MOSAIC: A Multilingual, Taxonomy-Agnostic, and Computationally Efficient Approach for Radiological Report Classification

Alice Schiavone and **Desmond Elliott** and **Melanie Ganz** *(1) Department of Computer Science, University of Copenhagen* {alsc, de}@di.ku.dk
Marco Fraccaro and **Rasmus Bonnevie** *(3) Unumed Aps, Denmark*
Lea Marie Pehrson and **Silvia Ingala** *(4) Department of Diagnostic Radiology, Copenhagen University Hospital*
Michael Bachmann Nielsen *(5) Department of Clinical Medicine, University of Copenhagen*
Vincent Beliveau *(7) Institute for Human Genetics, Medical University of Innsbruck*

Abstract

Radiology reports contain rich clinical information for training imaging models without costly manual annotation. Existing methods have key drawbacks: rules miss linguistic variability, supervised models need large labeled sets, and LLM-based systems often rely on closed-source or resource-intensive models unsuitable for clinical use. They are also mostly English-only and limited to single taxonomies. We present MOSAIC, a multilingual, taxonomy-agnostic, and efficient approach for radiology report classification. Based on the compact open-access MedGemma-4B model, MOSAIC supports zero-/few-shot prompting and lightweight fine-tuning, running on consumer GPUs. Evaluated across seven datasets in English, Spanish, French, and Danish, it achieves a mean F1 of 88 on chest X-rays, matching or surpassing expert-level baselines while requiring only 24GB GPU memory.

Keywords: Natural Language Processing, LLMs, AI in Healthcare

1 Introduction

Deep learning in medical imaging requires large expert-labeled datasets, which are costly and scarce. Radiology reports, however, offer structured clinical knowledge produced in routine care [Reichenpfader et al., 2024, Zhou et al., 2014]. Earlier approaches include rule-based systems and BERT classifiers [Irvin et al., 2019, Smit et al., 2020], which reach strong performance but demand handcrafted rules or extensive labels, making adaptation to new taxonomies or languages resource-heavy [Yang et al., 2023]. Recent LLMs enable zero-/few-shot classification, reducing manual labeling and improving adaptability [Gu et al., 2024, Dorfner et al., 2024]. Yet most rely on large or closed models, raising deployment and privacy concerns. Smaller open models are more practical

Dataset		Language	Number of Findings	Avg. Chars	Train	Dev	Test
MIMIC-CXR	M	en	14	760	535	50	100
PadChest-GR	P	es, en	49	115	1951	100	879
CASIA-CXR	C	fr	5	400	7677	100	3334
DanskCXR	D	da	48	312	1600	125	750
Reflacx^{I}	R^{I}	en	14	216	68	50	120
Reflacx^{II}	R^{II}	en	15	201	1046	52	1098
DanskMRI	B	da	3	1941	194	50	345

Table 1: Overview of the datasets used, including language, number of findings, average characters per report, and data split.

for healthcare. We propose MOSAIC, a lightweight, multilingual framework for radiology report classification. It runs locally on consumer GPUs, adapts to diverse taxonomies and modalities, and is open-source[1].

2 Methods

Few public radiological reports datasets are currently available, due to the risk of de-anonymization of patients or clinicians. For high-quality results, we consider only radiologist-validated sets: *MIMIC* [Johnson et al., 2019], *PadChest* [Castro et al., 2024], *CASIA* [Metmer and Yang, 2024], *REFLACX* [Bigolin Lanfredi et al., 2022], *DanskCXR* [Schiavone, 2025], and *DanskMRI* [Beliveau et al., 2024] (see Table 1). We compared Llama 3 (8B) [Grattafiori et al., 2024], MedGemma (4B) and Gemma 3 (12B) [Google, 2025], selecting their instruction-tuned versions for better structured outputs [Zhang et al., 2025]. Each model is fine-tuned on an NVIDIA RTX 4090 GPU (24GB) using its 4-bit quantized form along with Rank-Stabilized LoRA adapters. All the prompts and findings sets are written in English to leverage the models' stronger alignment to English instructions. The prompt structure is adapted based on the dataset's label taxonomy. For example, the prompt for *CASIA* is:

```
You are a helpful radiology assistant. Given a radiology report, classify
each abnormality into a class. Output a valid JSON with each abnormality
as key, and the class as value. The keys must be ['cardiomegaly', 'mass',
'pleural effusion', 'pneumonia', 'pneumothorax']. The values can be one of
[-1, 1]. The values have the following interpretation: (1) the abnormality
was mentioned, even with uncertainty, in the report e.g. 'A large pleural
effusion', 'The cardiac contours are stable.', 'The cardiac size cannot be
evaluated.'; (-1) the abnormality was not mentioned in the report, or the
abnormality was negatively mentioned in the report; e.g. 'No
pneumothorax.'.
```

