Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240776

Efficient Clinical Information Extraction from Breast Radiology Reports in French

Jamil ZAGHIR[*ab1], Belinda LOKAJ[*bc], Karen KINKEL[d], Amal-Dahila DJEMA[e], Hugues TURBÉ[ab], Mina BJELOGRLIC[ab], Valentin DURAND DE GEVIGNEY[c], Jérôme SCHMID[c], Christian LOVIS[ab] and Jean-Philippe GOLDMAN[ab]
[a] *Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland*
[b] *Department of Radiology and Medical Informatics, University of Geneva, Geneva Switzerland*
[c] *Geneva School of Health Sciences, HES-SO University of Applied Sciences and Arts Western Switzerland, Delémont, Switzerland*
[d] *Réseau Hospitalier Neuchâtelois, Neuchâtel, Switzerland*
[e] *Hirslanden - Clinique des Grangettes, Geneva, Switzerland*
[*] These authors contributed equally to this work

Abstract. Radiology reports contain crucial patient information, in addition to images, that can be automatically extracted for secondary uses such as clinical support and research for diagnosis. We tested several classifiers to classify 1,218 breast MRI reports in French from two Swiss clinical centers. Logistic regression performed better for both internal (accuracy > 0.95 and macro-F1 > 0.86) and external data (accuracy > 0.81 and macro-F1 > 0.41). Automating this task will facilitate efficient extraction of targeted clinical parameters and provide a good basis for future annotation processes through automatic pre-annotation.

Keywords. Breast cancer, NLP, Radiology report, text classification

1. Introduction

Recent advances in natural language processing (NLP) and machine learning (ML) have encouraged research for information extraction from complex and unstructured radiology reports. Research and predictive analysis could benefit from useful and clinically relevant information [1]. Potential applications of NLP in radiology with information extraction have been reported, such as clinical decision support by analyzing many reports to answer a specific clinical question or triaging improving clinical workflow [2,3], to monitor appropriate medical imaging use or protocol study [4], or to perform image labeling for computer vision [2]. Furthermore, it can involve the creation of a searchable database, enabling the creation of large sets of labeled data. These sets can be organized into diverse diagnoses enabling identification of cohorts for clinical trials or specific patient diagnoses [3,5,6]. Automatic information extraction is also a useful way of improving radiologists' reading efficiency by identifying key information

[1] Corresponding Author: Jamil Zaghir; E-mail: Jamil.Zaghir@unige.ch.

for patient management care, thus reducing manual search efforts [7]. Relatively good performances have been reported [1], and even similar performances to manual extraction by trained professionals [8]. In previous reviews of NLP applied to radiology reports, it was reported that included studies mostly aimed at classification, including classification of medical history of patient information, with approaches relying on manually defined rules, machine learning-based, or hybrid approaches [6,8].

In the context of a Swiss research study (SUBREAM) aiming to combine magnetic resonance imaging (MRI) data and non-imaging data for breast cancer diagnosis, the objective of this study was to evaluate breast radiology reports classification performances allowing targeted clinical parameter extraction, following the methodology of similar studies [9].

2. Methods

2.1. Corpus data and annotation step

We analyzed a total of 1,218 breast radiology reports from two clinical centers: 1,079 reports from Geneva (CCG) with 301 patient cases, and 139 reports from Neuchâtel (CCN) with 31 patient cases. The dataset spans nearly two decades, providing breast MRI reports in French, along with up to four previous breast imaging reports.

Manual annotation of CCG reports was performed by 3 juniors (A, B, C) and 1 expert (D) radiographer with Brat [10]. For CCN reports, annotation was made later and exclusively by annotator D. The annotators followed the guidelines describing the annotation scheme for six different classification tasks. These guidelines, constructed with the collaboration of the radiologist of CCN who has expertise knowledge in breast cancer, covered clinical parameters related to breast cancer risk factors and factors influencing contrast enhancement. These included menopausal status, contraception, personal and family history of breast cancer, BRCA mutation, and chemotherapy treatment status, typically found in the "*Indications*" section (Table 1). The annotation process began with a calibration annotation campaign for a subset of 50 cases, revealing suboptimal performance (F-scores: 0.56-0.83). This phase led to refining instructions and clarifying potential sources of confusion. In the second phase, corrections on the same set of 50 cases substantially improved inter-annotator agreement (F-scores: 0.9-0.95).

Therefore, a decision was made to segregate with overlap between annotators the remaining data. For the final annotation of all radiology reports, the agreement achieved was satisfactory when compared to the expert annotator D (F-scores: 0.9-0.93), with a minor loss due to two annotators forgetting one and three cases, and minimal observed mistakes. Despite high agreement, some ambiguity remained. To enhance consistency, only annotations by annotator D were used for ML experiments.

2.2. Machine learning step

Rule-based data preprocessing was performed to ensure that no patient identifiers were kept and to extract the content of the "*Indications*" section. This step also included lowercasing, retaining diacritics, punctuation removal, and stop-word removal except for words indicating negation. The text collection was then transformed into a matrix of word-counts. Bigrams of tokens were also incorporated into the matrix to ensure that the model is aware of existing collocations in the input.

