VALID: Verified AI for Limiting Inefficient Diagnostics

Shlok Panchal
McMaster University
panchs10@mcmaster.ca

Emily McElheran
McMaster University
mcelhere@mcmaster.ca

Sama Al-Oda
McMaster University
alodas@mcmaster.com

Akshaj Shrotri
McMaster University
shrotria@mcmaster.ca

Krish Bhagirath
McMaster University
bhagirak@mcmaster.ca

Krish Shah
McMaster University
shahk76@mcmaster.ca

Sehaj Ajimal
McMaster University
ajimals@mcmaster.ca

Sophia Duda
McMaster University
dudas1@mcmaster.ca

Trisha Raj
McMaster University
rajt1@mcmaster.ca

Yasmine Zitouni
McMaster University
zitouniy@mcmaster.ca

***Abstract*—The overuse of diagnostic imaging, particularly computed tomography (CT) scans, is a significant issue in Canadian healthcare, leading to increased costs, prolonged wait times, and unnecessary radiation exposure. Despite guidelines, up to 30% of scans are considered low value, highlighting a gap in clinical decision support. This study introduces VALID (Verified AI for Limiting Inefficient Diagnostics), a multimodal machine learning framework designed to prospectively predict CT scan necessity at emergency department triage using data (vitals and unstructured clinical notes) at time of admission to the ED. To address data leakage, we utilize a two-phase architecture, first extracting ground-truth labels retrospectively from radiology reports using ClinicalBERT, and then training an ensemble of classifiers (MLP, Random Forest, XGBoost) on the pre-scan data. Our ensemble model achieved a macro-average recall of 88% and precision of 91%, effectively balancing the identification of necessary scans with a high overall precision. VALID acts as an early-warning clinical decision support tool to safely reduce low-value imaging while maintaining diagnostic accuracy.**

I. Introduction

The overuse of diagnostic imaging, particularly computed tomography (CT) scans, has become a growing issue in Canadian healthcare systems. CT imaging is extremely valuable when used appropriately and is vital in identifying strokes, hemorrhages, and other life-threatening conditions. However, evidence shows that many scans are ordered without clear clinical justification, and up to 30% of medical tests and procedures are considered low value, offering no clinical benefit to the patient [1]. This pattern of superfluous diagnostic imaging contributes to unnecessary radiation exposure, increased healthcare costs, and prolonged waiting times for patients who need urgent diagnostic evaluation.

A. Motivation

A study analyzing 11,824 outpatient CT and 11,867 MRI scans across randomly selected hospitals in Ontario found that less than 2% of CT scans ordered for headaches identified abnormalities that could explain the patient's symptoms [2]. Moreover, the study revealed significant variation in imaging practices with as much as a 70-fold difference between hospitals in the frequency of scans ordered for the same indication [2]. Such variability suggests that imaging decisions are influenced by institutional practice patterns, not just clinical presentation.

Similarly, the Institute of Health Economics reported that in Alberta, 54% of lumbar spine MRI requisitions in the 2017-2018 fiscal year were considered inappropriate according to guideline-based criteria [3]. Importantly, only a small proportion of lower back pain cases (10%) require diagnostic imaging or specialist consultation, as most cases resolve with conservative management [3]. The discrepancy between recommended practice and real-world use demonstrates a significant and persistent gap between evidence-based guidelines and clinical decision-making.

Beyond financial inefficiency, unnecessary imaging exposes patients to ionizing radiation and increases cumulative lifetime cancer risk. CT scans deliver significantly higher radiation doses than conventional X-rays, and repeated exposure has been associated with increased risk of radiation-induced malignancies, especially in younger patients. Additionally, incidental findings can trigger several follow-up tests and invasive procedures that may not ultimately improve patient outcomes. Unnecessary imaging also contributes to congestion within radiology departments, which delays access for patients with time-sensitive conditions. When imaging capacity is limited, even moderate overuse can scale up into a system-wide strain.

To address this systemic bottleneck, our project introduces VALID (Verified AI for Limiting Inefficient Diagnostics), an intelligent decision-support framework that prospectively predicts the clinical utility of a CT scan at the point of triage, empowering physicians to safely reduce low-value imaging without missing critical diagnoses.