Fine-tuning is conducted using the Unsloth library [Han et al., 2023], while inference is performed with vLLM [Kwon et al., 2023]. Before inference, the LoRA adapters are merged into the base models in 16-bit precision. Cross-entropy loss is used as the objective function. Training configurations, prompt-

[1]Code: `github.com/aliswh/mosaic`, Models: `huggingface.co/AliceSch/mosaic-4b`

Dataset	M		P_E		P_S		C		D		*Average*	
Experiment	ZS	3S	ZS	3S	ZS	3S	ZS	3S	ZS	3S	ZS	3S
`Llama-8B`	54	61	77	76	69	67	70	75	61	63	66	68
`MedGemma-4B`	55	59	61	77	53	72	69	82	62	65	60	69
`Gemma-12B`	65	70	79	79	76	76	76	76	69	75	73	75
`Gemma-27B` $*$	68	69	81	81	81	82	75	77	71	75	75	77
`Llama-70B` $*$	69	72	78	79	74	70	68	79	68	73	71	75

Table 2: Classification performance of language models on chest X-ray radiological free-text reports, measured in F1 and ordered by model family and size. Models are tested under zero-shot (ZS) and three-shot (3S), with three examples drawn from corresponding training sets. We indicate with $*$ models fine-tuned on a 94GB H100, instead of a consumer-grade 24GB RTX 4090.

	M			P_E			P_S			C			D		
Datasets	M4	L8	G12	M4	L8	G12	M4	L8	G12	M4	L8	G12	M4	L8	G12
M	87	86	84	74	78	80	71	71	76	82	77	81	58	65	69
MP_E	78	**89**	85	92	**95**	**95**	84	85	91	83	74	80	65	65	70
MP_{E+S}	80	82	84	93	94	**95**	92	**94**	**94**	83	76	79	61	61	69
$MP_{E+S}C$ $\star$	76	85	84	92	**95**	94	89	**94**	**94**	**99**	**99**	**99**	66	63	69
$MP_{E+S}CD$	76	76	85	93	93	**95**	91	92	**94**	**99**	**99**	**99**	82	84	**86**

Table 3: F1 scores with incrementally expanded multilingual training configurations on chest X-ray radiological reports. In bold, the best result over each dataset. The symbol $\star$ indicates the training configuration used for MOSAIC. M4= `MedGemma-4B`, L8= `Llama-8B`, G12= `Gemma-12B`

ing strategy and hyperparameters are documented in the accompanying code repository. We use the mean macro F1 score to evaluate the extraction of findings for positively mentioned findings.

3 Results

On five chest X-ray datasets, 3-shot prompting outperformed zero-shot (Table 2). Gemma 12B gave the best overall results, while MedGemma 4B remained competitive despite its size, occasionally surpassing larger models. Llama 8B was stable but generally lower. Larger 27B/70B models further improved scores, but require high-end GPUs. Multilingual fine-tuning improved generalization across languages and taxonomies (Table 3). English-only training generalized to Spanish PadChest, and adding more datasets further improved scores. CASIA, with its large size and simple task, reached near-perfect performance. On MIMIC, MOSAIC achieved similar results to rule-based, BERT, and CheXbert baselines, achieving 0.88 F1 on positive findings.

M	P_E		R^I	$\Rightarrow R^I$	R^{II}	$\Rightarrow R^{II}$
77	100	Consolidation	67	91	65	92
93	✗	Pneumothorax	90	100	84	93
✗	68	Nodule	47	60	✗	✗
✗	100	Hiatal Hernia	✗	✗	89	100
✗	✗	Emphysema	71	50	✗	✗
✗	✗	Enlarged Hilum	✗	✗	74	75

Figure 1: Taxonomy adaptation in English of *MOSAIC* trained on $MP_{E+S}CD$, measured in F1. Left columns show performance on present findings in training sets M and P_E; right columns show generalization to unseen datasets R^I and R^{II} before and after fine-tuning ($\Rightarrow$R).

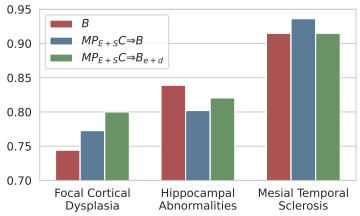

Figure 2: Performance of *MOSAIC* on the *DanskMRI* dataset, measured as F1 across three epilepsy-related abnormalities from MRI reports.

Table 1 looks at the performance of fine-tuned `MedGemma-4B` trained on unseen English datasets ($R^I$and R^{II}), both before and after fine-tuning on new data ($\Rightarrow$R). The left columns (M and P_E) reflect the model's initial task competency on English-language findings. The red "X" marks indicate that these specific findings are *not present* in that dataset taxonomy. On findings that are present in only one of the training sets ("Nodule" and "Pneumothorax"), fine-tuning improves the models' ability to adapt their existing task competency to new distributions. However, for findings *not* included in the initial training set taxonomy, the fine-tuned model shows a small improvement ("Enlarger Hilum") or, in the case of "Emphysema" (N=6), fine-tuning hurts performance.

