. N. Pham, M. Futakuchi, A. Bychkov, T. Furukawa, K. Kuroda, and J. Fukuoka, "Detection of lung cancer lymph node metastases from whole-slide histopathologic images using a two-step deep learning approach," *The American Journal of Pathology*, vol. 189, no. 12, pp. 2428–2439, 2019.

[12] M. Sadeghi, N. A. I Maldonado, A. B. J Haybaeck, and M. F. P Poudel, "Feedback-based self-improving cnn algorithm for breast cancer lymph node metastasis detection in real clinical environment," *Annu Int Conf IEEE Eng Med Biol Soc*, pp. 7212–7125, 2019.

[13] G. Skandadas and K. Dow-Mu, "Nodal staging," *Cancer Imaging*, pp. 104–111, 2009.

[14] H. Wang, Z. Zhou, Y. Li, Z. Chen, P. Lu, W. Wang, W. Liu, and L. Yu, "Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18f-fdg pet/ct images," *EJNMMI Research*, 2017.

[15] L.-Q. Zhou, X.-L. Wu, S.-Y. Huang, G.-G. Wu, H.-R. Ye, Q. Wei, L.-Y. Bao, Y.-B. Deng, X.-R. Li, X.-W. Cui, and C. F. Dietrich, "Lymph node metastasis prediction from primary breast cancer us images using deep learning," *Radiology*, vol. 294, no. 1, pp. 19–28, 2020.

[16] Z. Zhou, L. Chen, D. Sher, Q. Zhang, J. Shah, N. L. Pham, S. Jiang, and J. Wang, "Predicting lymph node metastasis in head and neck cancer by combining many-objective radiomics and 3-dimensioal convolutional neural network through evidential reasoning*," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2018, pp. 1–4.