Deep Learning Analytics for Diagnostic Support of Breast Cancer Disease Management

Tiancheng He, Mamta Puppala, Richard Ogunti, James J. Mancuso, Xiaohui Yu, Shenyi Chen, Jenny C. Chang, Tejal A. Patel, Stephen T.C. Wong

Abstract —**Breast cancer continues to be one of the leading causes of cancer death among women. Mammogram is the standard of care for screening and diagnosis of breast cancer. The American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) lexicon to standardize mammographic reporting to assess cancer risk and facilitate biopsy decision-making. However, because substantial inter-observer variability remains in the application of the BI-RADS lexicon, including inappropriate term usage and missing data, current biopsy decision-making accuracy using the unstructured free text or semi-structured reports varies greatly. Hence, incorporating novel and accurate technique into breast cancer decision-making data is critical. Here, we combined natural language processing and deep learning methods to develop an analytic model that targets well-characterized and defined specific breast suspicious patient subgroups rather than a broad heterogeneous group for diagnostic support of breast cancer management.**

I. Introduction

Breast cancer remains a leading cause of cancer death among women worldwide despite the advances that have been made in the identification of prognostic and predictive markers for breast cancer treatment. Mammographic reporting is the first step in the screening and diagnosis of breast cancer [1, 2]. Abnormal mammographic findings such as a mass, abnormal calcifications, architectural distortion, and asymmetric density can lead to a cancer diagnosis. The American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) lexicon, standardizing mammographic reporting to facilitate biopsy decision-making [2]. However, substantial inter-observer variability remains in the application of the BI-RADS lexicon, including improper term usage and missing data. This observer variability has led in part to considerable variation in the rate of biopsy across the US, with the vast majority of breast biopsies ultimately found to be benign lesions. Hence, there is the need for a method that can better stratify the risk of cancer and define an optimal threshold for biopsy.

The importance of risk stratifying cancer patients has led many research teams to focus on the application of machine learning methods. These techniques have also been utilized to model the risk of developing cancer and the progression and treatment of the disease. In addition, the ability of machine learning tools to detect key features from complex datasets reveals their significance in this modern era of personalized medicine. A variety of these techniques, including Bayesian Networks (BNs), Artificial Neural Networks (ANNs), and Support Vector Machines (SVMs) have been widely applied in clinical informatics research for the development of predictive models for clinical decision support systems. A number of these models for predicting the risk of breast cancer from mammographic features and demographic factors have been published [3-5]. While these studies suggest that more accurate decisions could be made using the probability of cancer as an outcome measure, these models only partly address the problem because they do not provide an optimal threshold for the decision to biopsy and most lack external clinical validations or prospective clinical evaluations.

In this paper, we present a novel analytic technique that outputs a risk assessment measure to aid radiologists and oncologists in breast cancer biopsy assessment and decisions for diagnostic support of breast cancer management. Our proposed technique first employs natural language processing method to extract pertinent clinical risk information related to breast cancer from vast amounts of radiologist-generated reports. Then it incorporates a new analytic model that deploys deep learning, an advanced subset of machine learning that is constructed from a set of neural networks that consists of three or more layers. By measuring similarities between the objects, a deep learning algorithm can find all of the objects with similar features, and, conversely, distinguish the different classes. The advantage of deep learning over previous neural networks and other machine learning methods is its capacity to extrapolate new features from a limited set of features contained in a training set. That is, it will search for and find other features that correlate to those already known and will discover new ways of detecting signals from noisy data. The ability of deep learning to generate new features without being explicitly instructed means that researchers working on risk assessment modeling can save time on feature selection and work with richer, more complex, and more comprehensive feature sets [6, 7].

To implement our study, we searched vast amounts of patient data and reports archived in our hospital clinical data warehouse. The validation results show that the deep learning algorithm used in our biopsy assessment and analytic model can help to accurately identify more subtle patterns. This will not only generate an optimal threshold for the decision to biopsy but also improve accuracy of further diagnosis support on the breast suspicious patient management.

This paper is organized as follows: Section II presents the mathematical methods of our prediction model. Section III shows the implementation results and the performance analysis. Finally, conclusions are provided in Section IV.

