# A Multimodal Deep Learning Ensemble Framework for Building a Radiology Triage System

Mahvash SIAVASHPOUR [a,1], Erin MCCABE [b], Andrew NATARAJ [c], Niketh PAREEK [d], Osmar R. ZAIANE [a], Douglas GROSS [b]

[a] *Computing Science, University of Alberta*
[b] *Rehabilitation Medicine, University of Alberta*
[c] *Medicine and Dentistry, University of Alberta*
[d] *McGill University*

**Abstract.** Spinal radiology reports serve as crucial resources for medical decision-making for patients experiencing low back and neck pain. Additionally, physician-completed questionnaires evaluating patients' conditions offer valuable insights for shaping medical care plans. However, the imbalance between the number of physicians and patients often slows the processing of these data, delaying referrals or medical decisions. This can result in individuals with severe conditions experiencing a deterioration in their health before receiving professional care. In this work, we propose an ensemble transformer-based framework built on top of pre-trained BERT-based models which can classify patients on their need for surgery given their different data modalities including radiology reports and questionnaires. We conducted our experiments using data collected from the Kaye Edmonton Spine Assessment Clinic. Our results show that our approach effectively leverages the diverse information across multiple data modalities, achieving an F1 score of 76.09 and surpassing the performance of competing methods.

**Keywords.** BERT, Transformer, Spinal Radiology Report, Medical Questionnaires, Patient Triage

## 1. Introduction

Radiology is a branch of medicine focused on diagnosing and treating diseases through imaging technologies such as X-rays, computed tomography (CT), and magnetic resonance imaging (MRI). These images are typically accompanied by radiology reports, which are written by radiologists based on their interpretation of the medical images. In addition to imaging, tabular data such as physician-completed questionnaires is another common format in the medical field. Radiology reports and questionnaires are critical for guiding medical interventions and tracking treatment progress. However, manual processing of the medical records for decision-making is time consuming and labor-intensive. Additionally, the variability in data style and format across healthcare centers,

[1] Corresponding Author: University of Alberta, Edmonton, Canada; E-mail: siavashp@ualberta.ca

as well as changes over time, presents significant challenges in developing an automated classification system for patient triage.

Machine learning has significantly advanced decision-making in precision health, particularly in the processing of radiology reports for various applications. While early approaches would mainly focus on rule-based methods [1], the advent of deep learning and the Transformer architecture marked a substantial step forward in automating these processes [2, 3, 4]. Leveraging these architectures, several studies have proposed pre-trained models by utilizing the vast amounts of biomedical texts available [5, 6], while others have developed models that incorporate radiology-specific texts into their training data to better capture the specialized vocabulary of radiology reports [7, 8].

Another line of work puts focus on processing tabular data in the medical domain, by utilizing feature selection methods such as information gain at each decision [9] or attention mechanisms [10]. However, the dynamic nature of medical questionnaires presents challenges for these approaches, as feature selection becomes less effective when the dataset's features change over time.

Moreover, while machine learning applications in radiology report processing have been extensively studied in areas like disease and abnormality detection [11], including the assessment of thyroid nodule malignancy scores [9], their application to patient triage systems remains comparatively underexplored. Developing an effective triage system requires a deeper understanding of key features influencing decision-making, as the data often lacks explicit answers. A related study examined the use of natural language processing (NLP) techniques, such as TF-IDF and word2vec embeddings, in combination with machine learning classifiers to automate the triaging of textual radiology reports [12]. However, their results suggested that simple clinical variables outperformed the NLP models in predictive accuracy.

This study aims to bridge the gap between decision-making based on different data modalities in the medical domain by developing an ensemble of expert models trained on a variety of data sources, including the *impression* and *findings* sections of radiology reports, as well as tabular questionnaires. By using the unique information from each modality, the individual models function as medical assistants, providing complementary insights that enhance the final decision-making process.

