2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Multi-modal Broad Learning System for Medical Image and Text-based Classification

Yanhong Zhou[1], Jie Du[1*], Kai Guan [1], Tianfu Wang [1]

***Abstract*— Automatic classification of medical images plays an essential role in computer-aided diagnosis. However, the medical images arise from the small number of available data and the improvement of existing data-enhancement methods are limited. In order to fulfil this demand, a Multi-Modal Broad Learning System (M^2-BLS) is proposed, which has two subnetworks for simultaneous learning of both medical images and the corresponding radiology reports. M^2-BLS provides two advantages: i) our M^2-BLS has closed-form solution and avoids iterative training, once the image feature is available; ii) benefit from the simultaneous learning of both image and text data, our M^2-BLS achieves high accuracy for medical classification. Experimental results on the publicly available datasets IU X-RAY and PEIR GROSS_895 show that our M^2-BLS highly improves the classification performance, compared to SOTA deep models that learn single-type of data information only.**

***Index Terms*— Medical Classification, Radiology Report, Simultaneous Learning, Broad Learning System**

I. Introduction

Effectively classifying medical images play an essential role in assisting clinical care and treatment due to doctors need to exam numerous medical images during the process of disease diagnosis [1],[2]. For example, radiologists read images and write textual radiology reports (e.g., Fig.1) [3] to record the findings regarding to every area of the chest. Based on these findings, physicians then give the corresponding diagnosis results accurately.

Compared with hand-designed method, the convolutional neural network (CNN) model was designed for image processing [4]. The well-known CNN-based deep models include VGG [5], ResNet [6], DenseNet [7], and so on. However, in order to construct a generalized deep model, huge amount of training data are usually very necessary due to their overfull model parameters, resulting to complex and time-consuming training process. Whereas the amount of available training data is usually very limited in medical image classification area. That is because the medical images cannot be collected from website (like nature images) to construct a large dataset and the medical experts have rare time to annotate medical images. Under these small datasets, these CNN-based deep models will suffer from overfitting [8]. In order to enrich the medical images, various data-enhancement methods are adopted [9], [10], such as rotations, flips, translations. However, these synthetic images are similar to the primal one and the improvements by using these enhancement methods are limited [11].

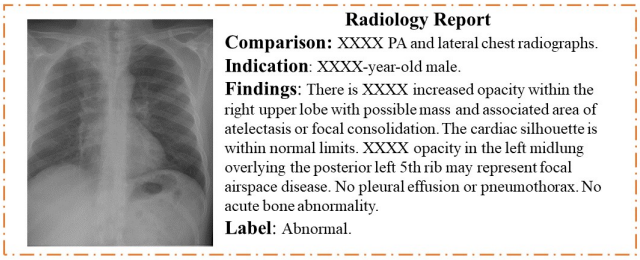

Fig. 1: An exemple of radiology report from the IU X-Ray dataset. The image on the left and the corresponding radiology report on the right. The *Comparison* section contains previous information about the patient (e.g., preceding medical exams); The *Indication* section contains reasons of examination(e.g., age); The *Findings* list the radiology observations. The *Label* section list the label of the image.

In this paper, we innovatively proposed to simultaneously learn both image and text information of medical data to overcome small data issue. In the literature, in order to learn text information, Recurrent Neural network (RNN) [12] is commonly used, which can recurrently learn words from text data but fails on long-term dependencies; LSTM [13] is an improved version of RNN, which has a more complicated structure inside, can select and adjust the transmitted information and keep long-term dependencies. However, RNN or LSTM mechanism only learn sequence information of words but lack word importance that is also very important for text classification [14]. Recently, Recurrent Board Learning System (R-BLS) [15] is proposed, which provides the way of simultaneous learning both sequence information and word importance in one network. Although R-BLS learns sufficient text information, its network cannot directly extract image information.