B. Related Works

Advances in natural language processing (NLP) have recently demonstrated the ability to extract clinically meaningful information from unstructured medical documentation. Recent research has shown that NLP models applied to radiology reports and post-imaging clinical notes can accurately identify diagnostic outcomes and patient status from textual data [4]. These findings indicate that physician documentation contains

structured patterns that can be assessed computationally to determine diagnostic relevance.

However, while existing systems are highly effective at retrospective analysis, identifying overuse after the scan has been performed and the report has been written, there remains a critical gap in prospective decision support.

The evidence indicates that variability and overuse in diagnostic imaging are systemic issues that stem from inconsistent decision support at the point of care. Our project aims to bridge this gap by developing a machine learning framework that integrates structured clinical variables with NLP-derived representations of physician notes to support evidence-based imaging decisions. The system is designed as a clinician-assistive decision support tool, with final imaging decisions to remain under physician judgment. Model calibration prioritizes patient safety and minimizing missed critical diagnoses. This approach aims to reduce unnecessary imaging while maintaining diagnostic accuracy and improving healthcare efficiency.

C. Problem Definition

While existing literature successfully leverages natural language processing (NLP) to retrospectively identify imaging overuse, these approaches fundamentally rely on post-scan documentation, such as the final radiology reports. This introduces significant data leakage if applied to point-of-care decision-making. At the time of an emergency department (ED) admission, a clinician does not have access to diagnostic imaging results, but rather a limited, noisy set of initial triage data.

Formally, let $X = \{T, V\}$ represent the multimodal input space available at the time of ED triage, where T denotes the unstructured text of the patient's chief complaint and V represents the structured array of initial physiological vital signs (e.g., heart rate, oxygen saturation). Let $y \in \{0, 1\}$ represent the objective ground-truth necessity of a CT scan, previously extracted via retrospective NLP, where $y = 1$ indicates an acute, actionable finding and $y = 0$ denotes a low-value or unnecessary scan.

Our objective is to learn a predictive mapping function $f(X; \theta)$ parameterized by θ that estimates the true diagnostic utility of a prospective scan,

$$P(y = 1|T, V) = f(T, V; \theta), \tag{1}$$

where the model outputs the probability of a scan yielding actionable clinical results. Because this system operates in a high-stakes medical environment, it is unacceptable to erroneously deny necessary imaging. Therefore, the final binary classification $\hat{y}$ is determined by a strategically calibrated decision threshold τ,

$$\hat{y} = \begin{cases} 1, & \text{if } P(y = 1|T, V) \geq \tau \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where τ is chosen to aggressively minimize false negatives.

Given a historical dataset of ED triage records mapped to their subsequent radiology outcomes, the core challenge is developing a multimodal model that learns $f(X; \theta)$ effectively without relying on post-scan data. The ideal solution should:

- Accurately identify and flag low-value CT scan orders ($\hat{y} = 0$) based exclusively on the prospective input space X.
- Ensure the system reliably approves high-value, necessary scans ($\hat{y} = 1$) for patients presenting with acute or critical conditions.
- Avoid overly aggressive filtering thresholds that could compromise patient safety, delay urgent care, or degrade the physician's clinical workflow.

II. Methodology

Figure 1 illustrates the overarching architecture of the VALID framework, detailing the flow of multimodal triage data from initial extraction to final predictive classification.

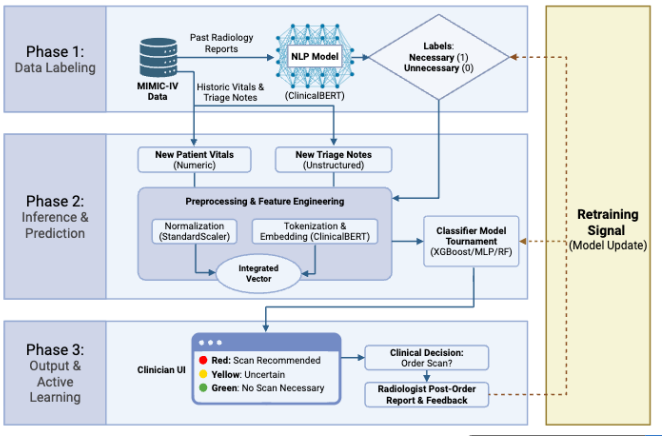

Fig. 1. VALID system overview: Multimodal triage data (vitals and chief complaint text) are processed and fused, then input to an ensemble of machine learning models to predict CT scan necessity at the point of care.

