CLASSICALLIBRARY : Pages:16–29

Original Research

Multi-Modal Clinical Document Understanding via Joint Text–Image Representations

Pratiksha Adhikari[1]

[1]Kathmandu Engineering College, Department of Information Technology, Kalimati Road, Kathmandu, Nepal.

Abstract
Multi-modal clinical document understanding has emerged as a critical area of investigation, aiming to improve patient outcomes, aid clinical decision-making, and streamline healthcare workflows by leveraging multiple sources of information. These sources include textual reports, physician notes, and diagnostic images such as X-ray, CT, and MRI scans. Traditional approaches for interpreting clinical data have predominantly focused on either text or images independently, missing valuable insights that can emerge from the synergy of textual and visual features. Recent advances in deep learning now enable the integration of diverse data streams, providing a more holistic view of patient conditions and reducing diagnostic uncertainty. However, effective multi-modal representation still poses several challenges, such as aligning high-dimensional data from heterogeneous domains, handling sparse and noisy clinical notes, and integrating large-scale datasets without overfitting. This work explores the theoretical foundations, methodological designs, and practical implementations of multi-modal systems for clinical document understanding, with a particular emphasis on joint text–image representations. By blending state-of-the-art natural language processing techniques with robust image feature extraction modules, we examine how models can capture latent relationships across modalities and how structured representations can be employed for domain-specific reasoning tasks. Our approach aspires to push the boundaries of current capabilities, ultimately enabling comprehensive and context-aware analyses of complex clinical datasets for improved patient care.

1. Introduction

Multi-modal clinical document understanding integrates the analysis of textual narratives, including clinical reports, discharge summaries, and progress notes, with the rich visual details contained in medical images such as radiographs or histopathology slides [1]. The overarching aim of this field is to consolidate multiple streams of information so as to enhance diagnostic accuracy, facilitate clinical decision-making, and foster personalized patient management strategies. Despite the remarkable progress made in both natural language processing and computer vision, multi-modal fusion remains a technically and conceptually challenging endeavor [2]. The principal challenge arises from the distinct modalities involved and the necessity to map high-level textual concepts to visual features that may be distributed over multiple spatial dimensions.

The concept of multi-modal embeddings has garnered substantial interest in medical image interpretation [3]. Let us denote an input text corpus as $\{t_1, t_2, \ldots, t_n\}$ and a corresponding set of medical images as $\{x_1, x_2, \ldots, x_m\}$. The goal of a joint embedding model is to learn a function

$$F : \big(\{t_1, t_2, \ldots, t_n\}, \{x_1, x_2, \ldots, x_m\}\big) \to \mathbf{R}^d,$$

where d is the dimension of the latent space [4]. The elements of the resulting feature vector are intended to capture shared semantics and domain-specific cues that can characterize a patient's diagnostic profile. For instance, let z be the embedded representation of text elements and v be the embedded representation of visual elements [5]. A well-designed model will map clinically related text–image pairs closer together in the latent space, reflecting their semantic affinity.

One of the core motivations behind multi-modal integration is the ability to disambiguate concepts that may be underdetermined when examined separately. For example, the phrase "ground-glass opacities" in a radiology report might necessitate direct examination of the corresponding region in a CT scan to fully confirm the presence and extent of the abnormality [6]. This synergy underscores the importance of designing joint representation models that can systematically align textual descriptors (e.g., "masses," "lesions," "consolidations") with corresponding visual biomarkers. However, effectively capturing such relationships demands careful curation of datasets and sophisticated modeling techniques capable of capturing relevant invariances. [7]

A fundamental theoretical question involves the precise nature of multi-modal alignment. In logic terms, suppose we express a statement $P(x)$ denoting that a certain pathology is visible in image x [8]. We also have a statement $Q(t)$ denoting that a text fragment t references the same pathology. In a coherent alignment, we want to ensure that whenever $P(x)$ is true for a particular pathology, and $Q(t)$ is true for the corresponding textual description, these statements should be recognized as describing the same underlying phenomenon [9]. Symbolically, we might say:

$$\forall x\, \forall t\, \big(P(x) \wedge Q(t) \implies A(x,t)\big), \text{[10]}$$

where $A(x, t)$ represents an alignment predicate in the joint embedding space. Capturing this alignment becomes a matter of defining appropriate loss functions and data sampling strategies such that co-occurring text–image pairs in the training set are consistently matched.

From a linear algebraic perspective, consider a textual embedding space spanned by a matrix $W_t \in \mathbb{R}^{k\times d}$ and an image embedding space spanned by a matrix $W_i \in \mathbb{R}^{l\times d}$, where k and l correspond to input dimension sizes for text and images, respectively, and d is the dimension of the shared latent space. A multi-modal alignment model may define transformations: [11]

$$z = f_t(t) = \sigma(W_t \cdot t^\top + b_t), \quad v = f_i(x) = \sigma(W_i \cdot x^\top + b_i),$$

where σ denotes a non-linear activation function such as ReLU or the hyperbolic tangent, and b_t, b_i are bias terms [12]. The fundamental challenge is to ensure that, for corresponding text–image samples that describe the same patient context, the distance $\|z - v\|$ is minimized, while non-matching pairs have a larger separation. This objective might be realized using a margin-based loss, cross-entropy loss, or a contrastive learning framework. [13]

Within the landscape of clinical NLP, the textual data is often rife with terminological variability, abbreviations, and domain-specific jargon. Concurrently, medical images can exhibit subtle visual features that require domain expertise to interpret [14]. Such idiosyncrasies highlight the need for domain adaptation and robust feature extraction modules that can handle linguistic irregularities and image artifacts. In many instances, a pre-trained language model (for example, a model specialized on biomedical corpora) is integrated with a convolutional neural network or a vision transformer fine-tuned on medical image sets. The synergy of these complementary networks must then be carefully calibrated to produce semantically coherent embeddings. [15]

Historically, research on multi-modal fusion in clinical contexts has explored the concatenation of learned text and image embeddings, the design of cross-modal attention mechanisms, and the use of graph-based methods that treat text and image features as interconnected nodes. Each approach carries its own merits and limitations [16]. Concatenation-based methods may lack fine-grained alignment capabilities, while attention-based models demand large amounts of data and computational resources. Graph-based approaches show promise in capturing relational patterns, but they require a clear definition of node and edge semantics [17, 18].

In practice, the representation of textual data often relies on token-level embeddings or subword embeddings that can capture morphological and semantic relationships among medical terms. These embeddings are then fed into transformer-based architectures, which have become de facto standards in natural language processing [19]. For images, convolutional neural networks or vision transformer backbones extract feature maps that can be pooled or flattened into an embedding vector. The challenge then becomes to define a joint function $F(z, v)$ that fuses or aligns the respective features [20]. Considering an attention-based mechanism, one might define:

$$\alpha_j = \frac{\exp\big(\beta(z, v_j)\big)}{\sum_{i=1}^{m} \exp\big(\beta(z, v_i)\big)},$$

where $v_1, \ldots, v_m$ are spatial image features, and β is a learnable compatibility function. In this manner, the text embedding z selectively attends to relevant image regions [21]. This notion can be inverted to allow images to attend to relevant textual tokens, facilitating cross-modal interplay.

