www.nature.com/scientificreports

scientific reports

OPEN

# Efficient labeling of french mammogram reports with MammoBERT

Nazanin Dehghani✉, Vera Saliba-Colombani, Aurélien Chick, Morgane Heng, Grégory Operto & Pierre Fillard

**Recent advances in deep learning and natural language processing (NLP) have broadened opportunities for automatic text processing in the medical field. However, the development of models for low-resource languages like French is challenged by limited datasets, often due to legal restrictions. Large-scale training of medical imaging models often requires extracting labels from radiology text reports. Current methods for report labeling primarily rely on sophisticated feature engineering based on medical domain knowledge or manual annotations by radiologists. These methods can be labor-intensive. In this work, we introduce a BERT-based approach for the efficient labeling of French mammogram image reports. Our method leverages both the expansive scale of existing rule-based systems and the precision of radiologist annotations. Our experimental results showcase the superiority of the proposed approach. It was initially fine-tuned on a limited dataset of radiologist annotations. Then, it underwent training on annotations generated by a rule-based labeler. Our findings reveal that our final model, MammoBERT, significantly outperforms the rule-based labeler while simultaneously reducing the necessity for radiologist annotations during training. This research not only advances the state of the art in medical image report labeling but also offers an efficient and effective solution for large-scale medical imaging model development.**

**Keywords** Deep learning, Natural language processing, Information extraction, Mammography report labeling, Breast cancer detection

Breast carcinoma ranks as the most common cancer among women worldwide[1], posing a significant health challenge. Early detection of breast cancer is crucial, as it significantly increases the survival rate and extends patients' lifespans. Mammography is an effective and low-cost imaging tool. It is noninvasive and plays an important role in the early diagnosis of breast diseases due to its high sensitivity. Advancements in deep learning have opened new avenues for developing more accurate and reliable models. These models can diagnose and treat breast cancer. However, training these models, especially in medical imaging, requires large labeled datasets. Given the high cost of radiologist annotation for radiology images, deep learning models for radiology image interpretation often rely on labels automatically extracted from accompanying reports, as noted by[2]. Mammography reports provide detailed insights into mammogram findings. They include potential cancer signs, breast tissue density, masses, calcifications, and architectural distortions. Reports also include relevant patient history, such as prior surgeries.

This paper focuses on extracting information about prior surgeries from mammography reports. We aim to identify the history of breast surgery in patients. We also want to determine whether the surgery occurred in the right or left breast, or both. It's important to exclude mammogram images of patients with prior breast surgery from training datasets, as surgical operation can change the tissue texture. This makes such images unsuitable for training breast cancer detection models on screening mammograms. This task is challenging because there are many invasive treatment methods. These include lumpectomy, mastectomy, surgical biopsy, radiotherapy, and cosmetic surgeries such as breast reduction and symmetrization, all of which alter breast tissue. Moreover, the surgical history is often mentioned sporadically in reports, or not at all. It may also include surgeries on other body parts or relatives. This complicates the design of rules for surgical history extraction. Traditional rule-based methods, though beneficial, produce noisy labels that can adversely impact the training of cancer detection models. Recent advancements with Transformers in end-to-end radiology report labeling, as demonstrated by Drozdov et al.[3] and Wood et al.[4], have shifted the focus from feature engineering to manual annotation, demanding considerable time and expertise. However, these methods often overlook the potential of

Therapixel Company, 1 Imp. Reille, 75014 Paris, France. ✉email: ndehghani@therapixel.com

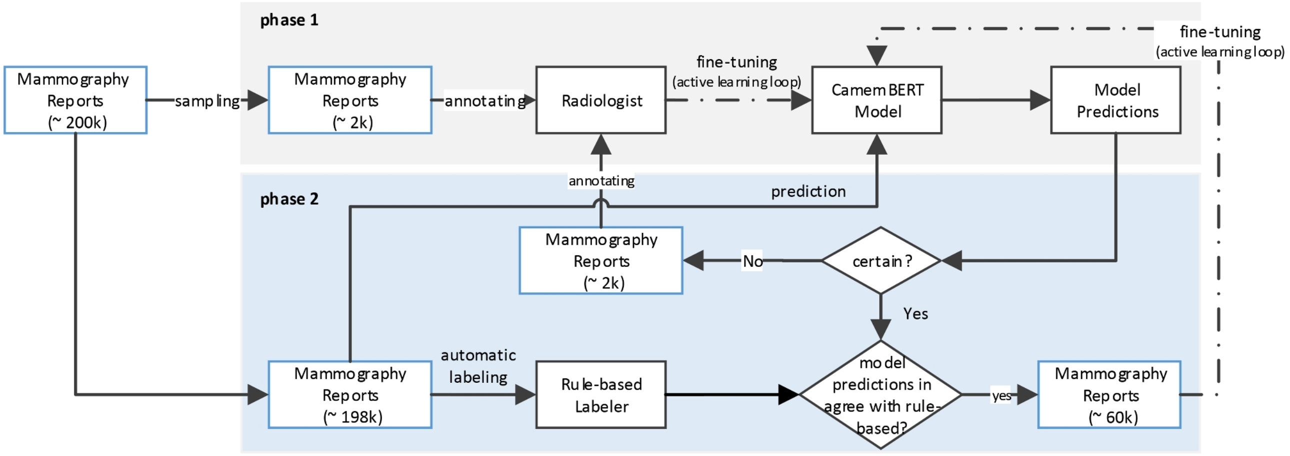
phase 1
Mammography Reports (~ 200k)
sampling
Mammography Reports (~ 2k)
annotating
Radiologist
fine-tuning (active learning loop)
CamemBERT Model
Model Predictions
fine-tuning (active learning loop)
phase 2
annotating
prediction
Mammography Reports (~ 2k)
No
certain ?
Yes
Mammography Reports (~ 198k)
automatic labeling
Rule-based Labeler
model predictions in agree with rule-based?
yes
Mammography Reports (~ 60k)