A. Data Acquisition and Description

VALID obtains data from the emergency department module of MIMIC-IV (MIMIC-IV-ED), a large-scale, publicly available database sourced from the electronic health record of the Beth Israel Deaconess Medical Center in Boston, Massachusetts [5]. MIMIC-IV was developed to support clinical research by providing structured, de-identified patient data collected during routine hospital care. Its combination of standardized tabular data and unstructured clinical notes makes it a widely used resource for developing and evaluating medical machine learning models [6].

B. Data Processing

Access to the MIMIC-IV database was obtained through the PhysioNet credentialing process, which requires completion of a data use agreement and human subjects research training. All data used in this study were fully de-identified in accordance with the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor provision [7]. Patient identifiers were replaced with randomized surrogates, and free-text fields were

processed using a hybrid de-identification algorithm to remove any protected health information (PHI).

To prepare the data for machine learning analysis, rigorous preprocessing and deduplication steps were performed. The tabular data were merged on the `subject_id` and `stay_id` primary keys. Missing physiological measurements in the triage data were handled using median imputation to ensure robustness, while categorical variables such as admission type and gender were one-hot encoded. Finally, continuous physiological variables were normalized using standard scaling to optimize the convergence of gradient-based classifiers.

C. Retrospective Ground-Truth Extraction via NLP

To address this systemic bottleneck, our project introduces VALID (Verified AI for Limiting Inefficient Diagnostics), an intelligent decision-support framework that prospectively predicts the clinical utility of a CT scan at the point of triage, empowering physicians to safely reduce low-value imaging without missing critical diagnoses. A fundamental challenge in training a model to predict CT scan necessity is the lack of explicit binary labels in raw medical records. To overcome this, VALID utilizes an automated Natural Language Processing (NLP) pipeline to retrospectively extract ground-truth labels from the final radiology reports.

Clinical notes often embed context, negation, and diagnostic nuance, which makes semantic extraction necessary to interpret them accurately. VALID extracts semantic meaning from physician narratives using ClinicalBERT, an application of the BERT model to clinical corpora [8]. BERT (Bidirectional Encoder Representations from Transformers) is a deep neural network that uses the transformer encoder architecture to learn embeddings for text . It is designed to learn deep bidirectional representations from unlabeled text by conditioning on both left and right context in all layers [9]. ClinicalBERT extends BERT by continuing its pre-training on large collections of clinical notes that allow it to better model medical jargon, abbreviations, and syntax [10].

Following the preprocessing steps of tokenization, cleanup, and normalization, the radiology reports are passed through ClinicalBERT's transformer encoder layers. A classification layer evaluates these embeddings to determine if the scan yielded an acute, actionable finding (labeled as 1, or "Necessary") or a normal/chronic finding (labeled as 0, or "Not Necessary"). To prioritize patient safety, the classification threshold for this pipeline was strictly calibrated to minimize false negatives, ensuring that critical diagnoses were not erroneously categorized as unnecessary.

D. Prospective Multimodal Classification

With the dataset accurately labeled by the retrospective NLP pipeline, VALID adopts an integrated feature engineering approach to construct a prospective predictive model. This model operates exclusively on information available the moment a patient is triaged in the Emergency Department.

To capture the semantic intent of the patient's presentation, the short, unstructured chiefcomplaint text from the triage notes is passed through ClinicalBERT. The final hidden-state vector of the [CLS] token is extracted to provide a sequence-level representation of the clinical narrative. To manage dimensionality, Principal Component Analysis (PCA) is applied to compress these embeddings from a 768 feature vector to a 20-vector for efficiency. This dense vector is then fused with the structured tabular patient data (triage vitals) to form a unified, multimodal feature vector for each patient.

These feature vectors are provided as input to three distinct machine learning classifiers to generate clinically meaningful CT scan necessity predictions: a Multi-Layer Perceptron (MLP), Random Forest, and XGBoost.

The MLP acts as our deep learning baseline, capturing complex, non-linear relationships within the multimodal data. During training, its parameters are optimized by minimizing the binary cross-entropy loss function:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \qquad (3)$$

where $y_i \in \{0,1\}$ represents the true label and p_i is the predicted probability of a necessary CT scan [11].