Six traditional ML techniques were used: Support Vector Machine (SVM) with radial basis function (RBF), another one with a linear function (LIN), Naive Bayes (NB), Logistic Regression (LR), Random Forrest (RF), and K-Nearest Neighbors (KNN). Hyperparameters were tuned, and a 5-fold cross-validation splitting scheme was employed (i.e. the test set represented 20% of the data). To evaluate these classifiers' performance, for both internal and external validation, we used accuracy. Furthermore, as the dataset was imbalanced for multiple entity types, the macro-F1 score was employed to ensure equal treatment of each class regardless of their support values.

3. Results

Table 1 illustrates the distribution of classes for CCN and CCG reports. The datasets showed a substantial imbalance, characterized by a significant prevalence of missing information denoted as "*no info*", indicating instances where the information was not mentioned within the reports. To mitigate the imbalance, some classes were consolidated, by merging perimenopausal and menopausal states, and "*no info*" and "*no*" for chemotherapy categories, since it is predominantly mentioned during active treatment or within the immediate post-treatment phase. Figure 1 depicts the results of interval validation (A, B) and external validation (C, D), with the test set being CCG (20% split) and CCN reports (full dataset) respectively. Majority class is represented in Figure 1 (B, D) to be compared to the accuracies. Among the six classifiers, LR stood out as the best-performing, slightly followed by NB and both SVM variants (RBF and LIN).

Table 1. Classes for each task, in CCG and CCN reports. The LR performances are also reported.

	CCG n=1079	**LR**	**CCN n=139**	**LR**
MENOPAUSAL STATUS (MenoSt)		(Macro F1 /		(Macro F1
Menopause absent	119 (11%)	accuracy)	1 (0.7%)	/ accuracy)
Menopause present	245 (22.7%)		2 (1.4%)	
Menopause with substitute	64 (5.9%)		2 (1.4%)	
No info	651 (60.3%)	0.88 / 0.95	134 (96.4%)	0.41 / 0.96
CONTRACEPTION (Contra)				
With contraception	20 (1.9%)		1 (0.7%)	
Without contraception	46 (4.3%)		1 (0.7%)	
No info	1013 (93.9%)	0.89 / 0.99	137 (98.6%)	1 / 1
FAMILY HISTORY OF BC (FamRisk)				
No risk	144 (13.3%)		4 (2.9%)	
Yes, family risk	281 (26%)		17 (12.2%)	
No info	654 (60.6%)	0.98 / 0.98	118 (84.9%)	0.79 / 0.95
PERSONAL HISTORY OF BC (PatRisk)				
Yes	605 (56.1%)		105 (75.5%)	
Other	68 (6.3%)		-	
No info	406 (37.6%)	0.93 / 0.97	34 (24.5%)	0.52 / 0.81
BRCA MUTATION (BRCA)				
Negative	21 (1.9%)		-	
Positive	57 (5.3%)		7 (5%)	
No info	1001 (92.8%)	0.86 / 0.99	132 (95%)	1 / 1
CHEMOTHERAPY (CHEMO)				
No	1043 (96.7%)		121 (87.1%)	
Yes	36 (3.3%)	0.91 / 0.99	18 (12.9%)	0.61 / 0.89

Bold style: Classification tasks. **Red**: Majority class (MC). **BC**: breast cancer