Despite these promising directions, multi-modal integration in clinical documents is still in its relative infancy compared to more general multi-modal tasks such as image captioning or visual question answering in open-domain settings [22]. There are both practical constraints, such as data privacy regulations that limit dataset sharing, and technical constraints, such as the difficulty of collecting large, high-quality text–image pairs that accurately represent clinical workflows. Consequently, domain adaptation, transfer learning, and careful model regularization remain integral to achieving robust performance. [23]

The forthcoming sections delve into the details of building joint text–image representations for clinical document understanding. We investigate how to structure the data, which neural architectures are best suited to this domain, and how advanced techniques in representation learning can be adapted for the nuanced demands of medical diagnostics [24]. By anchoring our discussion in theoretical underpinnings, practical heuristics, and empirical results, we aim to clarify the current state of the field and point toward future developments that promise to streamline integrative analysis of multi-modal clinical data.

2. Data Foundations and Representation

The design of effective multi-modal clinical document understanding systems relies on the complex interplay between data preparation, annotation strategies, and representation learning [25]. Clinicians often write lengthy narratives containing fragmented references to anatomical structures, pathologies, and procedures. Meanwhile, images come in diverse modalities, each with its own spatial resolution and contrast characteristics. Hence, the foundation of any rigorous model-building process involves carefully curated and annotated training datasets that capture the inherent variability of both text and images. [26]

One must consider the presence of domain-specific terms, abbreviations, and acronyms unique to clinical practice. Let us denote a corpus of clinical text by $T = \{t_i\}_{i=1}^{N}$ and an associated corpus of images by $X = \{x_j\}_{j=1}^{M}$. While in an ideal setting each t_i would be directly paired with an x_j describing the same clinical event or patient condition, the reality is often far more fragmented [27]. A single text document may reference multiple images, or several text documents may refer to the same image. Therefore, we must define a robust mapping strategy $\Phi : T \to X$, which indicates which text documents align with which images [28]. This mapping can be partial, injective, or surjective, depending on how clinical data is collected.

An essential step in annotation involves standardizing the textual data [29]. Clinical text often contains synonyms, e.g., “myocardial infarction” and “heart attack,” that should map to the same concept. Let us denote a standardizing function $\omega(t)$ that normalizes text input t to a canonical form via dictionary lookup or ontological mappings. Formally, if we let Ω be an ontology capturing medical concepts, we can define: [30]

$$\omega : t \mapsto c, \quad c \in \Omega,$$

thereby connecting raw text segments to well-defined domain concepts [31]. On the imaging side, each medical image requires an identification of regions of interest and relevant metadata such as imaging modality (CT, MRI, ultrasound) or anatomical site. This metadata can be denoted as $\gamma(x)$, which may include bounding boxes, segmentation masks, or morphological descriptors. [32]

When constructing embeddings that encapsulate text and image features, one approach is to build domain-specific dictionaries or vocabularies that concentrate on diseases, anatomical structures, and procedures. Another approach is to rely on unsupervised or self-supervised pre-training of large neural networks on a broad corpus of medical text and images, followed by fine-tuning on a smaller annotated dataset [33]. The advantage of pre-training emerges from the possibility of discovering low-level patterns in large volumes of unlabeled data. For instance, a large language model might learn robust representations of medical terminology, while a convolutional neural network might identify fundamental image primitives like edges, corners, and texture patterns [34]. For multi-modal tasks, one can combine these strategies by introducing contrastive or paired losses that align text tokens with image patches.

Multi-modal alignment can also be framed through the lens of manifold learning. Suppose $\mathcal{M}_t$ is the manifold underlying textual data, and $\mathcal{M}_i$ is the manifold underlying image data. The objective is to find a common manifold $\mathcal{M}$ such that there exist functions $f : \mathcal{M}_t \to \mathcal{M}$ and $g : \mathcal{M}_i \to \mathcal{M}$ for which the embeddings of corresponding text–image pairs are neighbors. More formally, if (t_k, x_k) is a matched pair, we want $\|f(t_k) - g(x_k)\| \le \epsilon$ for some small ϵ [35]. At the same time, for mismatched pairs, we want the embeddings to lie farther apart on the manifold. One might use topological constraints, such as requiring that each manifold be locally isometric to $\mathcal{M}$, though such constraints can be challenging to optimize in practice.

Structured representation techniques gain particular prominence in the context of clinical documents, as they enable the explicit modeling of relationships between medical concepts [36]. For instance, a **knowledge graph** may consist of nodes representing entities such as "patient," "diagnosis," "treatment," and "symptom," with edges encoding relations like "has_diagnosis" or "receives_treatment."

When extended to **multimodal data**, such as medical images, the graph can be augmented by associating each image—or specific image regions—with relevant clinical entities [37]. Symbolically, a knowledge graph can be expressed as a set of logical assertions [38]:

$$\{R(u, v)\},$$

where R denotes a binary relation and u, v are entities or concepts. [39]

In a **joint embedding framework**, these structured relationships impose soft or hard constraints on the alignment between textual embeddings and visual features. This results in representations that are not only data-driven but also informed by domain knowledge, encouraging semantic consistency across modalities [40]. The integration of such expert-defined relational priors improves the interpretability and robustness of the learned embeddings, particularly in settings where data is sparse, heterogeneous, or institutionally fragmented.

The data foundations stage concludes with thorough quality checks and an iterative refinement of both textual normalization strategies and image annotations. Missing or incomplete labels can degrade the quality of multi-modal models, as alignment objectives strongly depend on accurate matching [41]. In real-world clinical settings, partial data is common, and strategies such as weak supervision, semi-supervised learning, or data augmentation (e.g., text paraphrasing, image transformations) can mitigate these limitations. Ultimately, the careful development of these foundational steps paves the way for constructing more sophisticated architectures that can effectively interpret and reason about multi-modal clinical data. [42]

3. Architecture for Multi-Modal Fusion

Building on the robust data foundations, the next pivotal element in multi-modal clinical document understanding is the design of neural network architectures capable of fusing textual and visual information. These architectures can be conceptualized as pipelines that first transform raw text and raw images into lower-dimensional feature embeddings, and then integrate or align these embeddings through a fusion layer. [43]

We can denote the textual encoder by E_t and the image encoder by E_i. The textual encoder might be a pre-trained transformer specialized on medical text, or a recurrent neural network with specialized token embeddings [44]. For instance, let $t = (w_1, w_2, \ldots, w_n)$ be the sequence of tokens in a clinical note, and let:

$$z = E_t(t) \in \mathbb{R}^d,$$

where z is the aggregate text embedding. Similarly, let x be a clinical image, and: [45]

$$v = E_i(x) \in \mathbb{R}^d,$$

be the resulting image embedding. The design choices for E_t and E_i may range from pure convolutional backbones to attention-based image encoders (e.g., vision transformers) in the imaging pathway, and from smaller LSTM-based approaches to large language models for the textual pathway. [46]

A core innovation in multi-modal architectures lies in cross-attention mechanisms. These mechanisms aim to allow textual features to attend to relevant visual regions and, conversely, allow visual features to attend to salient textual tokens [47]. Formally, consider the sets of features $Z = \{z_1, z_2, \ldots, z_n\}$ and $V = \{v_1, v_2, \ldots, v_m\}$. A cross-attention module defines query, key, and value transformations for both text and image features. Let:

$$Q_z = W_q^z Z, \quad K_v = W_k^v V, \quad V_v = W_v^v V, \text{[48]}$$

where W_q^z, W_k^v, W_v^v are learnable parameter matrices. The attention from text to image features is computed as: [49]

$$\text{Attention}(Z, V) = \text{softmax}\Big(\frac{Q_z K_v^\top}{\sqrt{d}}\Big) V_v.$$

A parallel process can compute image-to-text attention. Through iterative stacking of such cross-attention layers, the model refines the representation of text by integrating visually grounded features, and refines the representation of image data by leveraging textual context.