**Fig. 1**. Architecture of the Mammography Report Labeling Approach. This two-phase method begins with Phase 1, where a pre-trained BERT-based model is initially trained on a small dataset of radiologist annotations (∼2k). In Phase 2, the model undergoes fine-tuning through an active learning loop, which integrates a combination of manual annotations (∼2k) selected via uncertainty sampling, and a larger set of automatic rule-based labels (∼40k) acquired through agreement sampling.

existing rule-based labelers, which remain state-of-the-art for many medical tasks. In this paper, we introduce a simple method that leverages the strengths of both rule-based labels and radiologist annotations. This approach, shown in Fig. 1, begins with a pretrained BERT-based model by[5], initially trained on a small set of radiologist annotations (phase 1). Then, it was fine-tuned through an active learning loop. This loop involved a mix of radiologist annotations selected by uncertainty sampling and a large set of automatic rule-based labels obtained through agreement sampling (phase 2). While the rule-based method was effective in capturing predefined patterns in the data, it faced limitations in handling the variability in natural language. To address these limitations, we developed this hybrid approach that integrates rule-based methods with machine learning models. The rule-based system was used to capture straightforward, well-defined patterns, while the machine learning model was employed to manage more complex cases that required understanding context and handling linguistic variability. This combination allowed us to leverage the precision of rules where applicable while relying on the adaptability of machine learning for more ambiguous cases. Our method, named MammoBERT, applies to the task of extracting surgery prior information from mammography reports. We demonstrate that MammoBERT outperforms traditional rule-based labelers, showing notable improvement on the F1 metric. This approach holds broad potential for natural language processing within the medical domain, where the collection of radiologist labels is costly and feature-engineered labels already exist for many tasks.

## Related work

The extraction of structured labels from free-text radiology reports has been a longstanding focus in natural language processing (NLP), with numerous systems developed over the past two decades[2,6–10]. These methods often rely on extensive feature engineering, incorporating controlled vocabularies and grammatical rules. Notably, tools like NegEx[11] and its extension, NegBio[2], CheXpert labeler[12], have played an important role in this domain, using regular expressions and universal dependencies for report labeling and detecting negations and uncertainties in radiology findings.

Deep learning revolutionized radiology report analysis, shifting from rule-based to data-driven approaches. Studies by[9] and[13] demonstrate the use of deep learning models, including CNNs with GloVe embeddings and networks combined with attention mechanisms, trained on expert-annotated radiology reports. However, these approaches face limitations in training set size, because of the constraints of radiologist time and expertise. The application of these advanced models has shown improvements over traditional rule-based systems, like peFinder[14].

Recent efforts in radiology report labeling have adopted Transformer-based models like BERT[15] and XLNet[16]. Drozdov et al.[3] trained classifiers on radiologist-labeled reports to detect normal and abnormal labels. Wood et al.[4] developed MRI head report classifiers using BioBERT[17] on a substantial number of radiologist-labeled reports. These studies have demonstrated notable improvements over simpler models, showing that Transformer-based architectures effectively capture the complexities of medical language.