In this work, we proposed a Multi-Modal Broad Learning System (M^2-BLS) to simultaneously learn information from both text and image data for accurate disease diagnosis. Our M^2-BLS contains two subnetworks: one extracts text information including sequence information and word importance; another extracts image information. These information are then used to analytically determine the output. The main contributions of this work are summarized as follows:

i) Our M^2-BLS enriches the medical data by taking text data information into consideration as well as image data

This work was supported by National Natural Science Foundation of China No.62006160, Educational Commission of Guangdong Province 2020KQNCX062, Shenzhen Fundamental Research Program 20200813102946001.

*Corresponding author

[1]School of Biomedical Engineering, Shenzhen University, National-Regional Key Technology Engineering Laboratory for Medical Ultrasound,Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Shenzhen, China. dujie@szu.edu.cn

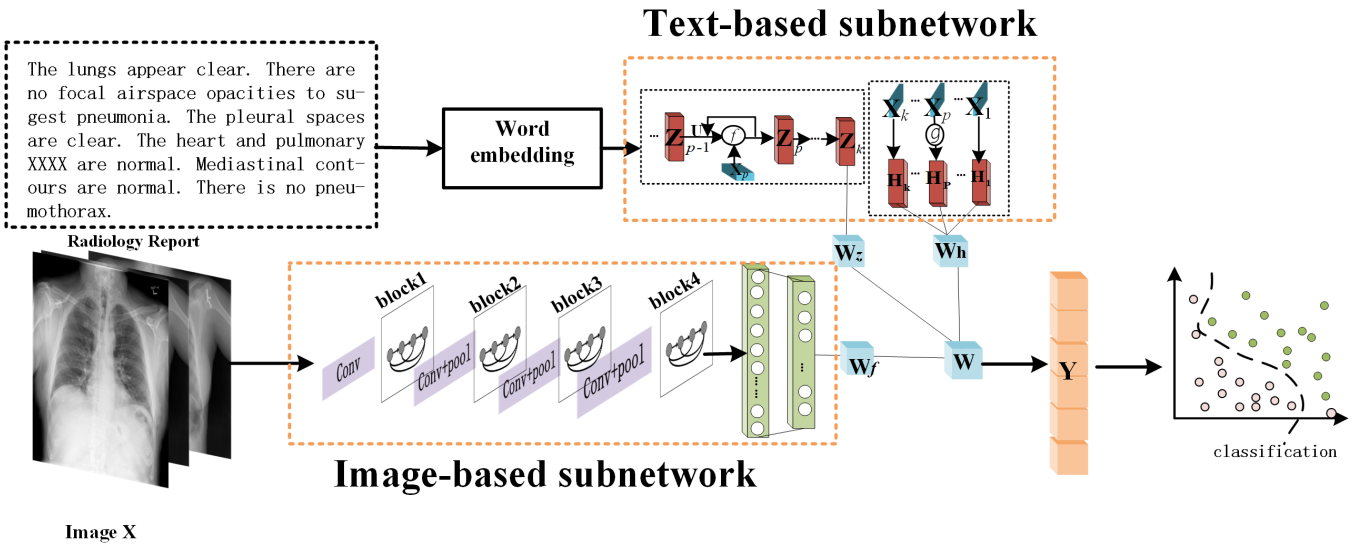

Fig. 2: The framework of the proposed M^2-BLS.

information;

ii) Our M^2-BLS has a closed-form solution that avoids the iterative training and shows high training efficiency;

iii) Through the simultaneous learning of both image and text data information, our M^2-BLS can significantly improve the testing accuracy, compared to SOTA methods learning single-type of data information.

II. PROPOSED METHOD

A. Image-based Subnetwork:

In order to extract image information, CNN-based deep model is firstly used to decode the 2d input image into 1d high-level feature representation $\mathbf{F}$ (i.e., the output of the fully connected layer). In this work, DenseNet is adopted due to its high feature extraction ability [16], as shown in Fig.2. After the training of DenseNet under image data only, $\mathbf{F}$ is then enhanced by the following non-linear operation as in Board Learning System (BLS) [17],

$$\mathbf{F}^{img} = g\left(\mathbf{F}\boldsymbol{\alpha}^{\boldsymbol{f}} + \boldsymbol{\beta}^{\boldsymbol{f}}\right) \tag{1}$$

Where g is an activation function, $\boldsymbol{\alpha}^{\boldsymbol{f}}$ and $\boldsymbol{\beta}^{\boldsymbol{f}}$ are also randomly generated as in original BLS.