Another promising avenue for multi-modal fusion is graph-based integration [50]. Suppose we represent each sentence or phrase in the clinical note as a node in one sub-graph, and each region of the image as a node in another sub-graph. We can then define edges that link text nodes to image nodes if they co-occur or if an attention mechanism deems them related [51]. Symbolically, let $G = (U, E)$ be a heterogeneous graph where $U = U_t \cup U_i$ is a union of text nodes and image nodes, and E is a set of edges. A graph neural network (GNN) can then propagate information across edges, resulting in contextually enriched node embeddings: [52]

$$h_u^{(l+1)} = \phi\Big(h_u^{(l)}, \{h_v^{(l)} : (v, u) \in E\}\Big),$$

where ϕ is a message-passing function. This approach has the capacity to represent explicit relationships such as "image region r correlates with mention m in text," leading to structured alignments. [53]

Fusion can also be performed by direct concatenation or pooling of the text and image embeddings, though such naive methods risk discarding the fine-grained interactions that might be crucial for diagnosis. A more refined method might involve a set of linear transformations: [54]

$$f(z, v) = \psi\big(\alpha \cdot z + (1 - \alpha) \cdot v\big),$$

where α is a learnable scalar or a gating function that adapts to the context. For instance, if a certain diagnosis is strongly cued by textual semantics, α might shift focus toward the text encoder's output, while if the diagnosis is visually distinctive, α might favor the image encoder's output [55]. Non-linear transformations, such as those realized through multi-layer perceptrons, can then project the fused embedding into a label space for classification tasks (e.g., diagnostic labels) or into a generative model framework for tasks like image captioning or report generation.

Mathematically, consider a multi-task objective function that includes both supervised classification loss and contrastive alignment loss [56]. Let $L_{\text{class}}(\theta)$ be the classification loss for diagnosing a condition based on the fused embedding, and let $L_{\text{align}}(\theta)$ be a contrastive loss that enforces alignment between matched text–image pairs. We can write:

$$L(\theta) = \lambda_{\text{class}} L_{\text{class}}(\theta) + \lambda_{\text{align}} L_{\text{align}}(\theta),$$

where λ_{class} and λ_{align} are hyperparameters that control the trade-off between classification accuracy and embedding alignment. This composite objective encourages the fused model to be both diagnostically accurate and semantically aligned [57]. One might also include auxiliary losses for tasks such as textual entailment or visual question answering, further enriching the representation.

In large-scale clinical settings, training such models often entails substantial computational overhead, especially if cross-attention modules or GNN-based approaches are utilized [58]. Consequently, distributed training paradigms and data-parallel strategies are typically employed. Moreover, because clinical datasets are frequently subject to data-sharing restrictions, federated learning approaches have been explored. In a federated scenario, each institution may maintain a local multi-modal model update without transferring raw data, only sharing parameter gradients [59]. This approach mitigates privacy concerns but increases the complexity of ensuring consistent alignment across geographically dispersed data sources.

Ultimately, the architectural choices for multi-modal fusion must balance complexity, data availability, computational constraints, and the specific nature of the clinical question at hand [60]. With robust, well-structured architectures in place, it becomes feasible to build upon them to tackle increasingly complex tasks, including automated radiology report generation, lesion detection guided by textual descriptions, and knowledge graph completion where the synergy of text and images can unveil novel insights into patient conditions.

4. Evaluation and Metrics

An integral part of any multi-modal clinical document understanding system is a rigorous evaluation framework that can objectively measure its effectiveness [61]. Unlike simpler classification tasks, multi-modal systems often demand a multi-pronged approach to evaluation that captures performance across textual understanding, image interpretation, and the synergy of both.

A fundamental evaluation step is to measure alignment quality between text and images [62]. Suppose we have a test set $\{(t_k, x_k)\}_{k=1}^{N}$ of matched text–image pairs. A common alignment metric is the retrieval-based approach. One computes the embedding $z_k = E_t(t_k)$ for each text and $v_k = E_i(x_k)$ for each image [63]. Then, a retrieval score is derived by selecting the top-ranked image for each text (or vice versa) based on cosine similarity. Metrics such as recall@K and mean rank measure how effectively the model retrieves the correct counterpart. A high recall@K indicates that matched pairs are embedded closely, reflecting strong alignment. [64]

For diagnostic tasks, classification performance provides additional insights. One might define a set of clinical labels C, such as specific pathologies or findings [65]. Given a fused embedding $f(z_k, v_k)$, the system outputs a label prediction $\hat{c}$ for the pair (t_k, x_k). Comparing $\hat{c}$ to the ground truth label c_k yields classification metrics such as accuracy, precision, recall, and F1-score. In more nuanced cases, one may adopt hierarchical metrics that reflect the severity or specificity of a diagnosis. For example, conflating "hypertensive heart disease" with "chronic heart failure" may be a lesser error than conflating "hypertensive heart disease" with "breast cancer." [66]

Beyond classification and retrieval, generative tasks, such as report generation, require specialized metrics. For a system that takes an image and partial textual input to produce a full radiology report, standard natural language generation metrics such as BLEU, ROUGE, or METEOR can be employed [67]. However, these metrics do not fully capture clinical correctness. Hence, a clinically oriented evaluation might require domain experts to rate the generated reports or apply specialized measures that check for key findings, correctness of stated pathologies, and coherence of the generated text [68]. From a more formal perspective, one might define a set of logical statements S_k that the generated report should satisfy, such as "mentions pathology X if present in the image." A logic-based metric can compute the fraction of these statements that hold true.

An emerging field of interest is explainability and interpretability of multi-modal models. Evaluating explainability involves determining whether the system can highlight relevant image regions when describing a finding, or whether it can reference the specific text spans that led to a particular conclusion [69]. One can use saliency-based measures or attention-weight visualization to ascertain how well the learned embeddings correlate with clinically meaningful features. Although this evaluation often remains qualitative, efforts to quantify interpretability can involve overlap measures with bounding boxes or segmentation masks [70]. Symbolically, for each text token w_i that references a pathology, one might evaluate a function $\eta(w_i, x)$ that indicates whether the model's attention mechanism focuses on the corresponding image region. Summarizing such overlaps across a dataset yields an average measure of interpretability. [71]

Handling uncertainty is also crucial, given that multi-modal models in clinical domains often output probabilistic estimates of pathology presence. Calibration metrics, such as the expected calibration error (ECE), can assess whether the model's predicted probabilities match empirical frequencies [72]. Suppose the model outputs a probability p_k of a pathology for a given text–image pair (t_k, x_k). A well-calibrated model ensures that for all pairs predicted with probability p, the actual fraction of positives is close to p. If we discretize the probability space into bins $B_1, \ldots, B_K$, each bin containing predictions around a certain probability value, the ECE is computed as: [73]

$$\mathrm{ECE} = \sum_{j=1}^{K} \frac{|B_j|}{N} \left| \bar{p}_{B_j} - \bar{y}_{B_j} \right|,$$

where $\bar{p}_{B_j}$ is the mean predicted probability in bin B_j, and $\bar{y}_{B_j}$ is the mean actual outcome. Low ECE indicates good calibration, an important property for models used in critical clinical decisions.