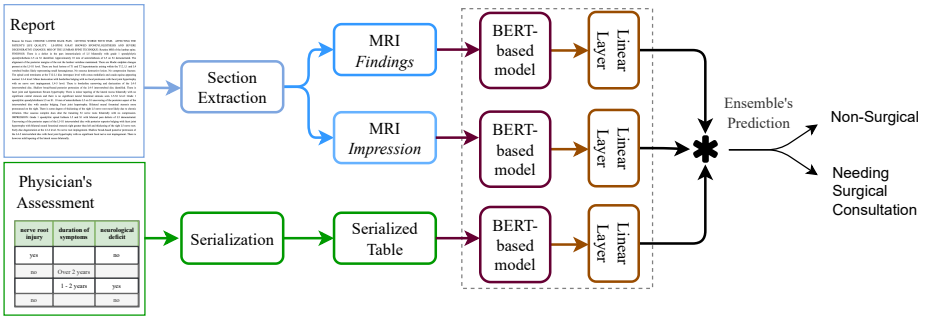



**Figure 1.** The overview of the classification pipeline

## 2. Methods

This study aims to develop a classification model to automatically categorize spinal injury patients into two groups: those requiring surgical referral and those who do not. Specifically, we propose a binary classification system that leverages both textual radiology reports and medical questionnaires. The radiology reports consist of unstructured free-text summaries detailing MRI findings of the spine, while the physicians' assessments are captured in tabular format through questionnaires, providing insights into the patient's primary issues. Each radiology report can be broken down into different sections [13], including the *scan type*, *comparisons* to the previous scans, *techniques* and the procedure used during the imaging, *findings* detailing observations from the patient's image, and *impression* summarizing all the information gathered by the radiologist. In this study, we focus on extracting the *impression* and the *findings* sections of the reports, as they provide the most clinically relevant information for determining the need for a surgical referral.

### *2.1. Data*

The data used in this study was obtained from the electronic medical records of patients at the Kaye Edmonton Spine Assessment Clinic. It includes records for 880 patients collected between 2020 and 2024. Out of all the patients, 487 were classified as needing surgical consultation, while 393 were determined to be non-surgical cases. Missing data accounted for 5.56% of the radiology reports and 36.47% of the physicians' assessments. Previous studies showed a positive correlation between a non-surgical case and missing data, meaning the patients are more likely not to be referred to a surgeon if some parts of their records are missing [12]. As a result, we only removed the patient where no data was available and kept partially missing data in the dataset.

After preprocessing the data, the *impression* and the *findings* sections of the reports are extracted. Furthermore, the tabular data from the physicians are serialized into textual format to match the pre-training data distribution of the BERT models. For serialization, we use a text template in the format of either *"column name is value"* or *"has column name"* for categorical and binary features, respectively.

### *2.2. Model Architecture*

All models used in the experiments are variations of the BERT architecture [14]. However, we report results only for the best-performing pre-trained BERT model. Figure 1 illustrates the overall architecture of our ensemble system. For each data modality, a pre-trained BERT-based model with a dedicated linear classification head is utilized. For the textual reports, we train separate BERT models for the *findings* and the *impression* section as previous research has demonstrated that each provides a distinct perspective, improving the final performance [3]. In each model, the [CLS] token output from BERT is passed to the classification head, generating class probabilities based on the data modality it processes. The final class prediction is determined by averaging the probabilities from the predictions made by each model.

### *2.3. Training*

We train three BERT-based models, one for each of the three data modalities: radiology *impression*, radiology *findings* and serialized tabular data. We train the models using the AdamW optimizer with a learning rate of $1e-5$ and cross entropy loss. Early stopping is applied to ensure that the best-performing model is selected after each training iteration. All model parameters are fine-tuned during the training process.

## 3. Results

In order to get robust results despite having a small dataset, we perform 5-fold cross-validation. Experiments are done on a number of BERT-based models including $\text{BERT}_{\text{base}}$ [14], RadBERT [7], BioBERT [5], BioClinicalBERT and BioDischargeSummary [8]. However, we only report results on RadBERT, which outperformed other BERT-based models on our dataset. We evaluate our pipeline against the publicly available pre-trained BERT-based model. We also compare our results to the word embedding averaging method combined with logistic regression, as explored in a previous study on the same surgical referral task [12]. As depicted in Table 1, our ensemble system achieves an F1 score of 76.09 and ROC AUC of 68.03, outperforming the baselines.