However, as detailed in Section I, the medical data are usually too small to obtain a generalized CNN-based deep model and the feature representation $\mathbf{F}^{img}$ may be not discriminant enough for disease diagnosis. Hence the feature representation $\mathbf{F}^{img}$ is then concatenated with text information $\mathbf{F}^{text}$ (detailed below) to improve the discriminant of features.

B. Text-based Subnetwork:

1) Learning sequence information: Given N training data $\{\mathrm{x}^i,\mathrm{y}^i\}$, i = 1 to N. Assume $\mathrm{x}_p^i \in \mathbb{R}^d$ (p = 1 to k) is a mapped vector representation under *word2vec*. The matrix representation of the $\mathrm{x}_p^i \in \mathbb{R}^d$ word in all N samples is $\mathbf{X_p}=\left[\mathrm{x_p^1}, \mathrm{x_p^2},..., \mathrm{x_p^i},..., \mathrm{x_p^N}\right]^{\mathrm{T}} \in \mathbb{R}^{N\times d}$, $\mathbf{Y} = \left[y^i\right] \in \mathbb{R}^{N\times m}$ is the label matrix for all N samples. Similar to RNN, every $\mathbf{Z}_p$ is determined by both current input $\mathbf{X}_p$ and the previous memory $\mathbf{Z}_{p-1}$,

$$\mathbf{Z}_p = f\left(\mathbf{X}_p\boldsymbol{\alpha}^{\boldsymbol{m}} + \mathbf{Z}_{p-1}\mathbf{U} + \boldsymbol{\beta}^{\boldsymbol{m}}\right) \tag{2}$$

Where f is an activation function such as sigmoid, and the weights $\boldsymbol{\alpha}^{\boldsymbol{m}}$,$\mathbf{U}$ and bias $\boldsymbol{\beta}^{\boldsymbol{m}}$ are randomly generated (as in Broad Learning System [17]).

2) Learning word importance: In order to effectively learn word importance, in the enhancement nodes (illustrated in Fig. 2), every words $\mathbf{X}_p$ are used as input instead of memory set $\mathbf{Z}_p$, each enhancement node $\mathbf{H}_p$ is then calculated by

$$\mathbf{H}_p = g\left(\mathbf{X}_p\boldsymbol{\alpha}^{\boldsymbol{o}} + \boldsymbol{\beta}^{\boldsymbol{o}}\right) \tag{3}$$

Where g is an activation function as well, weight $\boldsymbol{\alpha}^{\boldsymbol{o}}$ and bias $\boldsymbol{\beta}^{\boldsymbol{o}}$ are also randomly generated.

In this work, Eq.(2) is used to obtain sequence information Eq.(3) and is used to calculate each word information as in R-BLS. Hence, $\mathbf{F}^{text}$ is obtained by,

$$\mathbf{F}^{text} = [\mathbf{Z}_k|\mathbf{H}_1, \cdots, \mathbf{H}_k] \tag{4}$$

Where k is the number of words in the radiology report.

3) Output Weight: In Fig.3, $\mathbf{W}_f$, $\mathbf{W}_z$, and $\mathbf{W}_h$ respectively represents the importance of image information, sequence information and word importance for final disease diagnosis. As in original BLS, $\mathbf{W}_f$, $\mathbf{W}_z$ and $\mathbf{W}_h$ are not computed separately but these weights are firstly concatenated (i.e.,$\mathbf{W} = [\mathbf{W}_f|\mathbf{W}_z|\mathbf{W}_h]$) and then determined by the ridge regression approximation of pseudo inverse of (i.e.,$\left[\mathbf{F}^{img}|\mathbf{Z}_k|\mathbf{H}_1, \cdots, \mathbf{H}_k\right]$). Under this way, the iterative update of output weight $\mathbf{W}$ is eliminated and $\mathbf{W}$ can be calculated by a closed-form solution, which significantly reduces the training time ,