Finally, real-world validation often involves prospective studies or retrospective analyses with carefully selected patient cohorts [74]. These evaluations might measure clinical endpoints such as diagnostic time, misdiagnosis rates, or treatment outcomes. Although these end-to-end evaluations are more challenging to conduct and control, they provide the definitive measure of a system's utility in practical settings [75]. As multi-modal models become more pervasive in healthcare, regulators and institutions may mandate standardized testing protocols to ensure patient safety and consistent performance across diverse clinical environments.

5. Challenges and Future Directions

Despite substantial advancements in multi-modal clinical document understanding, numerous challenges remain that hinder widespread adoption of these techniques in everyday healthcare settings [76]. Perhaps the most pressing among these is the necessity for large, representative datasets that comprehensively capture the variability of clinical practice. In many clinical domains, images may be scarce, or the textual data might be incomplete, noisy, and filled with jargon [77]. Privacy regulations such as HIPAA in the United States or GDPR in the European Union further constrain data sharing, thereby limiting opportunities to train large-scale models across multiple institutions.

Another challenge lies in ensuring the reliability and interpretability of these multi-modal systems. While modern neural networks can achieve impressive accuracy, they often behave as black boxes, offering limited insight into how final predictions are reached [78]. This opacity becomes especially problematic in high-stakes clinical decision-making, where clinicians need to trust and understand model outputs. Efforts to incorporate attention maps, saliency methods, and localized explanations into multi-modal architectures are promising, but they remain insufficient to provide the rigorous interpretability demanded by medical practitioners [79]. One possible route to improved interpretability is the fusion of symbolic reasoning with distributed representations. Symbolic reasoning can enforce logical consistency and domain constraints, while neural embeddings capture more nuanced associations [80]. Let us define a consistency constraint κ as a set of Horn clauses or descriptive rules that clinical decisions must obey. Symbolically, κ might include statements like: [81]

$$\forall p \left(\text{Diagnosis}(p, \text{pneumonia}) \implies \text{Symptom}(p, \text{fever}) \vee \text{Symptom}(p, \text{cough})\right).$$

Such rules can be integrated into the learning process via penalty terms in the objective function, thereby guiding the model toward clinically coherent predictions.

Federated learning and distributed training approaches can mitigate some data-related constraints, but they require sophisticated orchestration and trust between institutions to ensure the correctness of updates and to prevent privacy leaks [82, 83]. Even with federated learning, the local data at each institution might be heterogeneous, with varying imaging protocols, different user interfaces for clinical text entry, and varying levels of annotation quality. This heterogeneity can degrade model performance unless domain adaptation or robust aggregation methods are implemented.

Another major frontier is the representation of temporal information [84]. Patient data evolves over time, with multiple imaging studies and textual entries recorded at different visits. Temporal modeling can significantly improve diagnostic accuracy, especially for chronic conditions or progressive diseases [85]. Formally, let us denote the text data at time steps $t_1, t_2, \ldots, t_n$ by $T_{t_1}, T_{t_2}, \ldots, T_{t_n}$ and the corresponding image sets by $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$. A multi-modal time-series approach must fuse information not only across modalities but also across time:

$$h_{t_k} = \psi\left(h_{t_{k-1}}, E_t(T_{t_k}), E_i(X_{t_k})\right),$$

where h_{t_k} is a hidden state summarizing the patient's condition up to time t_k. Graph-based or transformer-based models that incorporate temporal edges or positional encodings can track disease progression and improve prognostic predictions. [86]

The scarcity of well-annotated data also prompts new research on weakly supervised or self-supervised approaches. In a weakly supervised setting, a text document may contain a mention of a pathology without precise localization in the image [87]. Self-supervised learning strategies like masked language modeling or masked image modeling can leverage large unlabeled corpora, bridging data gaps. By defining suitable pretext tasks, such as predicting missing tokens in text or reconstructing partially occluded image regions, models can learn robust representations that later serve as foundations for downstream multi-modal tasks. Mathematically, let $\tilde{t}$ be a text sequence with randomly masked tokens, and $\tilde{x}$ be an image with masked regions. A reconstruction loss can be defined as: [88]

$$L_{\text{SSL}}(\theta) = \mathbb{E}\left[D\left(E_t(\tilde{t}), E_i(\tilde{x})\right)\right],$$

where D measures reconstruction error. The synergy of textual and visual embeddings in this self-supervised setting can lead to better alignment once actual paired data is introduced. [89]

Finally, real-time clinical applications demand efficient inference. A model that takes seconds per inference may be acceptable in certain settings like radiology, but in emergency care, near-instant predictions might be necessary [90]. Model compression, distillation, and quantization techniques can reduce inference time while minimally impacting performance. Let us define a teacher–student model configuration $\left(E^{(\text{teacher})}, E^{(\text{student})}\right)$, where the teacher is a large multi-modal model and the student is a compact version. A distillation loss can be introduced: [91]

$$L_{\text{distill}} = \mathrm{KL}\Big(\sigma(E^{(\text{teacher})}(t,x)), \sigma(E^{(\text{student})}(t,x))\Big),$$

where KL is the Kullback–Leibler divergence, and σ is a softmax or other transformation. This approach enables the smaller student model to inherit the teacher's knowledge, achieving near-teacher performance at reduced computational cost.

In sum, future directions in multi-modal clinical document understanding revolve around surmounting data limitations, ensuring interpretability and reliability, incorporating temporal dynamics, and developing real-time or near-real-time solutions [92]. The synergy of advanced neural architectures, domain-specific knowledge representations, and robust evaluation protocols stands to transform patient care by providing clinicians with integrative, context-rich insights that extend beyond the scope of unimodal analysis.

6. Conclusion

This work has explored the diverse theoretical and practical dimensions of multi-modal clinical document understanding through joint text–image representations. By integrating natural language processing techniques with sophisticated computer vision models, we can consolidate large volumes of heterogeneous information into unified embeddings that hold significant potential for improving clinical workflows, diagnostic accuracy, and patient outcomes [93]. The underlying motivation rests on the premise that medical text and images, taken together, can provide a more comprehensive and contextually rich depiction of the patient's condition, surpassing the limitations of single-modality analysis.

Our discussion emphasized data foundations, from ontology-based normalization of textual terms to structured annotations of medical images [94]. We presented advanced architectures that include cross-attention mechanisms, graph neural networks, and joint embedding methods capable of capturing the interplay between textual mentions and visual cues. These architectures, coupled with carefully designed training objectives involving classification, retrieval, and contrastive alignment, underscore the multifaceted nature of the problem [95]. Rigorous evaluation metrics, spanning alignment performance and clinically oriented diagnostics, are indispensable for assessing model utility and trustworthiness.