To complement the neural network, we deployed two tree-based ensemble methods: Random Forest, which establishes a robust, non-parametric decision boundary, and XGBoost, which maximizes predictive accuracy and interpretability by sequentially correcting errors and providing feature importance metrics.

By evaluating these three models simultaneously, VALID ensures robust generalization to accurately predict imaging utility without relying on post-scan documentation.

III. Results

Figure 2 compares the performance of several candidate classifiers for labeling radiology reports as CT Needed or CT Not Needed. Because the dataset contains a large proportion of reports without actionable findings, evaluation focuses on metrics that emphasize correct identification of the minority class (CT Needed), including recall, precision, F1 score, and PR-AUC.

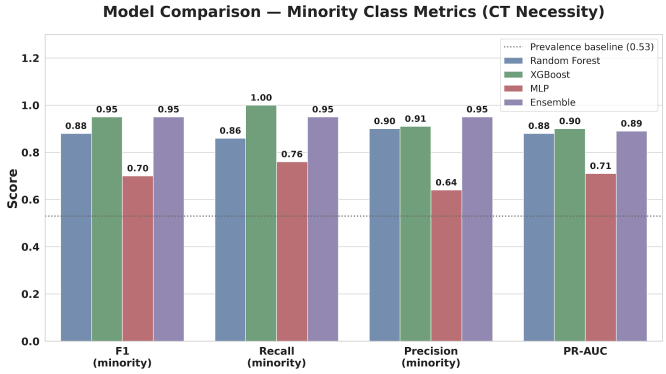

Fig. 2. Performance comparison of classifiers for labeling radiology reports as CT Needed or CT Not Needed.

Figure 3 shows the row-normalized confusion matrices for each classifier. Random Forest and XGBoost demonstrate strong performance in identifying necessary scans, while the MLP shows a higher rate of misclassification for this class. The ensemble model provides the most balanced performance, achieving high sensitivity for necessary scans while minimizing false negatives.

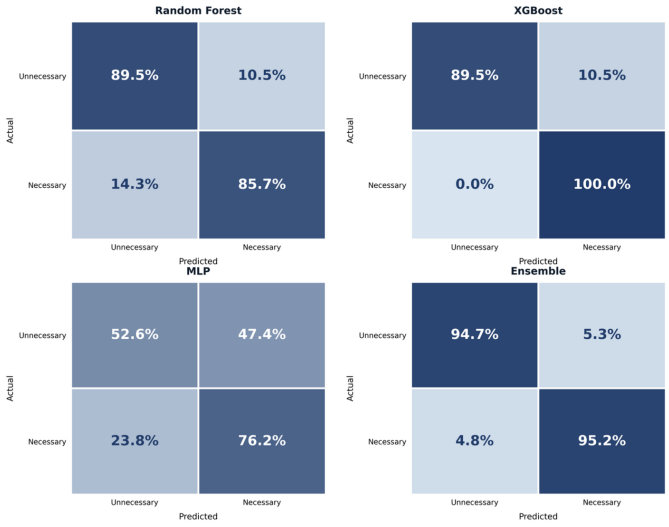

Fig. 3. Row-normalized confusion matrices for radiology report classification models (n = 40).

Figure 4 summarizes model performance for predicting the necessity of CT scans using a comprehensive set of metrics focused on the minority class (CT Needed). The prevalence baseline (dotted line) illustrates that the models significantly outperform naïve guessing.

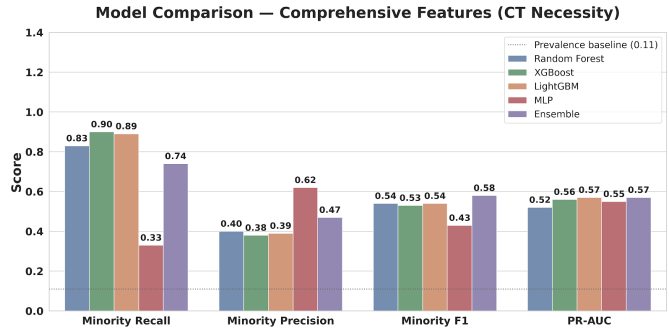

Fig. 4. Prospective prediction model performance for CT necessary

Because the dataset is dominated by CT not needed cases (113,376 out of 127,527 cases), weighted averages can mask poor performance on the minority class (CT needed). To accurately assess model effectiveness for identifying patients who truly require imaging, macro-average metrics are used alongside weighted averages. Tables 1 and 2 summarize the weighted-average and macro-average metrics for all models, emphasizing the importance of macro-average evaluation in assessing true performance on the minority class.