$$\begin{aligned}\mathbf{W} &= \mathbf{A}^{\dagger}\mathbf{Y} \\ &= \left(\mathbf{A}^{\mathbf{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathbf{T}}\mathbf{Y}\end{aligned} \tag{5}$$

where

$$\mathbf{A} = \left[\mathbf{F}^{img}|\mathbf{Z}_k|\mathbf{H}_1, \cdots, \mathbf{H}_k\right] \tag{6}$$

By substituting Eq.(6) into Eq.(5)

$$\begin{aligned}\mathbf{W} =& \left[\mathbf{F}^{img}|\mathbf{Z}_k|\mathbf{H}_1, \cdots, \mathbf{H}_k\right]^{\dagger}\mathbf{Y} \\ =& \left(\left[\mathbf{F}^{img}|\mathbf{Z}_k|\mathbf{H}_1, \cdots, \mathbf{H}_k\right]^{\mathrm{T}}\left[\mathbf{F}^{img}|\mathbf{Z}_k|\mathbf{H}_1, \cdots, \mathbf{H}_k\right]\right)^{-1} \\ & \left[\mathbf{F}^{img}|\mathbf{Z}_k|\mathbf{H}_1, \cdots, \mathbf{H}_k\right]^{\mathrm{T}}\mathbf{Y}\end{aligned} \tag{7}$$

III. EXPERIMENTS AND RESULTS

A. Data description

IU X-RAY [1]

We have conducted experiments on the publicly available dataset called the Indiana University Chest X-Ray Collection (IU X-Ray) [18]. The dataset consists of 7470 X-rays of size 512 × 624, among which 5177 X-rays are used for training and 1294 X-rays for testing after deleting the images that are not associated with the daignostic report (detailed in Table I). Each reports consisits of four parts: *Impression*, *Indication*, *Findings*, and *Label*. In this paper, we treat the contents in *Findings* as the text information and the Medical Text indexer (MIT) annotated label as the predicted.

PEIR GROSS_895 [2]

The Pathology Education Informational Resource (PEIR) digital library is a public access image database in medical education [19]. We collected 895 images and corresponding text descriptions from it. The dataset is split to 716 train and 179 test instances (detailed in Table I).

TABLE I: Data properties of compared datasets.

Dataset	Total	Num_class	Training	Testing
IU-XRAY	6471	2	5177	1294
PEIR-GROSS_895	895	2	716	179

B. Implementation details

Before experiments, the dataset IU X-RAY is preprocessed: first, we analysis the report and extract the *Findings* content and their label corresponding to each patient. Then, the X-rays pictures and id of the patients without *Findings* content are deleted. For the dataset PEIR GROSS_895, first we collected 895 pictures and their corresponding text reports from the official website. In addition, all raw data in text form are also preprocessed to become the numeric vector representations. In detail, the *trainWordEmbedding* function in Matlab environment (Text Analytics Toolbox) is used to train a word embedding. Then with the trained word embedding, *word2vec* function is used to map words to vectors. Since the samples or sentences in the radiology report are of different lengths, every sentence is practically truncated or padded to a fixed length of L words. Last, send the text and images into two sub-networks for training. For training ,batchsize=1 due to hardward limitation.

[1]https://openi.nlm.nih.gov/faq

[2]https://peir.path.uab.edu/library/index.php?/category/106

C. Results

The experiments are conducted in two aspects: 1) M^2-BLS is compared with SOTA CNN based deep models in terms of both test accuracy and training efficiency; 2) The comparison of M^2-BLS with SOTA Natural Language Processing (NLP) methods is also conducted.

TABLE II: The comparison results of our M^2-BLS with SOTA CNN-based deep models and NLP models on dataset IU X-RAY .