Nevertheless, substantial challenges remain [96]. Data scarcity, privacy restrictions, the necessity for interpretability, and difficulties in modeling temporal trajectories of patient data are hurdles that require continued innovation. Future directions point toward federated learning, advanced self-supervised strategies, and deeper integration of symbolic domain knowledge to yield systems that can navigate complex clinical scenarios with explainable reasoning. As computational power expands and collaborative initiatives grow, these multi-modal solutions have the potential to become integral tools in clinical decision support, shifting the paradigm from piecemeal analysis to cohesive, data-driven insights in patient care. [97]

References

[1] Y. Chillakuru, S. Munjal, B. Laguna, T. L. Chen, G. R. Chaudhari, T. Vu, Y. Seo, J. Narvid, and J. H. Sohn, "Development and web deployment of an automated neuroradiology mri protocoling tool with natural language processing.," *BMC medical informatics and decision making*, vol. 21, pp. 213–213, 7 2021.

[2] Z. Zhou, W. K. Chan, and J. H. Chow, "Agent-based simulation of electricity markets: a survey of tools," *Artificial Intelligence Review*, vol. 28, no. 4, pp. 305–342, 2007.

[3] M. Abouelyazid and C. Xiang, "Machine learning-assisted approach for fetal health status prediction using cardiotocogram data," *International Journal of Applied Health Care Analytics*, vol. 6, no. 4, pp. 1–22, 2021.

[4] R. Avula, "Architectural frameworks for big data analytics in patient-centric healthcare systems: Opportunities, challenges, and limitations," *Emerging Trends in Machine Intelligence and Big Data*, vol. 10, no. 3, pp. 13–27, 2018.

[5] F. J. Manion, M. R. Harris, A. G. Buyuktur, P. M. Clark, L. C. An, and D. A. Hanauer, “Leveraging ehr data for outcomes and comparative effectiveness research in oncology,” *Current oncology reports*, vol. 14, pp. 494–501, 9 2012.

[6] S. Uppuluri, A. Uppuluri, P. D. Langer, M. A. Zarbin, and N. Bhagat, “Enucleation in pediatric open globe injuries: demographics and risk factors.,” *Graefe’s archive for clinical and experimental ophthalmology = Albrecht von Graefes Archiv fur klinische und experimentelle Ophthalmologie*, vol. 260, pp. 3115–3122, 3 2022.

[7] H. Guo, Özgür Kafalı, A.-L. Jeukeng, L. Williams, and M. P. Singh, “Çorba: crowdsourcing to obtain requirements from regulations and breaches,” *Empirical Software Engineering*, vol. 25, pp. 532–561, 8 2019.

[8] B. Balducci and D. Marinova, “Unstructured data in marketing,” *Journal of the Academy of Marketing Science*, vol. 46, pp. 557–590, 6 2018.

[9] E. Holderness, N. Miller, P. B. Cawkwell, K. Bolton, M. Meteer, J. Pustejovsky, and M.-H. Hall, “Analysis of risk factor domains in psychosis patient health records,” *Journal of biomedical semantics*, vol. 10, pp. 1–10, 10 2019.

[10] B. Norgeot, K. Muenzen, T. Peterson, X. Fan, B. S. Glicksberg, G. Schenk, E. Rutenberg, B. Oskotsky, M. Sirota, J. Yazdany, G. Schmajuk, D. Ludwig, T. C. Goldstein, and A. J. Butte, “Protected health information filter (philter): accurately and securely de-identifying free-text clinical notes,” *NPJ digital medicine*, vol. 3, pp. 57–, 4 2020.

[11] R. T. Chen, J. C. Ho, and J.-M. S. Lin, “Extracting medication information from unstructured public health data: a demonstration on data from population-based and tertiary-based samples,” *BMC medical research methodology*, vol. 20, pp. 1–11, 10 2020.

[12] C. Shorten, T. M. Khoshgoftaar, and B. Furht, “Text data augmentation for deep learning.,” *Journal of big data*, vol. 8, pp. 101–, 7 2021.

[13] C. Caliskan and A. Kilicaslan, “Varieties of corona news: a cross-national study on the foundations of online misinformation production during the covid-19 pandemic.,” *Journal of computational social science*, vol. 6, pp. 191–243, 12 2022.

[14] J. Mower, T. Cohen, and D. Subramanian, “Complementing observational signals with literature-derived distributed representations for post-marketing drug surveillance.,” *Drug safety*, vol. 43, pp. 67–77, 10 2019.

[15] J. A. Sparks, W. Huang, B. Lu, S. Huang, A. Cagan, V. S. Gainer, S. Finan, G. Savova, D. H. Solomon, E. W. Karlson, and K. P. Liao, “Op0111 rheumatoid arthritis serologic phenotype at diagnosis and subsequent risk for pneumonia identified using machine learning approaches,” *Annals of the Rheumatic Diseases*, vol. 79, pp. 73–73, 2020.

[16] D. Abatemarco, S. Perera, S. Bao, S. Desai, B. Assuncao, N. Tetarenko, K. Danysz, R. Mockute, M. Widdowson, N. Fornarotto, S. Beauchamp, S. Cicirello, and E. Mingle, “Training augmented intelligent capabilities for pharmacovigilance: Applying deep-learning approaches to individual case safety report processing.,” *Pharmaceutical medicine*, vol. 32, pp. 391–401, 10 2018.

[17] C. Su, Z. Xu, J. Pathak, and F. Wang, “Deep learning in mental health outcome research: a scoping review,” *Translational psychiatry*, vol. 10, pp. 116–116, 4 2020.

[18] J. R. Machireddy, “Harnessing ai and data analytics for smarter healthcare solutions,” *International Journal of Science and Research Archive*, vol. 08, no. 02, pp. 785–798, 2023.

[19] A. N. Berman, C. Ginder, Z. A. Sporn, V. Tanguturi, M. K. Hidrue, L. B. Shirkey, Y. Zhao, R. Blankstein, A. Turchin, and J. H. Wasfy, “Natural language processing for the ascertainment and phenotyping of left ventricular hypertrophy and hypertrophic cardiomyopathy on echocardiogram reports.,” *The American journal of cardiology*, vol. 206, pp. 247–253, 9 2023.

[20] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, “Context-aware query performance optimization for big data analytics in healthcare,” in *2019 IEEE High Performance Extreme Computing Conference (HPEC-2019)*, pp. 1–7, 2019.

[21] A. Sharma and K. Goolsbey, “Identifying useful inference paths in large commonsense knowledge bases by retrograde analysis,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[22] B. Kochar, W. Cai, and A. N. Ananthakrishnan, “Inflammatory bowel disease patients who respond to treatment with anti-tumor necrosis factor agents demonstrate improvement in pre-treatment frailty.,” *Digestive diseases and sciences*, vol. 67, pp. 1–7, 5 2021.

[23] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, “Federated query processing for big data in data science,” in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 6145–6147, IEEE, 2019.

