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

doi:10.3233/SHTI240476

Automatic Classification of Conclusions from Multi-Tracer Reports of PET Brain Imaging in Cognitive Impairment

Jean-Philippe GOLDMAN[a,b,1], Pablo JANÉ[c,d], Jamil ZAGHIR[a,b], Eliluane PIRAZZO ANDRADE TEIXEIRA[c], Débora Elisa PERETTI[d], Valentina GARIBOTTO[c,d] and Christian LOVIS[a,b]

[a] *Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland*

[b] *Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland*

[c] *Division of NuclearMedicine and Molecular Imaging, Geneva University Hospitals, Geneva, Switzerland*

[d] *NIMTLab, Faculty of Medicine, University of Geneva, Center of biomedical imaging (CIBM), Geneva, Switzerland*

ORCiD ID: Jean-Philippe Goldman https://orcid.org/0000-0003-4000-4199

Abstract. The goal of this paper is to build an automatic way to interpret conclusions from brain molecular imaging reports performed for investigation of cognitive disturbances (FDG, Amyloid and Tau PET) by comparing several traditional machine learning (ML) techniques-based text classification methods. Two purposes are defined: to identify positive or negative results in all three modalities, and to extract diagnostic impressions for Alzheimer's Disease (AD), Fronto-Temporal Dementia (FTD), Lewy Bodies Dementia (LBD) based on metabolism of perfusion patterns. A dataset was created by manual parallel annotation of 1668 conclusions of reports from the Nuclear Medicine and Molecular Imaging Division of Geneva University Hospitals. The 6 Machine Learning (ML) algorithms (Support Vector Machine (Linear and Radial Basis function), Naive Bayes, Logistic Regression, Random Forrest, and K-Nearest Neighbors) were trained and evaluated with a 5-fold cross-validation scheme to assess their performance and generalizability. The best classifier was SVM showing the following accuracies: FDG (0.97), Tau (0.94), Amyloid (0.98), Oriented Diagnostic (0.87 for a diagnosis among AD, FTD, LBD, undetermined, other), paving the way for a paradigm shift in the field of data handling in nuclear medicine research.

Keywords. Nuclear Medicine, brain molecular imaging reports, text classification

1. Introduction

Electronic health records (EHR) serve as the primary information cornerstone within the clinical domain and are central for medical research efforts. Despite notable recent efforts aimed to transform the medical information landscape, the proportion of structured reporting documents within EHR databases, particularly in medical imaging, remains

[1] Corresponding Author: Jean-Philippe Goldman; E-mail: jglm@hcuge.ch.

remarkably low. Consequently, medical researchers are compelled to manually organize natural language texts from EHR into spreadsheets, a critical but cumbersome process that acts as a significant bottleneck. The advent of Machine Learning (ML)[1] and Natural Language Processing (NLP)[2][3] is ushering in a transformative era in medicine, revolutionizing both the field itself and our interactions with unstructured text[4][5][6][7], thus accelerating the implementation of real-world-data analyses in the clinic and setting the foundation for a new paradigm of personalized medicine.

This contribution presents a comprehensive methodology for document selection, conclusion extraction, data annotation, and conclusion classification, leveraging six established traditional machine learning techniques, following similar studies such as [8][9]. The methodology encompasses rigorous processes to ensure robustness and reliability at each stage of the analysis. First, document selection criteria and procedures are carefully defined to facilitate the acquisition of relevant data for subsequent analysis. Following document selection, a systematic approach to conclusion extraction is used to identify and extract key insights from the selected documents. Subsequently, a rigorous data annotation process is implemented to facilitate the training and validation of the machine learning models. Finally, six well-established traditional machine learning techniques are used to classify conclusions with precision and efficiency. The results of this methodology are then presented and discussed, providing valuable insights and perspectives on the efficacy and potential applications of the techniques employed.

2. Methodology

2.1. Data selection

The context of this study is a research project aimed at investigating molecular brain biomarkers in neurodegenerative disorders from existing amyloid-PET, tau-PET and/or [18F]FDG-PET images performed at the Nuclear Medicine Unit at the Geneva University Hospitals (HUG) since 2006. An initial list of 1434 brain PET exams has been compiled from patients who gave informed consent after approval by the Geneva Cantonal Ethics Committee (CCER). The inclusion criteria for the study are as follows: 1) clinical suspicion of a neurodegenerative disease; 2) [18F]FDG-PET, tau -PET and/or amyloid-PET performed at the Nuclear Medicine Unit at the HUG since 2006; 3) subjects >18 years of age. The exclusion criterion is a written or documented refusal by the patient to allow the re-use of his/her medical records.

This main dataset is divided into three sub-lists according to the PET examination modality as shown in Table 1. From this original list, only 1423 cases could be retrieved from the HUG datalake. After discarding further empty documents and non-PET/CT exams, the final list of examinations for this study contains 1386 brain PET/CT reports (95.6% of the original list) from 768 different patients.