In line with these advancements, specialized models have been developed to enhance the processing capabilities for non-English languages. CamemBERT by[5] has effectively adapted the BERT architecture for processing French across various NLP tasks, and DrBERT[18] is specifically developed for interpreting medical texts in French. Despite these developments, the field still faces significant challenges, particularly in the domain of medical radiology. Although there are successful models targeting specific medical topics in other languages, such as those in[19] and[20], French lacks a dedicated radiological language model. This gap underscores the need for developing targeted models that can be effectively used for French radiological reports.

| | NO_SURGERY | SURGERY_LEFT | SURGERY_RIGHT | SURGERY_BOTH | Total |
|---|---|---|---|---|---|
| Rule-based labels | 159,270 | 20,058 | 18,959 | 13,418 | 211,705 |
| Initial Radiologist labels | 888 | 2105 | 1861 | 1855 | 6705 |
| Extended hybrid labels | 42,437 | 7918 | 7219 | 4803 | 62,377 |
| Unseen institute ("Omicron") | 91 | 67 | 69 | 73 | 300 |

**Table 1**. Comparative statistics and distribution of labels among classes. This table presents label statistics obtained by four strategies: labels obtained from the rule-based method, initial labels provided by radiologists, labels from the extended hybrid method, and labels from an unseen institute provided by radiologists.

| | Surgery presence model* | | | Surgery laterality model* | | | |
|---|---|---|---|---|---|---|---|
| **Training strategy** | **P** | **R** | **F1** | **P** | **R** | **F1** | **Final output**** |
| EXP_Rule_Label | 0.817 | 0.874 | 0.845 | 0.648 | 0.518 | 0.575 | 0.82 |
| EXP_Rad_Label | 0.826 | 0.857 | 0.841 | 0.744 | 0.781 | 0.763 | 0.86 |
| EXP_GAN_CamemBERT | 0.866 | 0.879 | 0.872 | 0.793 | 0.806 | 0.799 | 0.88 |
| EXP_HYB_Label | 0.898 | 0.924 | 0.913 | 0.872 | 0.896 | 0.887 | 0.90 |
| EXP_HYB_REFINED_Label | 0.956 | 0.967 | 0.961 | 0.938 | 0.958 | 0.948 | 0.91 |
| EXP_HYB_REFINED_Label_DrBERT | 0.927 | 0.952 | 0.942 | 0.902 | 0.936 | 0.915 | 0.89 |
| EXP_HYB_REFINED_Label_AUG | 0.990 | 0.991 | **0.991** | 0.978 | 0.981 | 0.979 | 0.94 |
| **EXP_HYB_REFINED_Label_AUG_ORG** | 0.988 | 0.989 | 0.988 | 0.982 | 0.985 | **0.983** | **0.97** |

**Table 2**. Average performance metrics of our models in terms of Precision (P), Recall (R) and F1 score with 95% confidence intervals. *Performance is reported for the average of 5-fold cross-validation. **Performance is reported in terms of accuracy on unseen test data. Significant values are in bold.

| Institutes | NO_SURGERY | LEFT_SURGERY | RIGHT_SURGERY | BOTH_SURGERY | Total |
|---|---|---|---|---|---|
| zeta | 13,914 (79.1%) | 1404 (7.9%) | 1345 (7.6%) | 908 (5.1%) | 17,571 |
| chi | 12,108 (79.0%) | 1209 (7.8%) | 1178 (7.6%) | 829 (5.4%) | 15,324 |
| beta | 26,674 (65.6%) | 5567 (13.6%) | 5557 (13.6%) | 2851 (7.0%) | 40,649 |
| delta | 85,046 (83.4%) | 5935 (5.8%) | 5549 (5.4%) | 5428 (5.3%) | 101,958 |
| omicron | 25,724 (80.7%) | 2391 (7.5%) | 2293 (7.1%) | 1440 (4.5%) | 31,848 |

**Table 3**. Class distribution across five screening institutes using the EXP_HYB_REFINED_Label_AUG_ORG model. This table shows the outcomes of using the model for inference across four institutes conducting patient screenings and the Omicron institute. Following this, the class distribution across all datasets from each source is calculated.

In this context, our work introduces a BERT-based approach for efficient labeling of French mammogram reports that bridges the gap between rule-based systems and radiologist annotations. Our method, MammoBert, leverages the scale of rule-based labelers and the precision of radiologist annotations presenting a unique fusion of these two paradigms. MammoBert was initially fine-tuned on a small dataset of radiologist annotations. It was then further trained on annotations generated by a rule-based labeler. This hybrid approach addresses the limitations of both purely rule-based systems and models reliant solely on radiologist annotations.