[24] K. Bi, L. Scalschi, N. Jaiswal, T. Mengiste, R. Fried, A. B. Sanz, J. Arroyo, W. Zhu, G. Masrati, and A. Sharon, "The botrytis cinerea crh1 transglycosylase is a cytoplasmic effector triggering plant cell death and defense response.," *Nature communications*, vol. 12, pp. 2166–2166, 4 2021.

[25] L. Tang, Z. Sun, B. Idnay, J. G. Nestor, A. Soroush, P. A. Elias, Z. Xu, Y. Ding, G. Durrett, J. F. Rousseau, C. Weng, and Y. Peng, "Evaluating large language models on medical evidence summarization.," *NPJ digital medicine*, vol. 6, pp. 158–, 8 2023.

[26] S. K. Moulik, N. Kotter, and E. K. Fishman, "Applications of artificial intelligence in the emergency department.," *Emergency radiology*, vol. 27, pp. 355–358, 7 2020.

[27] R. Avula, "Applications of bayesian statistics in healthcare for improving predictive modeling, decision-making, and adaptive personalized medicine," *International Journal of Applied Health Care Analytics*, vol. 7, no. 11, pp. 29–43, 2022.

[28] M. T. Caton, W. F. Wiggins, S. R. Pomerantz, and K. P. Andriole, "Effects of age and sex on the distribution and symmetry of lumbar spinal and neural foraminal stenosis: a natural language processing analysis of 43,255 lumbar mri reports.," *Neuroradiology*, vol. 63, pp. 959–966, 2 2021.

[29] P. Pham, C. Cheng, E. Wu, I. Kim, R. Zhang, Y. Ma, C. Kortepeter, and M. A. Muñoz, "Leveraging case narratives to enhance patient age ascertainment from adverse event reports," *Pharmaceutical medicine*, vol. 35, pp. 307–316, 9 2021.

[30] V. Kashyap, C. Ramakrishnan, C. Thomas, and A. P. Sheth, "Taxaminer: an experimentation framework for automated taxonomy bootstrapping," *International Journal of Web and Grid Services*, vol. 1, no. 2, pp. 240–266, 2005.

[31] S. Šćepanović, M. Constantinides, D. Quercia, and S. Kim, "Quantifying the impact of positive stress on companies from online employee reviews.," *Scientific reports*, vol. 13, pp. 1603–, 1 2023.

[32] A. K. Saxena, "Evaluating the regulatory and policy recommendations for promoting information diversity in the digital age," *International Journal of Responsible Artificial Intelligence*, vol. 11, no. 8, pp. 33–42, 2021.

[33] M. Alobaidi, K. M. Malik, and S. Sabra, "Linked open data-based framework for automatic biomedical ontology generation," *BMC bioinformatics*, vol. 19, pp. 1–13, 9 2018.

[34] A. Sharma, *Structural and network-based methods for knowledge-based systems*. PhD thesis, Northwestern University, 2011.

[35] J. W. Berlin and R. Duszak, "The role of the radiologist in new payment systems.," *Abdominal radiology (New York)*, vol. 41, pp. 461–465, 3 2016.

[36] A. Jorge, V. M. Castro, A. Barnado, V. S. Gainer, C. Hong, T. Cai, T. Cai, R. J. Carroll, J. C. Denny, L. J. Crofford, K. H. Costenbader, K. P. Liao, E. W. Karlson, and C. H. Feldman, "Identifying lupus patients in electronic health records: Development and validation of machine learning algorithms and application of rule-based algorithms.," *Seminars in arthritis and rheumatism*, vol. 49, pp. 84–90, 1 2019.

[37] M. B. Imerman and F. J. Fabozzi, "Cashing in on innovation: a taxonomy of fintech," *Journal of Asset Management*, vol. 21, pp. 167–177, 5 2020.

[38] M. Yuan and A. Vlachos, "Zero-shot fact-checking with semantic triples and knowledge graphs," in *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pp. 105–115, 2024.

[39] V. Sangha, B. J. Mortazavi, A. D. Haimovich, A. H. Ribeiro, C. A. Brandt, D. L. Jacoby, W. L. Schulz, H. M. Krumholz, A. L. P. Ribeiro, and R. Khera, "Automated multilabel diagnosis on electrocardiographic images and signals.," *Nature communications*, vol. 13, pp. 1583–, 3 2022.

[40] R. Avula *et al.*, "Data-driven decision-making in healthcare through advanced data mining techniques: A survey on applications and limitations," *International Journal of Applied Machine Learning and Computational Intelligence*, vol. 12, no. 4, pp. 64–85, 2022.

[41] R. Castilla-Puentes, A. Dagar, D. Villanueva, L. Jimenez-Parrado, L. G. Valleta, and T. Falcone, "Digital conversations about depression among hispanics and non-hispanics in the us: a big-data, machine learning analysis identifies specific characteristics of depression narratives in hispanics.," *Annals of general psychiatry*, vol. 20, pp. 50–, 11 2021.

[42] A. Sharma and K. Forbus, "Modeling the evolution of knowledge in learning systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, pp. 669–675, 2012.

[43] Özlem Uzuner, Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de-identification," *Journal of the American Medical Informatics Association : JAMIA*, vol. 14, pp. 550–563, 6 2007.

[44] C. Northuis, M. Michalowski, and K. Lakshminarayan, "Abstract wp382: Using natural language processing algorithms to identify stroke cases and stroke subtypes from neuroimaging reports," *Stroke*, vol. 50, no. Suppl_1, 2019.

[45] B. J. Marafino, W. J. Boscardin, and R. A. Dudley, "Efficient and sparse feature selection for biomedical text classification via the elastic net," *Journal of biomedical informatics*, vol. 54, pp. 114–120, 2 2015.

[46] W. Shao, X. Luo, Z. Zhang, Z. Han, V. Chandrasekaran, V. Turzhitsky, V. Bali, A. R. Roberts, M. Metzger, J. Baker, C. L. Rosa, J. Weaver, P. Dexter, and K. Huang, "Application of unsupervised deep learning algorithms for identification of specific clusters of chronic cough patients from emr data.," *BMC bioinformatics*, vol. 23, pp. 140–, 4 2022.

[47] N. Rajeevan, K. M. Niehoff, P. Charpentier, F. L. Levin, A. C. Justice, C. Brandt, T. R. Fried, and P. L. Miller, "Utilizing patient data from the veterans administration electronic health record to support web-based clinical decision support: informatics challenges and issues from three clinical domains," *BMC medical informatics and decision making*, vol. 17, pp. 111–111, 7 2017.

[48] J. H. Garvin, M. Kalsy, C. Brandt, S. L. Luther, G. Divita, G. Coronado, D. Redd, C. Christensen, B. Hill, N. Kelly, and Q. Z. Treitler, "An evolving ecosystem for natural language processing in department of veterans affairs," *Journal of medical systems*, vol. 41, pp. 1–9, 1 2017.

