

ORIGINAL RESEARCH

Open access

Multimodal Transformer Architecture for ARDS Detection: A Framework Integrating Chest X-Ray, Clinical Notes, and Laboratory Values

Ravi Kumar^{1*}, Neha Sharma¹, Aniket Deshmukh², Arjun Nair¹, Meera Pillai²

Abstract

The Berlin definition of ARDS provides standardized diagnostic criteria based on acute onset within one week of a known insult, bilateral chest imaging opacities not explained by other causes, respiratory failure not due to cardiac issues or fluid overload, and impaired oxygenation measured by the $\text{PaO}_2/\text{FiO}_2$ ratio, enabling consistent identification in intensive care; however, its clinical use is limited by variability in imaging interpretation and the need for rapid decision-making, often causing delays and inconsistent diagnoses. Current practice relies heavily on subjective assessment of chest X-rays and limited integration of clinical notes and laboratory trends, resulting in moderate inter-observer agreement and reduced diagnostic reliability. To overcome these challenges, a multimodal transformer framework is proposed that integrates chest X-rays, clinical notes, and laboratory data using vision transformers, BERT-based text encoders, and temporally aware lab embeddings, with cross-modal attention enabling interaction across data types and a fusion module producing final ARDS probability estimates. This integrated approach improves diagnostic accuracy by combining complementary information, enhances interpretability through attention mechanisms, and offers a more objective and timely method for ARDS detection, with potential to support earlier intervention and better outcomes in critically ill patients.

Keywords ARDS detection, Multimodal transformer, Chest X-ray analysis, Clinical natural language processing, Laboratory value integration, Cross-modal attention

*Correspondence:

Ravi Kumar

ravi.kumar@outlook.com

¹ Department of Healthcare AI Systems, IIT Delhi, New Delhi, India

² Department of Clinical Analytics and Intelligent Systems, IIT Bombay, Mumbai, India

Introduction

The Berlin definition serves as the cornerstone for diagnosing acute respiratory distress syndrome, specifying four key criteria: timing of onset, bilateral chest imaging opacities, non-cardiogenic origin of pulmonary edema, and hypoxemia quantified by the $\text{PaO}_2/\text{FiO}_2$ ratio [1]. In intensive care units worldwide, ARDS manifests in a substantial proportion of mechanically ventilated patients, contributing to high mortality rates estimated between 35% and 45% across severity levels [2]. Accurate and timely identification remains pivotal for guiding supportive

therapies such as lung-protective ventilation and prone positioning. The standardized criteria have improved research consistency but highlight the ongoing need for enhanced diagnostic support tools in dynamic clinical environments [3].

Diagnostic challenges in ARDS stem primarily from inter-observer variability in interpreting chest radiographs for bilateral opacities consistent with the Berlin criteria [1]. Clinicians must exclude alternative explanations such as atelectasis or pleural effusions, which requires substantial expertise and can delay confirmation of the diagnosis [3].

This subjectivity is compounded by the heterogeneous presentation of ARDS in critically ill patients with multiple comorbidities. Consequently, there is a pressing demand for systems that augment human judgment with data-driven insights to improve reliability [2].

Unimodal artificial intelligence approaches exhibit inherent limitations when applied to ARDS detection. Chest X-ray classification models, while effective for opacity detection, overlook critical contextual information embedded in clinical narratives and physiological trends [4]. Similarly, natural language processing of clinical notes alone cannot verify imaging findings essential to the Berlin definition [5]. Laboratory value analysis, focused on oxygenation indices, lacks the visual and textual corroboration needed for comprehensive assessment [6]. These gaps underscore the necessity for integrated multimodal strategies [7].

Table 1 provides a structured analytical comparison highlighting the conceptual limitations of unimodal approaches relative to the proposed multimodal transformer framework.

