From Text to Insight: Classifying Microcalcifications in Radiology Reports with AI

Zardar Khan (Physical Sciences, Sunnybrook Research Institute, Canada)
Grey Kuling (Department of Medical Biophysics, University of Toronto)
Anne Martel (Department of Medical Physics, University of Toronto, Canada)

Introduction: Radiology departments generate millions of unstructured free-text reports containing valuable clinical information including cancer history, imaging modalities, and examination types, presenting significant untapped potential for research cohort discovery in breast imaging. However, manually extracting and analyzing this information remains time-consuming and prone to human error. The Breast Imaging Reporting and Data System (BI-RADS), developed by the American College of Radiology, provides standardized reporting guidelines that add inherent structure to these reports [1]. In this study, we focus specifically on automating the extraction of microcalcification information – a crucial indicator for early breast cancer diagnosis and treatment planning. Inspired by recent advances in artificial intelligence, particularly in natural language processing, we explored automated approaches to extract this structured information from these reports.

Methods: We evaluated three approaches to classify microcalcification status in breast imaging reports. The first two employed supervised learning with encoder models and classification heads to label patient status as positive, negative, or not stated. The first approach, conducted in prior work by Grey K., included segmenting reports according to BI-RADS structure to reduce input sequence length into BERT, a bi-directional encoder model that captures the context of words in all directions [2]. The second approach employed RadBERT, a model pre-trained on 4M radiology reports, processing reports with 512-token truncation. Our third approach explored zero-shot and few-shot capabilities of Large Language Models (LLMs) including Yi-34B, Mixtral 8x22B (MoE), Meditron-70B and Qwen-72B, all sourced from the HuggingFace repository [3]. These LLMs are known for their extensive pre-training on vast and diverse datasets. We evaluated LLM performance using unnormalized log likelihood scoring, while encoder models were assessed using classification accuracy.

Results: Performance varied across approaches, with encoder models achieving the highest accuracy. As shown in Figure 1, Yi-34B demonstrated strong baseline performance with 76% zero-shot accuracy, improving to 79% with few-shot learning. Other LLMs showed mixed results: Qwen-72B (46% to 60%), Mixtral 8X22B (50% to 72%), and Meditron-70B (72% zero-shot, declining to 34% few-shot, 61% with prompt-tuning). BERT models demonstrated superior performance, with Gatortron and RadBERT achieving 94% weighted accuracy, while the AWD-LSTM baseline reached 75%. Some few-shot experiments were selectively conducted based on initial performance and practical considerations.

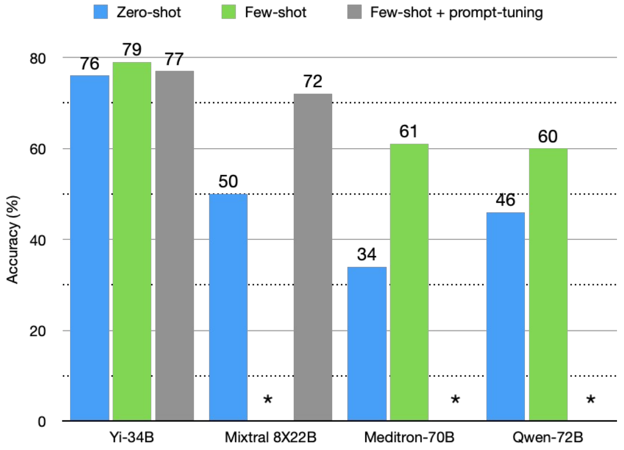

Figure 1 Performance comparison of LLMs (Yi-34B, Qwen-72B, Mixtral 8x22B [MoE], Meditron-70B) on microcalcification classification. Accuracy (%) shown for zero-shot, few-shot, and few-shot + prompt-tuning approaches. Missing data points (*) indicate experiments not conducted due to resource constraints or initial performance considerations.

Conclusions: Our study demonstrates that relying less on structured preprocessing and more on the latent capabilities of LLMs offers promising results for medical text classification. While BERT-based models achieved the highest accuracy at 94%, the strong performance of Yi-34B (79% accuracy) with minimal tuning suggests efficient paths for medical NLP deployment. Notably, larger parameter counts did not necessarily equate to better performance, as evidenced by Yi-34B outperforming larger models. Transformer-based architectures consistently outperformed traditional LSTM approaches in this medical text classification task. Future research will focus on optimizing these models for broader medical applications while maintaining deployment efficiency.

Acknowledgements: A special acknowledgment to the Biomatrix Team and Richard Mraz’s Information Services Team.

References: [1] (www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads) [2] BI-RADS BERT and Using Section Segmentation to Understand Radiology Reports (2022), Kuling et al. [3] (https://huggingface.co/models).