Tiancheng He, Mamta Puppala, Richard Ogunti, James J. Mancuso, Xiaohui Yu, Shenyi Chen, Stephen T.C. Wong are with the Systems Medicine and Bioengineering Department of Houston Methodist Research Institute and Informatics Development Department of Houston Methodist Hospital, Houston, TX 77030 USA.

Jenny C. Chang and Tejal A. Patel are with Houston Methodist Cancer Center of Houston Methodist Hospital, Houston, TX 77030 USA.

Corresponding author: Stephen T.C. Wong phone: 713-441-5884; fax: 713-441-7189; e-mail: stwong@houstonmethodist.org.

978-1-5090-4179-4/17/$31.00 ©2017 IEEE

II. METHOD

A. Pre-procedure

For building a better assessment model, the important pre-procedure is accurately extracting the cancer-related information from clinical reports. Researchers and cancer registrars generally depend on manual extraction of information, which is a time-intensive and costly process. Therefore, we developed the natural language processing (NLP) solution tool, named MOTTE, to automatically extract clinical report and patient demographics variables data for the further modeling [8-10]. Such an NLP tool can search and retrieve specific clinical information from free-text reports. In our method, mammographic reports are collected associated with breast biopsies that are referred for evaluation of a cancer prognosis. Normally, different clinical information is confined to specific sections of the report. Hence, we need to determine the general structure of the reports and recognize the data required from each section.

The framework of MOTTE as pre-procedure contained three steps, i.e. tokenization, stemming, and stop words removal. It is built on Stanford CoreNLP library [11], which integrates all Stanford NLP tools, including the part-of-speech (POS) tagger, the named entity recognizer (NER), the parser, the co-reference resolution system, and the sentiment analysis tools, and provides model files for analysis of English. Figure 1 illustrates the flow chart of the framework. The first step involves tokenization that translates to turn each clinical free text or semi-structured report into a stream of tokens. We used the following Bayesian model to process.

$$P(t|s) = \frac{P(t \cap s)}{P(s)} = \frac{P(t)}{P(t)P(m)} = \frac{1}{P(m)}$$

where t is the processed token, s is the input sentence, and m is the marks to be removed from the sentence. It is obvious that a sentence s is combined by the token t and removed marks m, i.e. $P(s) = P(t)P(m)$. In our method, m represents all punctuations and capitalizations in the text reports. We define a token t as a string of alphanumeric characters surrounded by white space.

The second step adopted was stemming. In the reports, one word always has different grammatical forms, and we only used the word's ground form in our method. For each token t, we used the following stem class model to remove the phase affixes and other lexical components and then refresh the remaining stem as the new token,

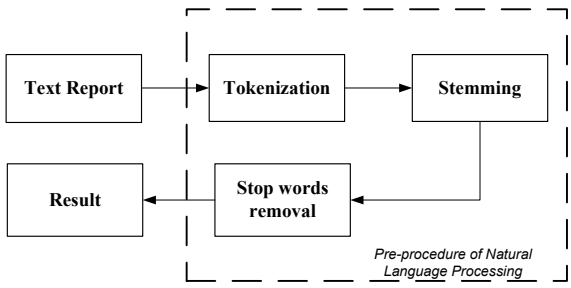

Figure 1. The flow chart of MOTTE-NLP framework for pre-procedure from free text clinical reports.

$$P(w|t) = \sum_{w_i \in E(w)} P(w_i|t),$$

where $E(w)$ is the scope of grammatical forms for word w_i, and w is ground form of token t. We also transfer the verbs, such as was or are, into the ground form, i.e. be.

The third step involves removing the stop words. The stop words in our method are words that do not carry any useful information for the text report analysis, such as the, this, which, etc. Most stop words are conjunctions, pronouns or prepositions. Hence, $P(B|w) = P(w)/P(X)$, where X is the stop word and B is the processed reports for modeling. A list of stop words was defined according to the previous NLP works, and we then used the list for stop words removal.

By exploring the language used in the mammographic reports and developing a query strategy, we extracted cancer-related information and then collaborated with the Houston Methodist Cancer Center to use the extracted data for breast cancer biopsy assessment modeling. Such an assessment created an estimate score that allows a clinician to stratify the patients by developing a more refined threshold for biopsy and safely reduces false-positive biopsies or unnecessary biopsies.