[49] M. S. Lim, T. Beyer, A. Babayan, M. Bergmann, M. Brehme, A. Buyx, J. Czernin, G. Egger, K. S. J. Elenitoba-Johnson, B. Gückel, A. Jačan, H. Haslacher, R. J. Hicks, L. Kenner, M. Langanke, M. Mitterhauser, B. J. Pichler, H. R. Salih, R. Schibli, S. Schulz, J. Simecek, J. Simon, M. Soares, U. Stelzl, W. Wadsak, K. Zatloukal, M. Zeitlinger, and M. R. Hacker, "Advancing biomarker development through convergent engagement: Summary report of the 2nd international danube symposium on biomarker development, molecular imaging and applied diagnostics; march 14–16, 2018; vienna, austria," *Molecular imaging and biology*, vol. 22, pp. 47–65, 5 2019.

[50] M. Conway, S. Keyhani, L. M. Christensen, B. R. South, M. Vali, L. C. Walter, D. L. Mowery, S. E. AbdelRahman, and W. W. Chapman, "Moonstone: a novel natural language processing system for inferring social risk from clinical narratives," *Journal of biomedical semantics*, vol. 10, pp. 6–6, 4 2019.

[51] L. He, M. Moldenhauer, K. Zheng, and H. Ma, "Analyzing free-text clinical narratives for veterans with lymphoid malignancies using natural language processing (nlp).," *Journal of Clinical Oncology*, vol. 41, pp. e13576–e13576, 6 2023.

[52] K. H. Goh, L. Wang, A. Yeow, H. M. N. Poh, K. Li, J. J. L. Yeow, and G. Y. H. Tan, "Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare," *Nature communications*, vol. 12, pp. 711–711, 1 2021.

[53] G. Arnold, M. Klasic, C. Wu, M. Schomburg, and A. York, "Finding, distinguishing, and understanding overlooked policy entrepreneurs," *Policy Sciences*, vol. 56, pp. 657–687, 11 2023.

[54] W. Hu, Y. Qu, and X.-Z. Sun, "Bootstrapping object coreferencing on the semantic web," *Journal of Computer Science and Technology*, vol. 26, pp. 663–675, 7 2011.

[55] A. Fong, N. Harriott, D. M. Walters, H. Foley, R. Morrissey, and R. R. Ratwani, "Integrating natural language processing expertise with patient safety event review committees to improve the analysis of medication events," *International journal of medical informatics*, vol. 104, pp. 120–125, 5 2017.

[56] S. K. Parr and G. T. Gobbel, "Considerations for advancing nephrology research and practice through natural language processing," *Kidney international*, vol. 97, no. 2, pp. 263–265, 2020.

[57] C. Xiang and M. Abouelyazid, "The impact of generational cohorts and visit environment on telemedicine satisfaction: A novel investigation," 2020.

[58] B. I. Reiner, "Quantitative analysis of uncertainty in medical reporting: Creating a standardized and objective methodology.," *Journal of digital imaging*, vol. 31, pp. 145–149, 12 2017.

[59] C. L. Overby, P. Tarczy-Hornoch, J. Hoath, I. J. Kalet, and D. L. Veenstra, "Feasibility of incorporating genomic knowledge into electronic medical records for pharmacogenomic clinical decision support," *BMC bioinformatics*, vol. 11, pp. 1–9, 10 2010.

[60] H. Kaur, S. Sohn, C. I. Wi, E. Ryu, M. A. Park, K. Bachman, H. Kita, I. T. Croghan, J. A. Castro-Rodriguez, G. A. Voge, H. Liu, and Y. J. Juhn, "Automated chart review utilizing natural language processing algorithm for asthma predictive index.," *BMC pulmonary medicine*, vol. 18, pp. 34–34, 2 2018.

[61] Z. B. Millman, D. J. Holt, M. S. Keshavan, L. Seidman, A. Breier, M. E. Shenton, D. Öngür, S. V. Abram, J. P. Hua, S. Nicholas, B. J. Roach, S. K. Keedy, J. A. Sweeney, D. H. Mathalon, J. M. Ford, S. Erhardt, Y. Gravenfors, C. Savage, G. Prettyman, P. Didier, J. W. Kable, T. D. Satterthwaite, C. Davatzikos, R. T. Shinohara, M. E. Calkins, M. Elliott, R. C. Gur, R. Gur, and D. H. Wolf, "Acnp 62nd annual meeting: Poster abstracts p501 – p753," *Neuropsychopharmacology*, vol. 48, pp. 355–495, 12 2023.

[62] B. W. Patterson, G. C. Jacobsohn, M. N. Shah, Y. Song, A. Maru, A. K. Venkatesh, M. Zhong, K. Taylor, A. G. Hamedani, and E. A. Mendonça, "Development and validation of a pragmatic natural language processing approach to identifying falls in older adults in the emergency department," *BMC medical informatics and decision making*, vol. 19, pp. 138–138, 7 2019.

[63] A. Sharma and K. D. Forbus, "Graph-based reasoning and reinforcement learning for improving q/a performance in large knowledge-based systems," in *2010 AAAI Fall Symposium Series*, 2010.

[64] M. Filannino and Özlem Uzuner, "Advancing the state of the art in clinical natural language processing through shared tasks.," *Yearbook of medical informatics*, vol. 27, pp. 184–192, 8 2018.

[65] F. Huang, S. Zhang, M. He, and X. Wu, "Clustering web documents using hierarchical representation with multi-granularity," *World Wide Web*, vol. 17, pp. 105–126, 1 2013.

[66] M. Hassani and S. D. Young, "Potential role of conversational agents in encouraging prep uptake.," *The journal of behavioral health services & research*, 5 2022.

[67] L. A. Lekham, Y. Wang, E. Hey, and M. T. Khasawneh, "Multi-label text mining to identify reasons for appointments to drive population health analytics at a primary care setting," *Neural Computing and Applications*, vol. 34, pp. 14971–15005, 5 2022.

[68] B. Roth, R. Kampalath, K. Nakashima, S. Shieh, T.-L. Bui, and R. Houshyar, "Revenue and cost analysis of a system utilizing natural language processing and a nurse coordinator for radiology follow-up recommendations.," *Current problems in diagnostic radiology*, vol. 52, pp. 367–371, 5 2023.

[69] H. Li, R. C. Gerkin, A. Bakke, R. Norel, G. Cecchi, C. Laudamiel, M. Y. Niv, K. Ohla, J. E. Hayes, V. Parma, and P. Meyer, "Text-based predictions of covid-19 diagnosis from self-reported chemosensory descriptions.," *Communications medicine*, vol. 3, pp. 104–, 7 2023.

[70] D. Gruhl, M. Nagarajan, J. Pieper, C. Robson, and A. P. Sheth, "Multimodal social intelligence in a real-time dashboard system," *The VLDB Journal*, vol. 19, pp. 825–848, 12 2010.

[71] R. Deka, J. K. Parsons, D. R. Simpson, P. Riviere, V. Nalawade, L. K. Vitzthum, A. K. Kader, C. J. Kane, C. S. Rock, J. D. Murphy, and B. S. Rose, "African-american men with low-risk prostate cancer treated with radical prostatectomy in an equal-access health care system: implications for active surveillance," *Prostate cancer and prostatic diseases*, vol. 23, pp. 581–588, 4 2020.

[72] Óscar Ferrández, B. R. South, S. Shen, F. J. Friedlin, M. H. Samore, and S. M. Meystre, "Evaluating current automatic de-identification methods with veteran's health administration clinical documents," *BMC medical research methodology*, vol. 12, pp. 109–109, 7 2012.