Table 1. Analytical Comparison of Unimodal versus Multimodal Paradigms for ARDS Detection

Dimension	Chest X-Ray Models	Clinical NLP Models	Latent Mechanisms
Primary Data Representation	Spatial imaging features	Sequential linguistic tokens	Temporal sequencing
Alignment with Berlin Criteria	Partial (opacity detection only)	Partial (symptom/context only)	Fused information (oxygenation metrics)
Dependency on Expert Interpretation	High	Moderate	Reduced dependency
Sensitivity to Missing Data	Low (image required)	Moderate	Highly context-aware
Ability to Capture Cross-	None	None	Integrated insights

Domain Relationships			
Interpretability Mechanism	Saliency maps	Attention weights	Fused information
Temporal Reasoning	Limited	Moderate	Seamless integration
Diagnostic Consistency	Moderate	Variable	Validated insights
Failure Mode	Misclassification due to visual ambiguity	Context misinterpretation	Loss of critical details

This conceptual framework proposes a multimodal transformer architecture that integrates chest X-ray images, clinical notes, and laboratory values to enable robust ARDS detection [8]. By employing cross-modal attention mechanisms, the model learns to fuse information across modalities in a principled manner [9]. The architecture offers a roadmap for advancing AI applications in critical care, beginning with background on ARDS and data modalities, followed by detailed framework design, input processing, and transformer components [10]. Ultimately, this design aims to support clinicians in achieving more objective and timely diagnoses [2].

Background

ARDS definition and diagnostic criteria

The Berlin definition of ARDS, jointly developed by international experts, delineates precise criteria to classify the syndrome into mild, moderate, and severe categories based on the PaO2/FiO2 ratio under positive end-expiratory pressure conditions [1]. It mandates that the onset occurs within one week of a known clinical insult, with chest imaging showing bilateral opacities not fully explained by other causes such as effusions or collapse [3]. Additionally, the respiratory failure must not be attributable to cardiac failure or fluid overload, often assessed via objective measures or clinical judgment [2]. These elements collectively provide a structured approach to

diagnosis that has been widely adopted in both research and practice.

Despite its standardization benefits, the Berlin criteria face practical limitations in application, particularly when clinical data are incomplete or ambiguous in fast-paced ICU settings [1]. Machine learning approaches have begun exploring enhancements to these criteria through computational phenotyping, yet they often remain reliant on single data streams [2]. Integration of multiple modalities could refine the application of the definition by providing corroborative evidence automatically [3]. This evolution toward data-driven refinement holds promise for reducing diagnostic uncertainty [2].

Chest X-Ray for ARDS detection

Chest X-rays play a central role in the Berlin definition by confirming the presence of bilateral opacities indicative of non-cardiogenic pulmonary edema [1]. Differential diagnosis on imaging includes conditions such as pneumonia, atelectasis, or pulmonary hemorrhage, which complicates accurate classification [11]. Inter-rater reliability for these assessments has historically been moderate, necessitating advanced computational aids [1]. Deep learning techniques for chest X-ray interpretation have demonstrated potential in automating opacity detection but require contextual supplementation [4].

While deep learning models excel at feature extraction from chest radiographs for classification tasks, they typically operate in isolation from other clinical data sources [12]. This unimodal focus limits their utility in fully satisfying the multifaceted Berlin criteria for ARDS [13]. Incorporating additional modalities could enhance the specificity and sensitivity of imaging-based predictions [14]. Conceptual advancements in this area emphasize the value of hybrid architectures for comprehensive analysis [11].

Clinical notes and NLP in critical care

Clinical notes in intensive care units contain rich narrative information on patient progress, including ventilator settings, physical examination findings such as auscultation results, and evolving symptoms relevant to ARDS development [5]. Natural language processing techniques enable the extraction of structured insights from these unstructured texts, supporting decision-making in critical care [15]. Variants of BERT tailored for clinical domains, such as ClinicalBERT, have proven effective in processing

domain-specific language [5]. These tools facilitate the identification of subtle indicators that complement imaging data [16].

BioBERT and related models further enhance performance by incorporating biomedical knowledge during pretraining, making them suitable for intensive care documentation analysis [17]. Section segmentation within notes allows targeted processing of relevant portions like assessment and plan sections [18]. However, standalone NLP approaches lack the visual confirmation provided by imaging, highlighting the need for multimodal integration [15]. This processing step forms a foundational component for synthesizing textual evidence in ARDS frameworks [5].