## Methodology

### Data

We obtained a comprehensive dataset comprising Digital Imaging and Communications in Medicine (DICOM) images and mammography reports from 397,100 studies. The studies were conducted on 226,295 patients with benign or malignant breast who underwent mammography examination between January 2008 and January 2021. The data were sourced from eight Radiology and Medical Imaging institutions in France. The institutions perform screening mammograms. For model training, data from seven institutions were used. The data was first shuffled and split into 5 folds: each fold contains 80% of data for training and 20% for testing. From the training subset, 10% was further used for validation. Models are individually evaluated in terms of the average macro-F1 score over their test set, and the averages across the 5 folds are reported in Table 2. The reports from the eighth institution (referred to as "Omicron" in this paper) were completely held out during training and cross-validation, serving as an unseen test set to assess the final model's generalization capabilities. Including data

| | Surgery presence model | | | Surgery laterality model | | |
|---|---|---|---|---|---|---|
| Fold | P | R | F1 | P | R | F1 |
| Fold 0: Zeta (Test), Chi + Delta (Train) | 0.982 | 0.982 | 0.982 | 0.915 | 0.915 | 0.914 |
| Fold 1: Chi (Test), Zeta + Delta (Train) | 0.982 | 0.982 | 0.982 | 0.965 | 0.963 | 0.965 |
| Fold 2: Delta (Test), Zeta + Chi (Train) | 0.971 | 0.967 | 0.969 | 0.942 | 0.944 | 0.942 |
| Average | 0.978 | 0.977 | 0.978 | 0.941 | 0.941 | 0.940 |

**Table 4**. Cross-validation performance across institutions. P: Precision, R: Recall, F1: macro-F1 Score.

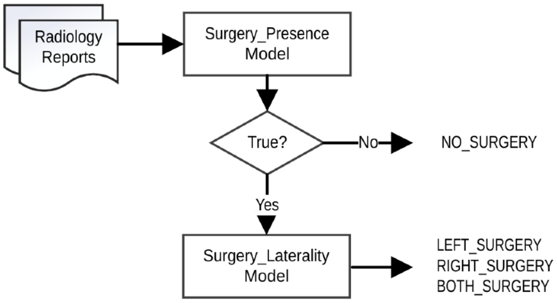

**Fig. 2**. Schematic Representation of the Two-Level Sub-Classification Model. The first level involves the Surgery_Presence model, which performs binary classification to determine the presence or absence of prior surgery. The second level, applicable only to cases with prior surgery history, involves the Surgery_Laterality model, which further classifies the surgery into three categories: left surgery, right surgery, or bilateral surgeries.

from multiple centers makes our approach more generalizable and robust. Data statistics and class distribution of the data is detailed in Table 1.

*Ethical and legal considerations*

After acquiring DICOM images and corresponding mammography reports, we initiated a thorough preprocessing stage. This involved de-identifying the mammography reports to protect patient and doctor privacy, ensuring compliance with ethical standards. Our study received authorization from the Health Data Hub and the Commission Nationale de l'Informatique et des Libertés (CNIL) for the data collected in France, confirming compliance with local regulations on health data collection and privacy. Additionally, we are HDS certified [https://esante.gouv.fr/produits-services/hds] and ISO 27001 [https://www.iso.org/standard/27001], which further underscores our commitment to data security and privacy. To further protect patient data, we developed a secure annotation system for radiologists. Access to this system was tightly controlled via a Virtual Private Network (VPN), ensuring that radiologists could only access and annotate data through this secure environment. This system ensured that data never left the secure server, and only authorized personnel were able to interact with the data.

## Model architecture

We employed a two-level sub-classification approach for learning surgery prior information from mammography report texts. The problem's hierarchical nature and the skewed, imbalanced data guided this choice. Then, we combined the trained classifiers to get the final output. As shown in Fig. 2, first a binary classification is performed to learn the presence or absence of prior surgery (Surgery_Presence model). Second, studies indicating a history of surgery are further classified into left surgery, right surgery, or bilateral surgeries (Surgery_Laterality model). All models use CamemBERT, a French language model based on a modified version of the BERT-base transformer architecture[15], incorporating a sequence classification head on top. The mammography report texts are tokenized, and the maximum number of tokens in each input sequence is constrained to 512.

## Labels

*Rule-based labels*

In our study, we implemented a rule-based method based on regular expressions to extract surgery prior information from the text of mammography reports. This approach involved designing specific patterns to identify whether a patient had undergone prior surgery. In cases where surgery was confirmed, the method further extract details regarding the laterality of breast surgery-whether it occurred on the left, right, or both breasts. We assigned one of four labels to each report using a post-processing step: NO_SURGERY, LEFT_SURGERY, RIGHT_SURGERY, or BOTH_SURGERY. The rule-based method offers the advantage that, once the patterns are designed, labels can be automatically obtained for all reports. Therefore, we automatically extracted the rule-based labels for all the reports in our dataset (see Table 1). However, the primary challenge of this approach is the time-consuming nature of the pattern design process. Additionally, labels can not generalize beyond those patterns. This means enhancing it requires constant pattern updates for each institution or the creation of new patterns for specific institutions. As shown in Table 2, this method achieves suboptimal performance on the unseen institution of "Omicron" due to these limitations.