Table 1. Breakdown of PET type in original and effective list for the study

	Amyloid	FDG	Tau	Total
Original list	645	450	339	1434
Final list	616	441	329	1386

In a second step, a second extraction of documents was carried out in order to complement specifically low supported classes of the original dataset. In this case, the extraction was performed by averaging two sets of specific keywords in English and

French (*DFT, FTD, fronto-temporale + Lewy, DLB, LBD, DCL)* found in the full reports of such "PET/CT CEREBRAL SANS CONTRASTE (PET-1)" regardless of the PET modality. Out of 421 documents found, 284 documents were retained (96 for LBD and 186 for FDT), leaving out documents from patients with unknown consent or explicit refusal, or who did not meet with the inclusion criteria.

For these 1670 reports (1386+284), the conclusion part has been extracted with regular expressions and was used for annotation and training in the study.

2.2. Data annotation

A manual classification of these conclusions was performed on the first dataset using `brat` (namely *brat rapid annotation tool*). The annotation scheme included a main set of 7 labels for metabolism/perfusion as well as two pairs of labels for amyloid and tau as shown in Table 2. The documents of the FDG subset were annotated for perfusion outcome only, whereas the documents of amyloid and tau subsets were annotated for both perfusion outcome and for their respective test (negative vs. positive).

Since early phase perfusion patterns of amyloid and tau radiotracers can serve as a proxy for pathological metabolic patterns, the documents of the FDG subset were annotated for neurodegenerative disorder pattern only, while the documents of amyloid and tau subsets were annotated for both neurodegenerative disorder pattern and for their respective test outcome (negative vs. positive)[10].

Table 2. Number of labels for the parallel annotation for the 3 tasks (perfusion, amyloid, tau), with the number of documents per tasks, the Cohen's kappa and the percentage of agreement between the two annotators, the detailed number of labels per annotators

	Label	# docs	Annot#1	Annot#2
Neurodegeneration	AD (Alzheimer's disease)	1386	147	144
labels	FTD (fronto-temporal dementia)	(κ=0.95)	42	43
	LBD (dementia with Lewy bodies)	(%a=0.98)	17	13
	other		22	12
	undetermined		173	164
	negative_perfusion		259	249
	non_applicable		21	21
Amyloid labels	negative_amy	616	265	261
	positive_amy	(κ=0.97)	326	331
	unannotated	(%a=0.99)	25	24
Tau labels	negative_tau	329	175	186
	positive_tau	(κ=0.93)	136	125
	unannotated	(%a=0.97)	18	18

The annotators had to manually annotate the relevant group of words. This precise annotation is not used in the study but could be used for future work in information extraction. In this classification task, labels were considered for the whole conclusion. The annotation times for the first annotator was 8.5 hours (resp. 12.3h for the second), yielding an annotation rate of 2.7 document per minute (resp. 1.9 doc/mn). The Cohen's kappa inter-annotator agreement is 0.95 for the perfusion annotation task (resp. 0.97/0.93 for the amyloid /tau task) showing the great homogeneity in the parallel annotations.

At this stage, some of the classes are clearly underrepresented (DFT, DCL, other) and would penalize the performance of ML approach. The annotation of the second dataset followed a different paradigm. Since the documents were selected by specific keywords, an automatic pre-annotation was performed by marking the whole conclusion with the corresponding annotation label (FTD for the 186 documents selected by the

keywords *DFT, FTD, fronto-temporale*, and LBD for the 96 documents selected by the keywords *Lewy, DLB, LBD, DCL*). In this case, the unique annotator had to accept or to correct the pre-annotation. This approach gave good results from a methodological point of view as only 34% (resp. 60%) of the 186 (resp. 96) documents pre-annotated as DFT (resp. DCL) had to be corrected, with an overall lower human annotation time (3 reports per minute compared to 1.8 for the first stage). The annotation of the specific relevant group of words was not considered for this dataset. This additional dataset allowed to rebalance the FTD and LBD under-populated classed, prior to the learning step.

In total the joint dataset contains 1668 documents from the two stages, with a breakdown by class for perfusion described in the Table 3, which shows a less unbalanced distribution between classes compared to Table 2. Finally, the FDG subset was also considered in a two-class description by merging all classes for positive perfusion (*MA, DFT, DCL, other, undetermined*) with 655 documents vs. *negative_perfusion* (with 290 documents).

Table 3. Support for perfusion classification

	Label	# docs
Perfusion labels	AD (Alzheimer's disease)	159
	FTD (fronto-temporal dementia)	134
	LBD (dementia with Lewy bodies)	54
	Other + undetermined	308
	negative_perfusion	290

2.3. Machine learning

This stage encompassed several preprocessing steps including lower case conversion, retention of diacritics, removal of punctuation and stop-words (excluding negation indicators). The text corpus was then converted into a matrix representing word frequencies, including bigrams to capture collocational information.