B. Deep Learning Model Architecture

After the pre-procedure, clinical reports from BI-RADS patients are converted to a structured dataset in order to build up the deep learning model, which would be organized as a report vector matrix. The rows of this vector matrix are the vector representations of each word in the report. In our method, we use the published model, i.e. word2vec model from Mikolov et al. [12], to represent each word in the report. We also define the dimensionality of the vector representations as d. If the number of vector representations vectors in the matrix is n, then the dimensionality of the report vector matrix is $d \times n$. The same zero-padding strategy [13] is then used to ensure the dimensionalities of the report vector matrices from all BI-RADS patients are identical.

After using the work of Collobert and Weston [14], we then treat the report vector matrix as a two-dimensional signal and perform convolution on it via linear filters. The rows represent discrete words in the report, so we use filters with widths equal to the dimensionality of the report vectors (i.e., d). According to the work of Ye [15], the number of adjacent rows considered jointly is defined as the height of the filter, and the height value varies according to the number of words in the report. We refer to the height of the filter as the region size of the filter. Suppose that there is a filter parameterized by the weight matrix $\mathbf{w}$ with region size h; $\mathbf{w}$ will contain $h \times d$ parameters to be estimated. The feature vector is denoted as the matrix by $\mathbf{F} \in \mathbb{R}^{d \times n}$, and we use $\mathbf{F}[i{:}j]$ to represent the sub-matrix of $\mathbf{F}$ from row i to row j. The output sequence $\mathbf{o} \in \mathbb{R}^{n-h+1}$ of the convolution operator is obtained by repeatedly applying the filter on sub-matrices of $\mathbf{F}$: $o_i = \mathbf{w} \bullet \mathbf{F}[i : i + h - 1]$, where $i = 1 \ldots n - h + 1$, and $\bullet$ is the dot product between the sub-matrix and the filter (a sum over element-wise multiplications). We add a bias term $b \in \mathbb{R}$ and an activation function f to each o_i, inducing the feature map $\mathbf{c} \in \mathbb{R}^{n-h+1}$ for this filter: $c_i = f(o_i + b)$.

To optimize our model, we then use multiple filters [16] for the same region size to learn complementary features from

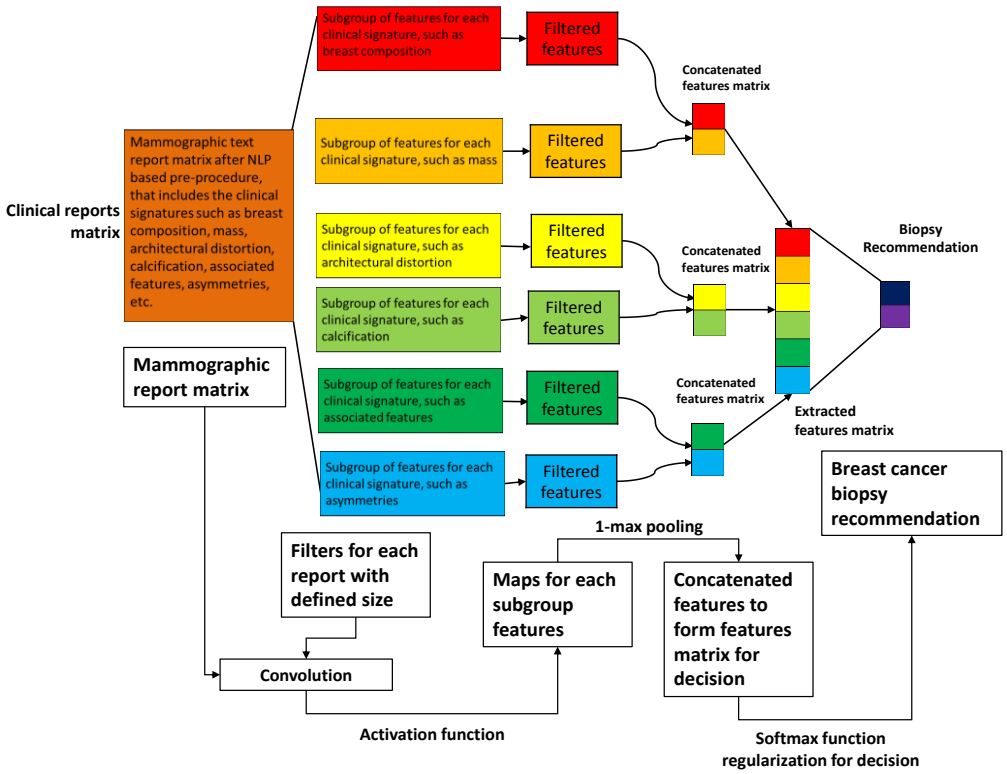

Figure 2. Flowchart of deep learning analytic for diagnosis support of breast cancer management.