*Initial radiologist labels*

From the training institutions (all except "Omicron"), we engaged radiologists to annotate the initial set of ($\sim$ 7k) reports with surgery-related information (see Table 1). Notably, these annotations were derived solely from reading the reports, without referencing the corresponding images. For patients with surgery priors, radiologists provided annotations indicating the laterality of the surgery, which could be the left breast, right breast, or both. This observation led us to model the problem as a two-level task. The initial level focuses on the surgery presence task. The second level addresses the surgery laterality task.

*Extended hybrid labels*

The labels generated by radiologists show high quality. However, their production comes at a considerable cost. On the other hand, the rule-based method offers a low-cost alternative for label production. However, it has limitations in terms of quality and generalization. On the other hand, a substantial dataset of high-quality labels is crucial to effectively train deep learning models in this domain.

- **Automatic extension of labels** To expand our dataset, we initially trained a baseline Surgery_Presence model on the radiologist-annotated labels. These labels are of high quality, ensuring a reliable basis for training our initial model. The small size of this dataset, however, poses a challenge for training highly accurate deep learning models. Using this model, we then predicted labels for the remaining mammography reports. We particularly focused on instances where the model's predictions aligned with the rule-based labels and were made with high confidence. This strategy aimed to reinforce the model's existing confidence and equip it with extensive language knowledge of mammography reports. At the time of writing, there is no pre-trained language model specifically for mammography reports. Thus, we employed a general domain pre-trained French language model. By augmenting the training set with an additional $\sim$60k samples identified through this method, we enabled the model to better learn and adapt to the specific linguistic nuances of mammography reports (see Table 1).
- **Extension of labels by radiologist annotations** To further refine our model, we engaged in a second round of label generation by radiologists, this time utilizing uncertainty sampling. We identified model predictions where the model has the highest uncertainty in its predictions and in disagreement with the rule-based labels. These uncertain and conflicting cases were then annotated by radiologists. We ensured diversity in our dataset by selecting $\sim$2k additional samples uniformly at random from eight distinct institutions. It method targeted the model's weaknesses, making the learning process more robust and accurate. It also enhanced the variety of the training data.To ensure the highest quality of our dataset, we implemented an iterative process of label extension, which combines automatic methods with manual refinements by radiologists. After the initial round of automatic label extension and subsequent refinement through radiologist annotations, we reassessed the dataset and model performance. This process was repeated until convergence, defined as the stage where further refinements no longer improved model performance. In our experiments, convergence was achieved after 3 iterations, resulting in *refined labels*. This systematic approach ensures that our dataset grows not only in size but also in quality, significantly enhancing the reliability and generalization of our models.

## Training procedure

For all our models, we fine-tune all layers of the CamemBERT model for 4 epochs. The Surgery_Presence model takes the text of the mammography report as input. It assigns the text the class label NO_SURGERY or YES_SURGERY. SURGERY_LEFT, SURGERY_RIGHT and SURGERY_BOTH are considered YES_SURGERY labels. This model is trained using binary cross-entropy loss with class weight and Adam optimization with a learning rate of $2e-5$. We used scikit-learn's compute_class_weight with the "balanced" option to automatically adjust class weights inversely proportional to class frequencies. These weights were applied in the loss function to emphasize underrepresented classes during training. The learning rate of $2e-5$ was chosen according to established guidelines for fine-tuning BERT-based models, as recommended by[15]. For standard fine-tuning, we set the maximum number of epochs to 4, following[21], and saved the best model based on validation set performance for testing. For hybrid label training, we extended the number of epochs to 8 due to improved performance observed on the validation set. The Surgery_Laterality model is trained using categorical cross-entropy loss, considering class weights to address the slight imbalance in the class distribution. Throughout the training process, we periodically assess the model on the development set and save the checkpoint corresponding to the highest performance. All models undergo training using 2 NVIDIA Tesla $P$100 SXM2 16GB GPUs with a batch size of 32.

## Evaluation

Models are evaluated on their average performance at the following levels:

- Evaluation of each model: Models are individually evaluated in terms of average of macro-f1 over 5-fold cross-validation.
- Evaluation of final output on unseen dataset: We compare the performance of all models to a radiologist benchmark. To achieve this, we selected 300 reports uniformly random from four classes within the rule-based labels of the "Omicron" unseen institution. Subsequently, we engaged radiologists to annotate this sub-

set, and the performance of all experiments is reported based on this radiologist-annotated set. Outputs of both models on the this set are combined to get the final prediction of surgery prior information among 4-classes of No_Surgery, Left_Surgery, Right_Surgery, Both_Surgery and evaluation in terms of accuracy is reported.

## Experiments

### Rule-based model

EXP_Rule_Label is derived from the labels obtained by the rule-based method.