Six traditional ML techniques were applied: Support Vector Machine (SVM) with radial basis function (RBF), or with a linear function (LIN), Naive Bayes (NB), Logistic Regression (LR), Random Forrest (RF), and K-Nearest Neighbors (KNN). Hyperparameters were optimized with a 5-fold cross-validation strategy (i.e., 20% of the data for testing). Moreover, given the class imbalance of the dataset, the macro-F1 score was utilized to ensure fair evaluation across all classes regardless of their support values.

3. Results

Table 4. Accuracy for the 6 ML approaches and for the 4 classifying tasks , with the number of classes per task (k) and the proportion of the majority class (maj.k). For the perfusion classifiers, the macro F1 score is also indicated in parenthesis.

	Docs	k	maj.k	SVM-RBF	SVM-LIN	NB	LR	RF	KNN
Perfusion	1668	6	0.42	**0.87(0.74)**	0.86(0.71)	0.81(0.71)	0.86(0.73)	0.74(0.55)	0.79(0.66)
Perf pres.	1668	3	0.42	**0.96(0.95)**	0.95(0.94)	0.93(0.91)	0.96(0.94)	0.89(0.85)	0.89(0.87)
Amyloid	616	2	0.53	0.96	**0.98**	**0.98**	0.97	**0.98**	0.97
Tau	329	2	0.51	**0.94**	0.92	0.85	0.92	0.84	0.85

Table 4 shows the accuracy (and the macro-F1 score) for the 6 classification approaches and for the 4 classification tasks.

4. Discussion

The SVM-RBF showed the best results, followed by LR. An attempt to use bigram of words did not give any improvement. The amyloid classification performed better compared with tau task, probably due to a higher number of document but also with a lot of recurrent wording in these conclusions. In any case, the accuracy is far above the percentage of the majority class.

5. Conclusions

This study of automatic classification of brain PET report conclusion showed great results, almost as high as the agreement between two human annotators. The support vector machine approach with radial basis function seems to be the most appropriate for this task. The experiment should be repeated with a larger amount and various types and provenance of reports to test generalizability of the method. Adding such structured semantic data to EHRs allows them to be incorporated into larger standardized corpora for research studies or into a graph database for larger purposes.

This study is approved the Geneva CCER: Research study of molecular and structural neuroimaging in neurodegenerative diseases (CCER number :2022-01520).

References

[1] Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: A systematic review. Radiology . 2016;279(2):329–43. Available from: http://dx.doi.org/10.1148/radiol.16142770

[2] Shaitarova A, Zaghir J, Lavelli A, Krauthammer M, Rinaldi F. Exploring the latest highlights in medical Natural Language Processing across multiple languages: A survey. Yearb Med Inform . 2023;32(1):230–43. Available from: http://dx.doi.org/10.1055/s-0043-1768726

[3] Zaghir J, Goldman J-P, Bjelogrlic M, Keszthelyi D, Gaudet-Blavignac C, Turbé H, et al. Performance of machine learning methods to classify French medical publications. Stud Health Technol Inform . 2022;294:874–5. Available from: http://dx.doi.org/10.3233/SHTI220613

[4] Reichenpfader D, Müller H, Denecke K. Large language model-based information extraction from free-text radiology reports: a scoping review protocol, BMJ Open. 13 (2023) e076865

[5] Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, Larson DB, et al. Deep learning to classify radiology free-text reports. Radiology . 2018;286(3):845–52. Available from: http://dx.doi.org/10.1148/radiol.2017171115

[6] Mozayan A, Fabbri AR, Maneevese M, Tocino I, Chheang T Akinci Dantonoli S, Stanzione A, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. RadioGraphics . 2021;41:1446–53. Available from: http://dx.doi.org/10.1148/rg.2021200113

[7] Akinci Dantonoli T, Stanzione A, Bluethgen C, Vernuccio F, Ugga L, Klontzas ME, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Dir . 0:0–0. Available from: http://dx.doi.org/10.4274/dir.2023.232417

[8] Lokaj B, Zaghir J, Kinkel K, Djema A-D, Schmid J, Lovis C, et al. Goldman Natural Language Processing for Efficient Clinical Patient Information Extraction from Breast Radiology Reports. Studies in health technology and informatics. 2024;

[9] Goldman J-P, Mottin L, Zaghir J, Keszthelyi D, Lokaj B, Turbé H, et al. Classification of oncology treatment responses from french radiology reports with supervised machine learning. Stud Health Technol Inform . 2022;294:849–53. Available from: http://dx.doi.org/10.3233/SHTI220605

[10] Boccalini C, Peretti DE, Ribaldi F, Scheffler M, Stampacchia S, Tomczyk S, et al. Early-phase18F-florbetapir and18F-flutemetamol images as proxies of brain metabolism in a memory clinic setting. J Nucl Med . 2023;64(2):266–73. Available from: http://dx.doi.org/10.2967/jnumed.122.264256