Laboratory values in ARDS

Laboratory values are integral to the Berlin definition, particularly the PaO₂/FiO₂ ratio which stratifies ARDS severity and guides oxygenation support [3]. Inflammatory markers such as C-reactive protein or lactate levels provide additional context on systemic involvement and potential etiologies [2]. Serial measurements enable trend analysis that can signal progression or response to therapy in real time [10]. Proper embedding of these temporal sequences is essential for predictive modeling in critical care [6].

Handling laboratory data requires attention to missing values and normalization across different measurement timestamps [19]. Multimodal frameworks can incorporate these numerical features alongside imaging and text to create a holistic patient profile [7]. Trend features derived from sequential labs enhance the detection of acute changes consistent with ARDS onset [3]. Such integration promises to strengthen the overall diagnostic process beyond isolated metrics [2].

Framework Overview

High-level architecture

The proposed framework adopts a high-level architecture consisting of three distinct input modalities processed through dedicated encoders before converging in a cross-modal transformer [8]. Chest X-ray images, clinical notes, and laboratory values feed into their respective modality-specific components, enabling specialized feature extraction [7]. The outputs are then aligned via transformer layers that facilitate information exchange across domains

[10]. A final fusion stage aggregates these representations for downstream classification [19].

Figure 1 illustrates the hierarchical multimodal transformer architecture integrating chest X-ray imaging, clinical notes, and laboratory values through modality-specific encoders and cross-modal attention for ARDS detection.

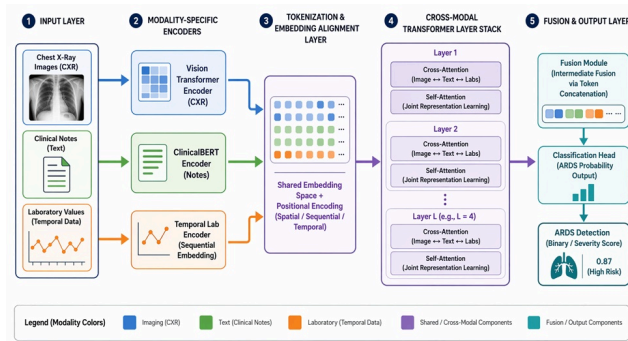


Figure 1. Hierarchical Multimodal Transformer Architecture for ARDS Detection Integrating Imaging, Textual, and Laboratory Modalities

This sequential design from modality encoders to joint transformer ensures that modality-unique characteristics are preserved initially while allowing for rich interactions later [20]. The architecture supports end-to-end learning without manual feature engineering, aligning with modern AI paradigms in healthcare [21]. By structuring the flow in this manner, the framework achieves scalability for diverse ICU data streams [7]. Conceptual benefits include improved robustness to variations in data quality across modalities [19].

Core assumptions

The framework assumes the availability of paired multimodal data, including contemporaneous chest X-rays, clinical notes, and laboratory results from the same patient encounters in ICU settings [15]. Sufficient labeled data for ARDS outcomes, derived from established criteria, is also presumed to enable supervised training [2]. These assumptions reflect typical electronic health record ecosystems in modern hospitals [3]. Data quality and completeness are expected to meet thresholds suitable for deep learning applications [7].

Labeled datasets from large-scale ICU repositories provide the foundation for model development under these assumptions [20]. Strategies for data augmentation can further enhance generalizability without violating the paired

data premise [22]. The framework does not require perfectly aligned timestamps but accommodates reasonable temporal proximity [10]. This flexibility supports practical deployment in real clinical workflows [19].

Design principles

Core design principles emphasize modality-agnostic processing to accommodate varying data types without custom pipelines for each [19]. Attention-driven mechanisms are prioritized to capture dependencies both within and across modalities effectively [9]. The architecture promotes explainability by design, allowing inspection of contribution from each input source [23]. These principles ensure the framework remains adaptable to future extensions or additional modalities [24].

Explainability is further supported through attention weight visualization, aiding clinician trust and validation [25]. The design avoids over-reliance on any single modality by enforcing balanced cross-modal learning [26]. Overall, these principles align with best practices in artificial intelligence for healthcare systems [8]. They facilitate the creation of a robust and interpretable tool for ARDS detection [7].