- **Results**, As previously noted, the rule-based method facilitates the automatic acquisition of labels for all reports, however, the quality of labels is limited. The performance of the rule-based model, particularly in terms of Precision, Recall, and F1 Score on the training sets, is detailed in Table 2. It's important to note that the rules for this model were handcrafted, relying only on insights derived from the training sources' reports. This ensures fairness and relevance in the development of the rule-based model. As shown in Table 2, EXP_Rule_Label achieves an accuracy of 0.82 on unseen data. Training models solely on rule-based labels is impractical due to their limited quality and reliability. These labels, created from simple hand-crafted rules often fail to reflect the complexity of real data, resulting in potential biases and unreliable outcomes. A neural network trained on such labels would likely replicate these inaccuracies, limiting its effectiveness and generalizability.

### Training on initial radiologist labels

The EXP_Rad_Label is obtained by training the model on the initial radiologists labels, and fine-tuning of all weights. Performance metrics are reported in terms of average macro-F1 score across 5-fold cross-validation. This baseline method served as a reference point for evaluating the efficacy of more advanced models in our research.

- **Results**, As shown in Table 2, EXP_Rad_Label achieves an average F1 of 0.84 for the Surgery_Presence model and an average F1 of 0.76 for the Surgery_Laterality model. Although the labels were prepared by radiologists, the small size of the training set limited the learning capacity of the model.

### GAN-CamemBERT: semi-supervised with adversarial learning

We conducted EXP_GAN_CamemBERT using the GAN-BERT model, as proposed by[22], which combines Generative Adversarial Networks (GANs) with BERT for text classification. We selected this model as a state-of-the-art baseline for semi-supervised text classification due to its demonstrated effectiveness. Specifically, we adapted the model by employing CamemBERT to handle our French reports, with a sequence length of 512 trained for 20 epochs. The GAN component was utilized to generate synthetic data that mirrors the distribution of the unlabeled data, while CamemBERT acted as the discriminator. Radiologist annotations served as labeled data, with the remainder treated as unlabeled.

- **Results**, As presented in Table 2, the GAN-CamemBERT model outperforms the supervised model using only CamemBERT, yet it still does not achieve better performance than the EXP_HYB_Label model. Notably, GAN-BERT has shown more improvements in previous studies when applied to shorter text, such as single sentences or tweets[22].

### Training on extended hybrid labels

The EXP_HYB_Label is obtained by training the model on the hybrid of radiologists and rule-based labels described in section "Labels", and fine-tuning of all weights. Performance metrics are reported in terms of average macro-F1 score across 5-fold cross-validation. The training of the model extended to 8 epochs. This decision was made because it performed slightly better on the development set. Typically, BERT-based fine-tuning tasks use the standard 4 epochs[21].

- **Results**, As shown in Table 2, EXP_HYB_Label achieves an average F1 of 0.91 and 0.88 for the Surgery_Presence model and the Surgery_Laterality model respectively, that is significantly higher than the performance of the baselines. These results confirm that the devised semi-automated proposed method effectively increased the size of the training data. It preserved label quality and achieved notable diversity and coverage. In the process of making these hybrid labels, a large portion was added automatically, which may contain noise. After we got the results of the EXP_HYB_Label, we gave the misclassification in each fold to radiologists for review. Then, we updated the refined labels in the dataset. We trained the models again with refined labels. We gave the misclassifications in each fold to radiologists for review. We repeated this process three times until the other labels did not change. As shown in Table 2, using these refined labels in EXP_HYB_REFINED_Label improved F1 by $\sim$5 points. This confirms that the model is sensitive to label quality.

### Biomedical pre-trained model

We investigate the impact of utilizing models pretrained on biomedical data. DrBERT[18] is the language model for French Biomedical applications. It is based on the RoBERTa architecture and pretrained on the French Biomedical corpus NACHOS. We initialized the weights of EXP_HYB_REFINED_Label_DrBERT by adopting weights from DrBERT instead of CamemBERT. We used the same training procedure as EXP_HYB_Label.

- **Results**, EXP_HYB_REFINED_Label_DrBERT as shown in Table 2, achieves an average F1 of 0.94 for the Surgery_Presence model and an average F1 of 0.91 for the Surgery_Laterality model. The decrease in performance observed with the model EXP_HYB_REFINED_Label_DrBERT could potentially be due to the use of a different vocabulary, varying sequence lengths, and other configuration differences.

### Data augmentation

In examining misclassified cases, a predominant pattern emerged. It was particularly prevalent in reports where family-related surgeries were reported. For instance, consider the following excerpt from a report about a patient without a personal history of surgery but with a grandmother who underwent breast surgery: “History of breast surgery in a paternal grandmother around the age of 65”. To address the model’s difficulty in correctly classifying reports with family-related surgery mentions, we augmented the training data by creating 1,000 synthetic samples. These samples were drawn from reports originally labeled as NO_SURGERY. We manually inserted references to family members’ surgeries into these reports while maintaining the NO_SURGERY label. This targeted augmentation improved the model’s ability to distinguish between a patient’s own surgical history and that of their relatives, resulting in better overall classification performance.