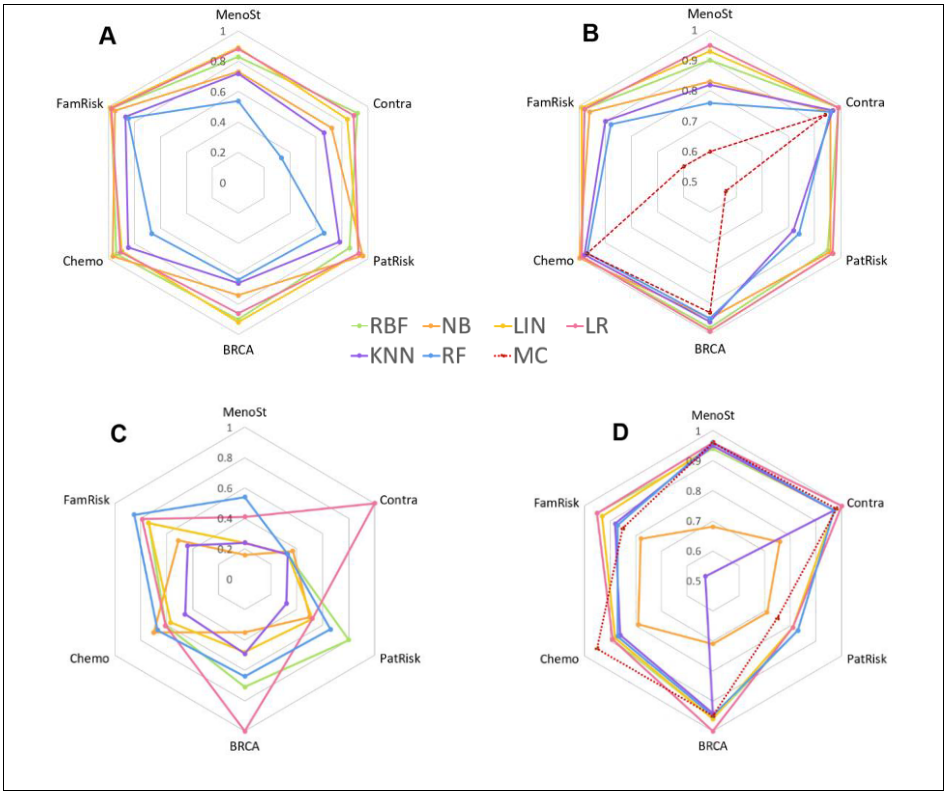

Figure 1. Performances of each classifier for each classification task. Macro F1 (A) / Accuracy (B) for CCG internal data, and Macro F1 (C) / Accuracy (D) for CCN external data. Red dashed lines: Majority class (MC)

LR's accuracies were consistently above 0.95 for CCG reports, and 0.8 for CCN ones. However, due to highly imbalanced datasets, with some classes having low support, it was crucial to use a metric that takes this parameter into account such as macro F1. The macro-F1 performances followed the same trend as the ones for accuracy in the internal validation. However, this is less straightforward for some classes in external validation. The higher proportion of chemotherapy cases in CCN (12.9%) compared to CCG (3.3%) may result in lower performance because the model was trained on CCG data. Additionally, excellent performance is noted for BRCA and contraception classification. Concerning personal history of BC, the poor results are attributed to abbreviations used for cancer designation in CCN reports that were absent in the CCG training data.

4. Discussion

While not perfect, these models proved to be excellent candidates for future pre-annotation tasks. LR classifier consistently outperformed other models in both internal and external validation. However, the Macro-F1 score, particularly in external validation, showed slightly lower performance, likely due to variations in radiology reporting practices between CCG and CCN. The dataset used for external validation showed no annotation biases, with consistent adherence to guidelines by the same annotator. The only discernible bias was attributed to the data source, impacting the model's performance. Additionally, the expert annotator reported that the two centers did not follow the same way of reporting patient information, highlighting the need for standardized information reporting. Future research should compare our findings using large language models to leverage prior knowledge for improved generalization. Although comparison across dissimilar classification tasks presents challenges, the results of this study are comparable to the current literature for report classification [1,2,9].

5. Conclusions

This study demonstrated the potential for automatic breast report classification in CCG, according mainly to breast cancer risk factors. Thus, allowing efficient clinical parameter extraction for research purposes. Inclusion of reports from more clinical centers in the training set, and clinical validation could improve model generalization. .

This work is part of the SUBREAM project funded by the Swiss Cancer Research (KFS-5460–08-2021-R) and approved by the Geneva Cantonal Ethics Committee (ID: 2019-00716). Informed consent was obtained from each patient for re-use of anonymized breast reports.

References

[1] Short RG, Bralich J, Bogaty D, Befera NT. Comprehensive Word-Level Classification of Screening Mammography Reports Using a Neural Network Sequence Labeling Approach. J Digit Imaging. 2019;32(5):685-92.

[2] Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, Langlotz CP, Amrhein TJ, Lungren MP. Deep Learning to Classify Radiology Free-Text Reports. Radiology. 2018;286(3):845-52.

[3] Mozayan A, Fabbri AR, Maneevese M, Tocino I, Chheang S. Practical Guide to Natural Language Processing for Radiology. RadioGraphics. 2021;41(5):1446-53.

[4] Akinci D'Antonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, Cuocolo R, Cannella R, Koçak B. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. dir. 2023;0(0):0-0.

[5] Yala A, Barzilay R, Salama L, Griffin M, Sollender G, Bardia A, Lehman C, Buckley JM, Coopey SB, Polubriaginof F, Garber JE, Smith BL, Gadd MA, Specht MC, Gudewicz TM, Guidi AJ, Taghian A, Hughes KS. Using machine learning to parse breast pathology reports. Breast Cancer Res Treat. 2017;161(2):203-11.

[6] Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiology. 2016;279(2):329-43.

[7] Dada A, Ufer TL, Kim M, Hasin M, Spieker N, Forsting M, Nensa F, Egger J, Kleesiek J. Information extraction from weakly structured radiological reports with natural language queries. Eur Radiol. 2023.

[8] Saha A, Burns L, Kulkarni AM. A scoping review of natural language processing of radiology reports in breast cancer. Front Oncol. 2023;13:1160167.

[9] Goldman JP, Mottin L, Zaghir J, Keszthelyi D, Lokaj B, Turbé H, Gobeil J, Ruch P, Ehrsam J, Lovis C. Classification of Oncology Treatment Responses from French Radiology Reports with Supervised Machine Learning. Stud Health Technol Inform. 2022;294:849-53.

[10] Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. Brat: a Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the EACL. Avignon, France: ACL; 2012. p. 102-7.