Multimodal Input Processing

Chest X-Ray processing

Chest X-ray processing in the framework utilizes a vision transformer backbone or CNN-based feature extractor to handle image inputs effectively [4]. Input images are divided into patches with positional encodings added to retain spatial information [27]. This patching strategy allows the model to learn local and global patterns associated with ARDS-related opacities [11]. Preprocessing steps such as normalization and resizing ensure compatibility with the encoder architecture [12].

The vision transformer approach captures long-range dependencies within the image that are crucial for identifying diffuse bilateral changes [21]. Positional encoding enhances the model's understanding of anatomical structures across the radiograph [8]. Conceptual integration of these techniques supports accurate feature representation for subsequent cross-modal fusion [14]. This processing pipeline prepares high-quality embeddings tailored to the imaging modality [13].

Clinical note processing

Clinical note processing begins with tokenization using clinical-specific vocabularies followed by embedding via models such as ClinicalBERT [5]. Notes are segmented into sections like history, examination, and assessment to focus on ARDS-relevant content including ventilator parameters and auscultation findings [15]. This structured approach improves the quality of textual representations [17]. Pretrained embeddings capture semantic nuances inherent to critical care documentation [16].

Advanced variants like BioBERT incorporate domain knowledge to better handle medical terminology in notes [17]. The processing accounts for temporal aspects by ordering notes chronologically when multiple entries exist [18]. Resulting embeddings encode contextual information that complements imaging and lab data [5]. Such detailed input handling is vital for the multimodal transformer's effectiveness [15].

Laboratory value processing

Laboratory value processing involves embedding numerical measurements with associated timestamps to capture temporal dynamics [10]. Each lab result is transformed into a vector representation, incorporating trend features such as deltas or rates of change for markers like PaO2/FiO2 [6]. Missing values are handled through imputation or masking techniques integrated into the encoder [19]. This method preserves the sequential nature of laboratory monitoring in ARDS cases [3].

Time-stamped embeddings allow the model to weigh recent versus historical values appropriately in the context of acute respiratory changes [7]. Normalization and scaling ensure consistency across different laboratory panels [2]. The processing pipeline generates modality-specific features ready for cross-modal interaction [22]. Conceptual design here emphasizes robustness to incomplete laboratory profiles common in ICU data [19].

Transformer Architecture

Modality-specific encoders

Modality-specific encoders form the initial stage of the transformer architecture, with separate pathways for imaging, text, and laboratory data to extract specialized features [8]. The chest X-ray encoder employs vision

transformer layers, while the clinical note encoder leverages transformer-based language models [5]. Laboratory encoders utilize feed-forward networks tailored for sequential numerical inputs [10]. This separation preserves the unique structural properties of each modality prior to fusion [7].

Independent encoding prevents early dilution of modality-specific signals that are critical for ARDS-related patterns [20]. Each encoder can be initialized with pretrained weights from domain-relevant tasks to accelerate convergence conceptually [4]. The outputs are projected into a common embedding space for compatibility in later stages [21]. Such architecture design enhances the overall capacity for multimodal representation learning [19].

Cross-modal transformer layers

Cross-modal transformer layers implement self-attention mechanisms that operate across the combined token sequences from all modalities [9]. Learnable cross-attention modules allow queries from one modality to attend to keys and values from others, facilitating information flow [24]. This enables the model to dynamically align, for instance, specific image regions with corresponding phrases in clinical notes [23]. The layers stack to build increasingly abstract joint representations [25].

Table 2 conceptualizes the functional roles of cross-modal attention mechanisms and their correspondence to clinically meaningful interactions in ARDS diagnosis.