the same regions. We also specify multiple kinds of filters with different region sizes (e.g., 'heights'). The dimensionality of the feature map generated by each filter may vary as a function of the sentence length and the filter region size. A pooling function is applied to each feature map to induce a fixed-length vector. A common strategy is 1-max pooling [17], which extracts a scalar from each feature map. Together, the outputs generated from each filter map can be concatenated into a feature vector, which is then fed through a softmax function to generate the final decision.

At the softmax layer, the dropout model [18] is applied as a means of regularization. This entails randomly setting values in the weight vector to 0. One may also impose an $l2$ norm constraint, i.e., linearly scale the $l2$ norm of the vector to a pre-specified threshold when it exceeds this.

Figure 2 provides a schematic diagram to illustrate the aforementioned deep learning model architecture. A training objective to be minimized is the categorical cross-entropy loss. We then used SGD[19] and back-propagation[20] to estimate the parameters in the model. These parameters include the weight vectors of the filters, the bias term in the activation function, and the weight vector of the softmax function optimization.

III. Implementation

In the implementation, MOTTE was first applied to mammographic reports to extract contexts including breast suspicious information. Then the deep learning method was used to establish the analytic model. The inputs of the model are mammographic and demographic text reports that are stored in our database while the outcomes of the model are the possibilities of cancer types from the biopsy report-based pathology findings. We defined the five types from the biopsy report, benign, Atypia, LCIS (Lobular Carcinoma In Situ), DCIS (Ductal Carcinoma In Situ), and carcinoma. The decision function in our deep learning model was assigned with the possibilities of each cancer type. The data from 699 breast suspicious patients with BI-RADS 5 retrieved from Houston Methodist enterprise data warehouse, METEOR[8], were used for model training and retrospective testing, in which data from 400 patients data were used for initial deep learning model training. When training the model, our clinical experts were also monitoring the data in order to ensure the inputs of the model were accurate. Table I lists the 400 breast suspicious patients' data that are categorized according to the defined cancer types to make sure that each patient's outcome has four cancer type possibilities.

In the validation, we evaluated the new analytic model by comparing the predictive accuracy with current radiologist-based BI-RADS 5 biopsy recommendations using pathology report as benchmark. Invasive breast cancer or DCIS as a histological diagnosis was considered as a positive result. Any other pathology diagnosis was considered as free of breast cancer and was considered as negative validation results. A binary classification was performed to determine the

TABLE I. NUMBER OF BI-RADS 5 PATIENTS THAT WERE INCLUDED IN THE TRAINING DATASET

Number of subjects	Pathology Findings (Based on Biopsy Report)				
	Population with Cancer Type: Benign (%)	Population with Cancer Type: Atypia (%)	Population with Cancer Type: LCIS (%)	Population with Cancer Type: DCIS (%)	Population with Cancer Type: Carcinoma (%)
400	21(5%)	4(1%)	59(15%)	32(8%)	284(71%)

performance of our model.

With the testing data from 299 patients, our clinical experts also helped to identify the breast cancer subtype of each patient. The manual review was performed for the set of 299 patients to confirm our model accuracy. The results of manual review from these patients were used as ground truth. Within the stated breast cancer subtypes, 23 patients were identified with DCIS, 12 patients were identified with benign breast biopsies, 45 patients were identified with LCIS, and 219 patients have available pathology reports with carcinoma findings. Our trained model performed on the testing data and provided the possibilities of breast cancer subtypes based on the information from mammograms and demographic text reports. The results with high possibilities on carcinoma and DCIS were considered biopsy recommended patients. By comparing with the ground truth, Figure 3 illustrates the ROC curve of our validations and the AUC of the ROC curve is 0.79.