	Method	ACC (%)	Time (s)
CNN	ResNet50	65.76	10894.58
	ResNet101	66.38	22280.92
	DenseNet121	68.39	13223.60
	DenseNet169	69.09	19993.54
NLP	LSTM	84.00	359.7
	R-BLS	85.70	0.61
	OURS	**87.56**	**0.87**+19993.54

TABLE III: The comparison results of our M^2-BLS with SOTA CNN-based deep models and NLP models on dataset PEIR GROSS_895.

	Method	ACC (%)	Time (s)
CNN	ResNet50	66.67	1343.18
	ResNet101	72.22	1919.82
	DenseNet121	78.28	1218.11
	DenseNet169	79.16	726.79
NLP	LSTM	85.46	7.92
	R-BLS	95.75	0.09
	OURS	**99.44**	**0.03**+726.79

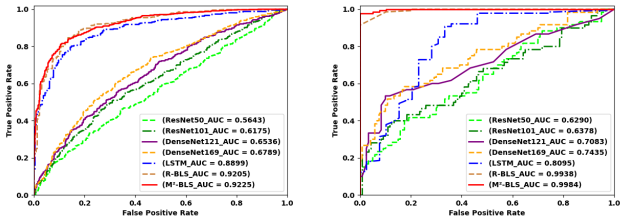

Fig. 3: ROC curves for the IU X-RAY dataset (left) and PEIR GROSS_895 dataset (right).

1) Test Accuracy: As introduced in Section I, medical data are usually very limited to construct a generalized deep model and hence the classification accuracy may be not satisfactory, which is also verified in our experiments. As illustrated in Table II&III, all SOTA CNN-based deep models get unsatisfactory classification results (about 69% on IU X-Ray and 79% on PEIR GROSS_895). On the contrary, our M^2-BLS simultaneously learns information from both image and text data to enhance the learned features and achieves 87.56% and 99.44% of accuracy on IU X-Ray and PEIR GROSS_895 respectively, even though the medical image data is a small one.

Similarly, SOTA Natural Language Processing (NLP) methods only learn information from text data, which is also insufficient for medical classification [20]. As illustrated

in Table II&III, compared to LSTM and R-BLS, our M^2-BLS also achieves best performance and improves accuracy by 3.56% and 1.86% respectively on IU X-RAY dataset and 13.98% and 3.69% respectively on PEIR GROSS_895 dataset. Whereas, the improvement is limited compared to that with CNN-based deep models. The main reason is that the DenseNet169 may suffer from overfitting due to small data issue when training and the extracted image feature is not good enough. In the future, we would like to design a deep model which can directly extract information from image and simultaneously learn both image and text data.

2) Training time : Since our M^2-BLS has closed-form solution, iterative training is eliminated. Once the image feature is extracted by deep model, our M^2-BLS only takes 0.87s and 0.03s to extract information from text data and simultaneously learns both image and text information to accurately classify samples (i.e., patients), which is a significant improvement. In summary, our M^2-BLS takes additional 0.87s and 0.03s to improve the classification accuracy of DenseNet169 by 18.47% and 20.28% on datasets IU X-RAY and PEIR GROSS_895 respectively.

Compared to NLP models, once the image feature is ready for learning, our M^2-BLS only takes additional 0.26s (i.e., 0.87s-0.61s) to improve the classification accuracy, compared to R-BLS on IU X-RAY dataset. For PEIR GROSS_895, our M^2-BLS takes less training time but improves the classification accuracy by 3.69% ,compared to R-BLS.

3) ROC curve: In Fig. 3, the ROC curve is plotted for the seven methods. It can be seen intuitively that our method is superior to other methods.

IV. CONCLUSION

In this paper, a Multi-Modal Broad Learning System (M^2-BLS) is proposed, which can simultaneously learn information from both image and text data and effectively resolve the problem of small datasets in medical image processing. In M^2-BLS, image-based subnetwork extracts image information, while text-based subnetwork simultaneously learns sequence information and word importance of text data. All extracted information are then used to analytically determine the output. Our M^2-BLS has two advantages: i) higher accuracy due to the simultaneous learning of both image and text data, even though the data is a small one; ii) faster training time due to its closed-form solution. Experiments on the dataset IU X-Ray show that our M^2-BLS additionally takes only 0.87s for training but improves the classification performance up to 21.80%. For PEIR GROSS_895 dataset, although there are only a few hundred of images, our M^2-BLS achieves 99.44% of accuracy, which only takes additional 0.03s for training.