TABLE I
WEIGHTED AVERAGE PRECISION, RECALL, AND F1 FOR EACH PROSPECTIVE PREDICTION MODEL.

Model	Precision	Recall	F1
Random Forest	84%	84%	87%
XGBoost	83%	83%	85%
LightGBM	83%	83%	86%
MLP	90%	90%	89%
Ensemble	88%	88%	89%

TABLE II
ACCURACY AND MACRO AVERAGE PRECISION, RECALL, AND F1 FOR EACH PROSPECTIVE PREDICTION MODEL.

Model	Accuracy	Precision	Recall	F1	PR-AUC
Random Forest	84%	69%	84%	72%	0.524
XGBoost	83%	68%	86%	71%	0.559
LightGBM	83%	69%	86%	72%	0.567
MLP	90%	72%	82%	75%	0.552
Ensemble	88%	91%	88%	89%	0.568

While the MLP achieved the highest overall accuracy (90%), its poor macro-average metrics reveal a dangerously high false negative rate (67%), demonstrating that weighted averages can drastically overstate clinical performance. Similarly, tree-based models (Random Forest, XGBoost, LightGBM) exhibited high weighted precision (91–92%) but lower macro-average recall (69–71%), reflecting a trade-off between avoiding false positives and missing necessary scans.

The ensemble model proved to be the most clinically appropriate option, significantly improving minority-class detection to achieve a macro-average recall of 88% and precision of 91%. Finally, Figure 5 presents row-normalized confusion matrices that reinforce the need to minimize false negatives in practice, highlighting XGBoost's low false negative rate (10.4%) in stark contrast to the MLP.

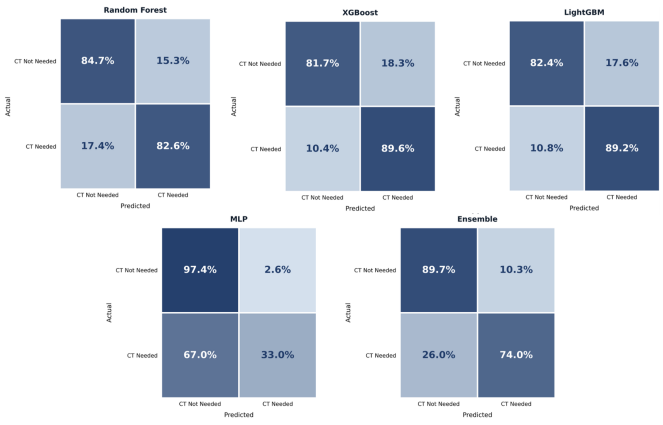

Fig. 5. Row-normalized confusion matrices for prospective CT necessity prediction (n=127,527)

Figure 6 illustrates a prototype clinical dashboard designed for emergency department physicians. Supported by a deployable API to facilitate seamless integration with existing electronic health record (EHR) systems, the interface processes

unstructured triage notes and physiological vitals to prospectively assess the clinical necessity of a CT scan.

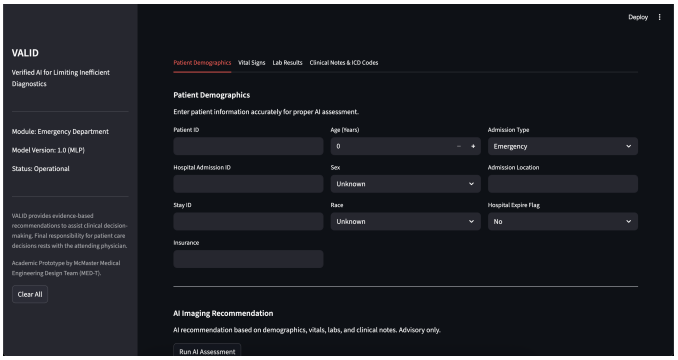

Fig. 6. Prototype clinical dashboard utilizing the VALID multimodal ensemble model to assess CT scan necessity at the point of triage.

IV. Conclusion

This project developed and evaluated multiple machine learning models to predict the necessity of CT scans at emergency department triage using multimodal clinical data. The results demonstrate that while all models perform well in identifying low-value CT scans, detecting high-value, necessary scans remains more challenging due to class imbalances and clinical complexity.