**Table 1.** Performance comparison of the proposed ensemble model and the baselines

| **Model** | **Variation** | **F1 Score** | **ROC AUC** |
|---|---|---|---|
| Word Embedding + Logistic Regression | Impression | 66.84 | 58.28 |
| | Findings | 67.65 | 58.38 |
| RadBERT + Logistic Regression | Impression | 67.53 | 62.51 |
| | Findings | 64.43 | 60.73 |
| Fine-tuned RadBERT | Impression | 74.46 | 63.09 |
| | Findings | 72.34 | 65.11 |
| | Tabular | 70.46 | 64.80 |
| Our Fine-tuned RadBERT Ensemble Model | Impression + Findings + Tabular | **76.09** | **68.03** |

## 4. Discussion

Comparing the results of our ensemble approach with those of individual BERT-based models demonstrates that incorporating multiple modalities of medical data leads to more informed and accurate decisions than relying on a single data source. Furthermore, models trained on the *impression* section of the radiology reports on average showed better performance than those trained on the *findings* section. This implies that the *impression* section provides more critical information for the task of determining surgical referrals.

## 5. Conclusions

In this study, we proposed an ensemble system that averages the class probabilities obtained from three fine-tuned BERT-based models, each trained on a different data modal-

ity. Our approach demonstrated improved performance compared to using isolated fine-tuned BERT-based models. Additionally, when compared to a previous study on similar data from the Spine Assessment Clinic, our method showed a significant performance improvement.

## References


[1] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics. 2001 Oct;34(5):301-10.

[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. arXiv; 2023. ArXiv:1706.03762 [cs].

[3] Han Y, Yuan Y, Mou L. A Dual-View Approach to Classifying Radiology Reports by Co-Training. arXiv; 2024. ArXiv:2406.05995 [cs].

[4] Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. arXiv; 2020. ArXiv:2004.09167 [cs].

[5] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb;36(4):1234-40.

[6] Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare. 2022 Jan;3(1):1-23. ArXiv:2007.15779 [cs].

[7] Yan A, McAuley J, Lu X, Du J, Chang EY, Gentili A, Hsu CN. RadBERT: Adapting Transformer-based Language Models to Radiology. Radiology: Artificial Intelligence. 2022 Jul;4(4):e210258.

[8] Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, McDermott M. Publicly Available Clinical BERT Embeddings. In: 2nd Clinical Natural Language Processing Workshop; 2019. p. 72-8.

[9] Santos T, Kallas ON, Newsome J, Rubin D, Gichoya JW, Banerjee I. A Fusion NLP Model for the Inference of Standardized Thyroid Nodule Malignancy Scores from Radiology Report Text. AMIA Annual Symposium Proceedings. 2022 Feb;2021:1079-88.

[10] Arik SO, Pfister T. TabNet: Attentive Interpretable Tabular Learning. arXiv; 2020. ArXiv:1908.07442 [cs, stat].

[11] Lau W, Lybarger K, Gunn ML, Yetisgen M. Event-Based Clinical Finding Extraction from Radiology Reports with Pre-trained Language Model. Journal of Digital Imaging. 2023 Feb;36(1):91-104.

[12] Krebs B, Nataraj A, McCabe E, Clark S, Sufiyan Z, Yamamoto SS, Zaïane OR, Gross DP. Developing a triage predictive model for access to a spinal surgeon using clinical variables and natural language processing of radiology reports. European Spine Journal. 2023 Feb.

[13] Kahn CE, Langlotz CP, Burnside ES, Carrino JA, Channin DS, Hovsepian DM, Rubin DL. Toward best practices in radiology reporting. Radiology. 2009 Sep;252(3):852-6.

[14] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2019. ArXiv:1810.04805.