- **Results**, EXP_HYB_REFINED_Label_AUG, as shown in Table 2, achieves an average F1 of 0.99 for the Surgery_Presence model and an average F1 of 0.97 for the Surgery_Laterality model, which are higher than those of EXP_HYB_REFINED_Label, the latter being the same model without augmentation.

### History of surgery in other body organs

During the examination of misclassified samples, we identified an additional source of error wherein mammography reports contained histories of surgeries on other organs of the body. As demonstrated in the example, “History of ovarian cancer treated by hysterectomy in 2012”, this might lead to model confusion. To address this issue, we compiled a list of other body organs. During the pre-processing stage, we systematically excluded sentences mentioning these organs from the text of the report.

- **Result**, EXP_HYB_HYB_REFINED_AUG_ORG as shown in Table 2, achieves an average F1 of 0.99 for the Surgery_Presence model and an average F1 of 0.98 for the Surgery_Laterality model higher than that of EXP_HYB_HYB_REFINED_AUG.

### Analysis of class distribution across institutes

For each institute within our dataset, a part of its data was used for training purposes. We performed inferences using the EXP_HYB_REFINED_Label_AUG_ORG model to get the surgery priors of the remaining data from each source. Subsequently, we computed the class distribution over the full dataset of each source. The analysis of class distribution in Table 3 reveals consistent patterns in patients’ surgical history across five different institutes that conduct patient screenings. This consistency validates the model’s robustness. It indicates that the model could be effectively generalized to other institutes in the future. The distinct distribution observed in the beta institute’s data can be attributed to its combination of screening and diagnostic data. Unlike datasets that only include screening mammography for early detection in the general population without symptoms, the beta institute’s data also contains diagnostic mammography. Diagnostic mammography is performed following the detection of suspicious signs or symptoms, or after a positive screening mammogram. This means that the population subset within the beta institute’s data inherently has a higher likelihood of presenting conditions that may require surgical intervention.

To further assess the model’s generalizability and robustness, we run an experiment using a modified cross-validation method. Since ’Omicron’ is also a screening institution, we implemented 3-fold cross-validation where each fold corresponds to a specific screening institution (’zeta’, ’chi’, ’delta’) being used as the test set. The performance results of this cross-validation are added in Table 4, showing how well the model performs across different institutions, providing insights into its robustness and generalizability beyond the ’Omicron’ institution.

### Training times

For all our experiments using the hybrid labels, training Surgery_Presence model on $\sim$60k reports takes $\sim$80 minutes and training Surgery_Laterality model on $\sim$20k reports takes $\sim$30 minutes.

## Limitations

Our study has several limitations. First, our hybrid approach requires pre-existing rule-based labels. Second, our report labeler has a maximum input token size of 512 tokens. However, it may be easily extended to work with longer mammography reports. Third, a single board-certified radiologist determined the ground truth radiologist labels. Using more radiologists could provide a more accurate comparison to the radiologist benchmark. Fourth, we do test performance on a dataset from an institution unseen in training. Additional datasets across institutions could further establish the model’s ability to generalize.

## Conclusion and future work

In this study, we developed a new method for mammography report labeling. We integrated radiologist annotations with existing automatic rule-based labelers. Our approach uses a pretrained BERT-based model. The model was initially trained on a small, high-quality set of radiologist annotations. The model is then fine-tuned using an active learning loop. This loop merges radiologist annotations, selected through uncertainty sampling, with a large volume of automatic rule-based labels, obtained by agreement sampling. This method addresses the limitations of rule-based labels, which tend to be noisy and restrict model performance, and the challenges of obtaining high-quality but scarce radiology labels, which are costly and time-consuming to produce.

Our extensive experiments revealed two major sources of errors. The first was confusion between patients' and their relatives' medical histories. The second was the misinterpretation of surgeries on other body organs. We effectively addressed these by applying data augmentation for the former and preprocessing data to exclude irrelevant surgical history for the latter. Our results demonstrate that combining the quality of radiologist annotations with the scale of rule-based labels significantly enhances model performance. MammoBERT's architecture allows for adaptation to other medical report types and languages. By retraining on specific datasets and adjusting extraction rules, the model can be applied to different medical contexts, such as lung or brain imaging reports. Language-specific BERT models, like German-BERT[23] and Finnish-BERT[24], can facilitate adaptation to other languages, broadening MammoBERT's applicability across various medical domains and linguistic settings. Additionally, exploring the impact of image integration on the model's stability and accuracy could provide further insights into optimizing performance and be a valuable direction for future research.