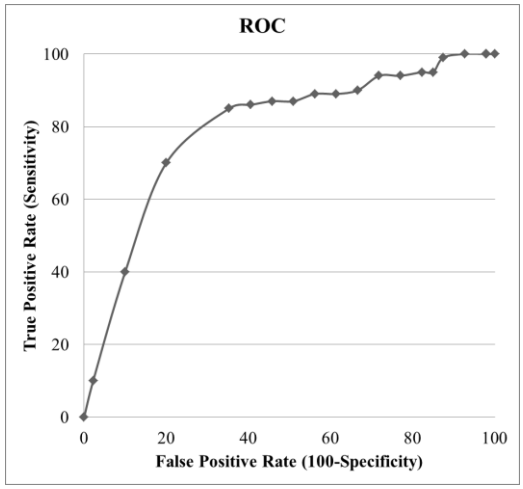

Figure 3. ROC of experimental results using our analytic model.

IV. CONCLUSION

In this paper, we present a well-defined, novel natural language processing based deep learning analytic model for breast cancer management incorporating hospital warehouse datasets. This model outputs a probability measure of biopsy recommendation that is more clinically relevant and informative than the traditional BI-RADS scores. It is the first diagnosis support model for breast suspicious patient data using NLP based deep learning method and will enhance engagement between the patient and clinician in making an informed decision on whether to biopsy or not. We compared the proposed model with manual review results and showed that the deep learning method maintains high accuracy. Further work includes incorporating more data with different BI-RADS scores that can be lined up with radiological reporting as well as further optimization of assessment, and extensive evaluation. We will also validate the model in the prospective study, and compare with other models such as GMM, PCA, SVM and etc. More users will be allowed to test our model.

ACKNOWLEDGMENT

This research is supported by John S Dunn Research Foundation and TT & WF Chao Center for BRAIN. The authors would like to thank our physician collaborators and hospital IT colleagues at the Houston Methodist Hospital during the implementation of the presented framework.

REFERENCES

[1] T. A. Patel, M. Puppala, R. O. Ogunti, J. E. Ensor, T. He, J. B. Shewale, *et al.*, "Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods," *Cancer,* 2016.

[2] C. D. Lehman, A. Y. Lee, and C. I. Lee, "Imaging management of palpable breast abnormalities," *American Journal of Roentgenology,* vol. 203, pp. 1142-1153, 2014.

[3] T. Ayer, Q. Chen, and E. S. Burnside, "Artificial neural networks in mammography interpretation and diagnostic decision making," *Computational and mathematical methods in medicine,* vol. 2013, 2013.

[4] A. Stojadinovic, C. Eberhardt, L. Henry, J. Eberhardt, E. A. Elster, G. E. Peoples, *et al.*, "Development of a Bayesian classifier for breast cancer risk stratification: a feasibility study," DTIC Document2010.

[5] E. S. Burnside, J. Davis, J. Chhatwal, O. Alagoz, M. J. Lindstrom, B. M. Geller, *et al.*, "Probabilistic computer model developed from clinical data in national mammography database format to classify mammographic findings 1," *Radiology,* vol. 251, pp. 663-672, 2009.

[6] Y. Bengio, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning,* vol. 2, pp. 1-127, 2009.

[7] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013.

[8] M. Puppala, T. He, S. Chen, R. Ogunti, X. Yu, F. Li, *et al.*, "METEOR: an enterprise health informatics environment to support evidence-based medicine," *IEEE Transactions on Biomedical Engineering,* vol. 62, pp. 2776-2786, 2015.

[9] T. He, R. Ogunti, M. Puppala, S. Chen, X. Yu, J. J. Mancuso, *et al.*, "A smartphone app framework for segmented cancer care coordination," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2016, pp. 372-375.

[10] M. Puppala, T. He, X. Yu, S. Chen, R. Ogunti, and S. T. Wong, "Data security and privacy management in healthcare applications and clinical data warehouse environment," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2016, pp. 5-8.

[11] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in *ACL (System Demonstrations)*, 2014, pp. 55-60.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111-3119.

[13] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882,* 2014.

[14] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 160-167.

[15] Y. Zhang and B. Wallace, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification," *arXiv preprint arXiv:1510.03820,* 2015.

[16] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, pp. 436-444, 2015.

[17] Y.-L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 111-118.

[18] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research,* vol. 15, pp. 1929-1958, 2014.

[19] T. M. Breuel, "The effects of hyperparameters on SGD training of neural networks," *arXiv preprint arXiv:1508.02788,* 2015.

[20] X. Li and D. Roth, "Learning question classifiers," in *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, 2002, pp. 1-7.