The low prevalence of necessary CT cases, relative to unnecessary scans, makes it difficult for models to achieve high recall without increasing false positives. Weighted-average metrics can obscure poor performance on these critical minority cases, highlighting the importance of evaluating macro-average precision, recall, and F1 to assess true effectiveness. Among the models tested, the ensemble approach provided the most clinically appropriate trade-off, achieving improved recall for necessary CT scans while maintaining high precision, suggesting it may be most suitable for a decision-support system aimed at minimizing false negatives and ensuring patient safety.

Despite these promising results, several challenges and limitations remain. First, the input available at triage is limited, restricting the amount of predictive information available to the models. Also, the ground-truth labels were retrospectively extracted from radiology reports using NLP, which may introduce errors or bias inherent in clinical documentation. Finally, the strong class imbalance remains a persistent challenge, as the minority CT Needed cases are both clinically important and difficult to detect reliably.

Future work should focus on improving minority-class detection and enhancing real-world applicability. Potential directions include optimizing decision thresholds and model calibration to further reduce false negatives without generating excessive false positives, and integrating additional patient information such as comorbidity history. Prospective validation in live clinical settings is also necessary to assess safety, usability, and workflow integration. Additionally, enhancing model explainability will be needed for clinician adoption, allowing healthcare providers to understand and trust model recommendations. Overall, this framework demonstrates the potential for machine learning to reduce unnecessary imaging, improve efficiency, and support patient safety.

References

[1] K. Grant. (2017) Up to 30 per cent of medical care canadians receive is unnecessary: Report. Department of Medicine, University of Toronto. Republished from the *Globe and Mail* and based on a report by the Canadian Institute for Health Information and Choosing Wisely Canada. [Online]. Available: https://deptmedicine.utoronto.ca/news/30-cent-medical-care-canadians-receive-unnecessary-report

[2] J. J. You, I. Purdy, D. M. Rothwell, R. Przybysz, J. Fang, and A. Laupacis, "Indications for and results of outpatient computed tomography and magnetic resonance imaging in ontario," *Canadian Association of Radiologists Journal*, vol. 59, no. 3, pp. 135–143, jun 2008. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/18697720/

[3] D. Chojecki, E. Wright, E. Kirwin, N. Razavilar, M. Karkhaneh, C. Yan, J. Round, B. Guo, and C. Moga, "Optimizing the use of low back pain and spine condition-related interventions and procedures," 2022. [Online]. Available: https://ihe.ca/files/optimizing_the_use_of_low_back_pain_and_spine_condition_related_interventions_and_procedures.pdf

[4] P. Causa Andrieu, J. S. Golia Pernicka, R. Yaeger, K. Lupton, K. Batch, F. Zulkernine, A. L. Simpson, M. Taya, L. Gazit, H. Nguyen, K. Nicholas, N. Gangai, V. Sevilimedu, S. Dickinson, V. Paroder, D. D. B. Bates, and R. Do, "Natural language processing of computed tomography reports to label metastatic phenotypes with prognostic significance in patients with colorectal cancer," *JCO Clinical Cancer Informatics*, vol. 6, p. e2200014, 2022. [Online]. Available: https://ascopubs.org/doi/10.1200/CCI.22.00014

[5] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, B. Moody, B. J. Gow, L.-W. H. Lehman, L. A. Celi, and R. G. Mark, "Mimic-iv, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/36596836/

[6] A. Bennett, J. Wiedekopf, H. Ulrich, P. van Damme, P. Szul, J. Grimes, and A. E. W. Johnson, "Mimic-iv on fhir (version 2.1)," PhysioNet, 2024, version 2.1, RRID:SCR_007345. [Online]. Available: https://physionet.org/content/mimic-iv-fhir/2.1/

[7] A. Johnson, L. Bulgarelli, T. Pollard, L. A. Celi, R. Mark, and S. Horng, "Mimic-iv-ed (version 2.2)," *PhysioNet*, jan 2023, emergency Department dataset. [Online]. Available: https://physionet.org/content/mimic-iv-ed/2.2/

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint*, vol. abs/1810.04805, 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[10] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint*, vol. abs/1904.05342, 2019. [Online]. Available: https://arxiv.org/abs/1904.05342

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.