V. COMPLIANCE WITH ETHICAL STANDARDS

We use the open source datasets. We wish to confirm that there are no known conflicts of interest associated with this publication. This research study is approved by the ethical review board of the institute.

REFERENCES

[1] Georgina Cosma, David Brown, Matthew Archer, Masood Khan, and A Graham Pockley, "A survey on computational intelligence approaches for predictive modeling in prostate cancer," *Expert Systems with Applications*, vol. 70, pp. 1–19, 2017.

[2] Weibin Wang, Dong Liang, Qingqing Chen, Yutaro Iwamoto, Xian-Hua Han, Qiaowei Zhang, Hongjie Hu, Lanfen Lin, and Yen-Wei Chen, "Medical image classification using deep learning," in *Deep Learning in Healthcare*, pp. 33–51. Springer, 2020.

[3] Changhwan Lee, Jongseong Jang, Seunghun Lee, Young Soo Kim, Hang Joon Jo, and Yeesuk Kim, "Classification of femur fracture in pelvic x-ray images using meta-learned deep neural network," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[4] Quande Liu, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng, "Semi-supervised medical image classification with relation-driven self-ensembling model," *IEEE Transactions on Medical Imaging*, vol. 39, no. 11, pp. 3429–3440, 2020.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computer Science*, 2014.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[7] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[8] Fei Gao, Hyunsoo Yoon, Teresa Wu, and Xianghua Chu, "A feature transfer enabled multi-task deep learning model on medical imaging," *Expert Systems with Application*, vol. 143, no. Apr., pp. 112957.1–112957.11, 2020.

[9] Jun Ding, Bo Chen, Hongwei Liu, and Mengyuan Huang, "Convolutional neural network with data augmentation for sar target recognition," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 3, pp. 364–368, 2016.

[10] Lei Yang, Xiaorong Wang, M. A. Fangjing, Junxi Gao, and Department Of Ultrasound, "Value of contrast-enhanced ultrasound in differential diagnosis of benign and malignant breast lesions," *Journal of Clinical Ultrasound in Medicine*, 2019.

[11] Zhiwen Huang, Xingxing Zhu, Mingyue Ding, and Xuming Zhang, "Medical image classification using a light-weighted hybrid neural network based on pcanet and densenet," *IEEE Access*, vol. 8, pp. 24697–24712, 2020.

[12] Jeffrey L Elman, "Finding structure in time," *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[13] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] Shuang Yang and Yan Tang, "Text classification based on convolutional neural network and attention model," in *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE, 2020, pp. 67–73.

[15] Jie Du, Chi-Man Vong, and CL Philip Chen, "Novel efficient rnn and lstm-like architectures: Recurrent and gated broad learning systems and their applications for text classification," *IEEE Transactions on Cybernetics*, 2020.

[16] Jeyaprakash Hemalatha, S Abijah Roseline, Subbiah Geetha, Seifedine Kadry, and Robertas Damaševičius, "An efficient densenet-based deep learning model for malware detection," *Entropy*, vol. 23, no. 3, pp. 344, 2021.

[17] CL Philip Chen and Zhulin Liu, "Broad learning system: An effective and efficient incremental learning system without the need for deep architecture," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 1, pp. 10–24, 2017.

[18] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.

[19] Baoyu Jing, Pengtao Xie, and Eric Xing, "On the automatic generation of medical imaging reports," *arXiv preprint arXiv:1711.08195*, 2017.

[20] Pratiksha R. Deshmukh and Rashmi Phalnikar, "Anatomic stage extraction from medical reports of breast cancer patients using natural language processing," *Health and Technology*, , no. 12, 2020.