[73] T. Cai, Z. He, C. Hong, Y. Zhang, Y.-L. Ho, J. Honerlaw, A. Geva, V. A. Panickan, A. King, D. R. Gagnon, M. Gaziano, K. Cho, K. Liao, and T. Cai, "Scalable relevance ranking algorithm via semantic similarity assessment improves efficiency of medical chart review.," *Journal of biomedical informatics*, vol. 132, pp. 104109–104109, 6 2022.

[74] S.-A. Hussain, E. Sezgin, K. Krivchenia, J. Luna, S. Rust, and Y. Huang, "A natural language processing pipeline to synthesize patient-generated notes toward improving remote care and chronic disease management: a cystic fibrosis case study.," *JAMIA open*, vol. 4, pp. ooab084–, 7 2021.

[75] P. Damacharla, P. Dhakal, S. Stumbo, A. Y. Javaid, S. Ganapathy, D. A. Malek, D. C. Hodge, and V. K. Devabhaktuni, "Effects of voice-based synthetic assistant on performance of emergency care provider in training," *International Journal of Artificial Intelligence in Education*, vol. 29, pp. 122–143, 3 2018.

[76] A. J. Alsheikh, S. Wollenhaupt, E. A. King, J. Reeb, S. Ghosh, L. R. Stolzenburg, S. Tamim, J. Lazar, J. W. Davis, and H. J. Jacob, "The landscape of gwas validation; systematic review identifying 309 validated non-coding variants across 130 human diseases.," *BMC medical genomics*, vol. 15, pp. 74–, 4 2022.

[77] J. Xu, S. Kim, M. Song, M. Jeong, D. Kim, J. Kang, J. F. Rousseau, X. Li, W. Xu, V. I. Torvik, Y. Bu, C. Chen, I. A. Ebeid, D. Li, and Y. Ding, "Building a pubmed knowledge graph.," *Scientific data*, vol. 7, pp. 205–, 6 2020.

[78] A. Edinger, D. Valdez, E. Walsh-Buhi, and J. Bollen, "Deep learning for topical trend discovery in online discourse about pre-exposure prophylaxis (prep).," *AIDS and behavior*, vol. 27, pp. 443–453, 8 2022.

[79] M. M. P. Kuijten, M. H. Degeling, J. W. Chen, G. R. Wojtkiewicz, P. Waterman, R. Weissleder, J. Azzi, K. Nicolay, and B. A. Tannous, "Multimodal targeted high relaxivity thermosensitive liposome for in vivo imaging.," *Scientific reports*, vol. 5, pp. 17220–17220, 11 2015.

[80] F. Rustam, Z. Imtiaz, A. Mehmood, V. Rupapara, G. S. Choi, S. Din, and I. Ashraf, "Automated disease diagnosis and precaution recommender system using supervised machine learning," *Multimedia Tools and Applications*, vol. 81, pp. 31929–31952, 4 2022.

[81] J. Singleton, C. Li, P. D. Akpunonu, E. L. Abner, and A. M. Kucharska-Newton, "Using natural language processing to identify opioid use disorder in electronic health record data.," *International journal of medical informatics*, vol. 170, pp. 104963–104963, 12 2022.

[82] R. J. Carroll, W. K. Thompson, A. E. Eyler, A. M. Mandelin, T. Cai, R. Zink, J. A. Pacheco, C. S. Boomershine, T. A. Lasko, H. Xu, E. W. Karlson, R. G. Perez, V. S. Gainer, S. N. Murphy, E. Ruderman, R. M. Pope, R. M. Plenge, A. N. Kho, K. P. Liao, and J. C. Denny, "Portability of an algorithm to identify rheumatoid arthritis in electronic health records," *Journal of the American Medical Informatics Association : JAMIA*, vol. 19, pp. e162–9, 2 2012.

[83] J. R. Machireddy, "Automation in healthcare claims processing: Enhancing efficiency and accuracy," *International Journal of Science and Research Archive*, vol. 09, no. 01, pp. 825–834, 2023.

[84] M. J. Cook, L. Yao, and X. Wang, "Facilitating accurate health provider directories using natural language processing.," *BMC medical informatics and decision making*, vol. 19, pp. 80–80, 4 2019.

[85] S. A. Graham, C. A. Depp, E. E. Lee, C. Nebeker, X. M. Tu, H.-C. Kim, and D. V. Jeste, "Artificial intelligence for mental health and mental illnesses: an overview.," *Current psychiatry reports*, vol. 21, pp. 116–116, 11 2019.

[86] E. O. Pyzer-Knapp, J. W. Pitera, P. W. J. Staar, S. Takeda, T. Laino, D. P. Sanders, J. Sexton, J. R. Smith, and A. Curioni, "Accelerating materials discovery using artificial intelligence, high performance computing and robotics," *npj Computational Materials*, vol. 8, 4 2022.

[87] J. P. Ferraro, H. Daumé, S. L. DuVall, W. W. Chapman, H. Harkema, and P. J. Haug, "Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation," *Journal of the American Medical Informatics Association : JAMIA*, vol. 20, pp. 931–939, 3 2013.

[88] A. S. Villar and N. Khan, "Robotic process automation in banking industry: a case study on deutsche bank," *Journal of Banking and Financial Technology*, vol. 5, pp. 71–86, 5 2021.

[89] H. Hochheiser, M. Castine, D. J. Harris, G. Savova, and R. S. Jacobson, "An information model for computable cancer phenotypes," *BMC medical informatics and decision making*, vol. 16, pp. 121–121, 9 2016.

[90] H. Chen, S. M. Lundberg, G. Erion, J. H. Kim, and S.-I. Lee, "Forecasting adverse surgical events using self-supervised transfer learning for physiological signals.," *NPJ digital medicine*, vol. 4, pp. 167–, 12 2021.

[91] A. Sharma and K. M. Goolsbey, "Learning search policies in large commonsense knowledge bases by randomized exploration," 2018.

[92] M. Robila and S. A. Robila, "Applications of artificial intelligence methodologies to behavioral and social sciences," *Journal of Child and Family Studies*, vol. 29, pp. 2954–2966, 12 2019.

[93] C. Baechle and A. Agarwal, "A framework for the estimation and reduction of hospital readmission penalties using predictive analytics," *Journal of Big Data*, vol. 4, pp. 1–15, 11 2017.

[94] C. Robertson, G. Mukherjee, H. Gooding, S. Kandaswamy, and E. Orenstein, "A method to advance adolescent sexual health research: Automated algorithm finds sexual history documentation.," *Frontiers in digital health*, vol. 4, pp. 836733–, 7 2022.

[95] P. Lakhani and C. P. Langlotz, "Automated detection of radiology reports that document non-routine communication of critical or significant results," *Journal of digital imaging*, vol. 23, pp. 647–657, 10 2009.

[96] H. Forsvik, V. Voipio, J. Lamminen, P. Doupi, H. Hyppönen, and R. Vuokko, "Literature review of patient record structures from the physician's perspective," *Journal of medical systems*, vol. 41, pp. 1–10, 12 2016.

[97] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, "Automatic visual recommendation for data science and analytics," in *Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 2*, pp. 125–132, Springer, 2020.