## Data availability

The datasets used and analysed within this study are not publicly available due to legal restrictions but are available from the corresponding author on reasonable request.

## Code availability

The source code as well as the pre-trained models will be available on github.

Received: 22 April 2024; Accepted: 14 October 2024
Published online: 22 October 2024

## References

1. Ferlay, J. et al. Cancer statistics for the year 2020: An overview. *Int. J. Cancer* **149**, 778–789 (2021).
2. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. & Summers, R. M. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2097–2106 (2017).
3. Drozdov, I. et al. Supervised and unsupervised language modelling in chest X-ray radiological reports. *PLoS One* **15**, e0229963 (2020).
4. Wood, D. A., Lynch, J., Kafiabadi, S., Guilhem, E., Al Busaidi, A., Montvila, A., Varsavsky, T., Siddiqui, J., Gadapa, N., Townend, M. et al. Automated labelling using an attention model for radiology reports of mri scans (alarm). In *Medical Imaging with Deep Learning, PMLR*, pp. 811–826 (2020).
5. Martin, L., Muller, B., Ortiz Suarez, P. J., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D. & Sagot, B. Camembert: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020).
6. Pons, E., Braun, L. M., Hunink, M. M. & Kors, J. A. Natural language processing in radiology: a systematic review. *Radiology* **279**, 329–343 (2016).
7. Yadav, K. et al. Automated outcome classification of computed tomography imaging reports for pediatric traumatic brain injury. *Acad. Emerg. Med.* **23**, 171–178 (2016).
8. Hassanpour, S., Langlotz, C. P., Amrhein, T. J., Befera, N. T. & Lungren, M. P. Performance of a machine learning classifier of knee mri reports in two large academic radiology practices: a tool to estimate diagnostic yield. *Am. J. Roentgenol.*, 750–753 (2017).
9. Chen, M. C. et al. Deep learning to classify radiology free-text reports. *Radiology* **286**, 845–852 (2018).
10. Bozkurt, S., Alkim, E., Banerjee, I. & Rubin, D. L. Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm. *J. Digit. Imaging* **32**, 544–553 (2019).
11. Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F. & Buchanan, B. G. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. Biomed. Inform.* **34**, 301–310 (2001).
12. Irvin, J. et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 590–597 (2019).
13. Bustos, A., Pertusa, A., Salinas, J.-M. & De La Iglesia-Vaya, M. Padchest: A large chest X-ray image dataset with multi-label annotated reports. *Med. Image Anal.* **66**, 101797 (2020).
14. Chapman, B. E., Lee, S., Kang, H. P. & Chapman, W. W. Document-level classification of CT pulmonary angiography reports based on an extension of the context algorithm. *J. Biomed. Inform.* **44**, 728–737 (2011).
15. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
16. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **32** (2019).
17. Lee, J. et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).
18. Labrak, Y., Bazoge, A., Dufour, R., Rouvier, M., Morin, E., Daille, B. & Gourraud, P.-A. DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL'23), Long Paper* (Association for Computational Linguistics, 2023).
19. Buonocore, T. M., Parimbelli, E., Tibollo, V., Napolitano, C., Priori, S. & Bellazzi, R. A rule-free approach for cardiological registry filling from Italian clinical notes with question answering transformers. In *International Conference on Artificial Intelligence in Medicine*, 153–162 (Springer, 2023).
20. Jantscher, M. et al. Information extraction from German radiological reports for general clinical text and language understanding. *Sci. Rep.* **13**, 2353 (2023).
21. Sun, C., Qiu, X., Xu, Y. & Huang, X. How to fine-tune bert for text classification?. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, 194–206 (Springer, 2019).
22. Croce, D., Castellucci, G. & Basili, R. GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online* (eds. Jurafsky, D. et al.), 2114–2119 (2020). https://aclanthology.org/2020.acl-main.191. https://doi.org/10.18653/v1/2020.acl-main.191
23. Chan, B., Schweter, S. & Möller, T. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online)* (eds. Scott, D. et al.), 6788–6796 (2020). https://aclanthology.org/2020.coling-main.598. https://doi.org/10.18653/v1/2020.coling-main.598
24. Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F. & Pyysalo, S. Multilingual is not enough: Bert for finnish, arXiv preprint arXiv:1912.07076 (2019).

## Author contributions

N.D. designed and developed the models, carried out the experiments and wrote the manuscript. V.S. processed the experimental data and analysed the results. A.C. managed the team, verified the experiments and analysed the results. M.H. processed and analysed the data. G.O. processed and analysed the data. P.F. supervised the project. All authors discussed the results, provided critical feedback and helped shape the research, analysis and manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to N.D.