The *DanskMRI* dataset consists of Danish MRI reports annotated for three epilepsy-related brain abnormalities. Unlike chest X-ray datasets, these findings relate to neurological imaging, introducing both clinical and linguistic shifts. As shown in Figure 2, adaptive fine-tuning on external chest X-ray datasets ($MP_{E+S}C$) improves performance over the base model for Focal Cortical Dysplasia and Mesial Temporal Sclerosis. B_{e+d} is the same dataset, extended with the same reports machine-translated to English. This data augmentation technique improves consistency across all findings. In particular, it recovers performance on Hippocampal Abnormalities. These results highlight the benefit of lightweight augmentation when adapting to new modalities, especially under language and data constraints, as only 194 examples are provided for fine-tuning.

4 Conclusion

We introduce MOSAIC, a framework for classifying radiology reports across languages, taxonomies, and modalities. Built on compact open models, it runs on consumer GPUs. Evaluations on English, Spanish, French, and Danish datasets show robust performance comparable to previous methods. We invite the community to extend MOSAIC for broader clinical applications.

References

V. Beliveau, H. Kaas, M. Prener, C. N. Ladefoged, D. Elliott, G. M. Knudsen, L. H. Pinborg, and M. Ganz. Classification of radiological text in small and imbalanced datasets in a non-english language. *arXiv preprint arXiv:2409.20147*, 2024.

R. Bigolin Lanfredi, M. Zhang, W. F. Auffermann, J. Chan, P.-A. T. Duong, V. Srikumar, T. Drew, J. D. Schroeder, and T. Tasdizen. Reflacx, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays. *Scientific data*, 9(1):350, 2022.

D. C. Castro, A. Bustos, S. Bannur, S. L. Hyland, K. Bouzid, M. T. Wetscherek, M. D. Sánchez-Valverde, L. Jaques-Pérez, L. Pérez-Rodríguez, K. Takeda, et al. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *arXiv preprint arXiv:2411.05085*, 2024.

F. J. Dorfner, L. Jürgensen, L. Donle, F. A. Mohamad, T. R. Bodenmann, M. C. Cleveland, F. Busch, L. C. Adams, J. Sato, T. Schultz, et al. Is open-source there yet? a comparative study on commercial and open-source llms in their ability to label chest x-ray reports. *arXiv preprint arXiv:2402.12298*, 2024.

Google. Medgemma hugging face. `https://huggingface.co/collections/google/medgemma-release-680aade845f90bec6a3f60c4`, 2025. Accessed: [Insert Date Accessed, e.g., 2025-05-20].

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

J. Gu, H.-C. Cho, J. Kim, K. You, E. K. Hong, and B. Roh. Chex-gpt: Harnessing large language models for enhanced chest x-ray report labeling. *arXiv preprint arXiv:2401.11505*, 2024.

D. Han, M. Han, and U. team. Unsloth. `http://github.com/unslothai/unsloth`, 2023. Software.

J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

A. Johnson, M. Lungren, Y. Peng, Z. Lu, R. Mark, S. Berkowitz, and S. Horng. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*, 101:215–220, 2019.

W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

H. Metmer and X. Yang. An open chest x-ray dataset with benchmarks for automatic radiology report generation in french. *Neurocomputing*, 609:128478, 2024.

D. Reichenpfader, H. Müller, and K. Denecke. A scoping review of large language model based approaches for information extraction from radiology reports. *npj Digital Medicine*, 7(1), Aug. 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01219-0. URL http://dx.doi.org/10.1038/s41746-024-01219-0.

A. e. a. Schiavone. Effective machine learning techniques for non-english radiology report classification: A danish case study. *AI*, 6(2), 2025. ISSN 2673-2688. doi: 10.3390/ai6020037. URL https://www.mdpi.com/2673-2688/6/2/37.

A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *CoRR*, abs/2004.09167, 2020. URL https://arxiv.org/abs/2004.09167.

E. Yang, M. D. Li, S. Raghavan, F. Deng, M. Lang, M. D. Succi, A. J. Huang, and J. Kalpathy-Cramer. Transformer versus traditional natural language processing: how much data is enough for automated radiology report classification? *The British Journal of Radiology*, 96(1149):20220769, 2023.

S. Zhang, P. Sun, S. Chen, M. Xiao, W. Shao, W. Zhang, Y. Liu, K. Chen, and P. Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. In *European Conference on Computer Vision*, pages 52–70. Springer, 2025.

Y. Zhou, P. K. Amundson, F. Yu, M. M. Kessler, T. L. S. Benzinger, and F. J. Wippold. Automated classification of radiology reports to facilitate retrospective study in radiology. *Journal of Digital Imaging*, 27(6):730–736, dec 2014. doi: 10.1007/s10278-014-9708-x.