Table 2. Cross-Modal Attention Functions and Clinical Correspondence in ARDS Detection

Attention Interaction	Query Source	Key/Value Source	Clinical Interpretation	Diagnostic Correlation
Image → Text	CXR patches	Clinical note tokens	Aligns radiographic opacities with documented respiratory findings	Eventual diagnosis
Text → Image	Clinical phrases	Image regions	Links descriptions (e.g., "bilateral infiltrates") to spatial features	Anatomical region

Lab → Image	Lab embeddings	Image patches	Associates hypoxemia trends with visual lung pathology	S
Image → Lab	Image features	Lab sequences	Connects severity of opacities with oxygenation decline	s
Text → Lab	Clinical notes	Lab values	Relates narrative observations to quantitative trends	in
Lab → Text	Lab sequences	Clinical tokens	Anchors lab abnormalities to documented symptoms	c
Multi-Head Fusion	All modalities	All modalities	Simultaneous multi-perspective reasoning	hc rep

Multi-head attention in these layers captures diverse types of cross-modal relationships relevant to ARDS diagnosis [26]. Residual connections and layer normalization maintain training stability across the deep architecture [8]. The design draws from established transformer principles adapted for healthcare multimodal data [10]. Conceptual application here supports nuanced integration beyond simple concatenation [9].

Positional encoding across modalities

Positional encoding is adapted differently for each modality to reflect their inherent structures within the transformer [27]. For laboratory values, temporal positional encodings indicate the sequence of measurements over time [10]. Image patches receive spatial positional encodings based on their location in the chest X-ray [8]. Textual tokens utilize standard sequential positional embeddings from the language model [5].

To align modalities, additional learnable embeddings may indicate modality type or relative timing across inputs [25].

This cross-modal positional strategy ensures the transformer understands correspondences, such as a lab trend coinciding with a note entry [19]. The encoding scheme supports effective attention computation in the joint space [23]. Overall, it enables coherent processing of heterogeneous data streams [9].

Classification head

The classification head aggregates representations using a CLS token or mean pooling from the final transformer layer outputs [8]. Dense layers with appropriate activation functions process the fused embedding to produce logits for ARDS classification [2]. The head outputs probabilities corresponding to ARDS presence or severity categories conceptually [10]. Dropout and regularization techniques are incorporated to prevent overfitting in this component [7].

This design allows the model to generate a single probability score for ARDS detection based on the integrated multimodal evidence [3]. The head can be extended for multi-task learning if auxiliary predictions are desired in future conceptual expansions [19]. Softmax or sigmoid activation ensures interpretable outputs aligned with clinical decision needs [2]. Integration with the transformer backbone completes the end-to-end architecture for practical use [15].

Cross-Modal Attention

Attention mechanism design

The attention mechanism in the framework employs a cross-modal design where queries derived from one modality attend to keys and values from the others, enabling dynamic alignment of features across chest X-ray patches, clinical note tokens, and laboratory embeddings [9]. This query-key-value structure allows the model to weigh the relevance of specific image regions to textual descriptions of respiratory status or laboratory trends indicative of hypoxemia [24]. Multi-head attention further captures diverse relational patterns, such as linking bilateral opacities to mentions of ventilator adjustments in notes [23]. The mechanism operates within stacked transformer layers to progressively refine joint representations suitable for ARDS detection [25].

Scalability is maintained through efficient attention computations that avoid quadratic complexity bottlenecks in multimodal sequences [26]. Learnable parameters in the

cross-attention modules adapt to the unique characteristics of healthcare data, ensuring robust information exchange [8]. This design draws from established cross-modal techniques to foster synergistic learning without modality dominance [9]. The overall architecture supports flexible attention flows that mirror clinical reasoning processes in critical care [24].

Interpretability via attention

Interpretability arises naturally from the attention weights, which highlight correspondences between image regions showing opacities and specific phrases in clinical notes describing auscultation findings or laboratory shifts in oxygenation [24]. Clinicians can visualize these weights to understand how the model links a particular PaO2/FiO2 trend to radiographic evidence, thereby building trust in the ARDS classification [25]. Such attention maps provide post-hoc explanations aligned with the Berlin criteria components [23]. The framework conceptualizes attention as a transparent layer that augments rather than replaces human oversight [26].

By examining cross-modal attention patterns, potential biases in individual modalities become evident, allowing for targeted refinements in the conceptual design [9]. This interpretability feature distinguishes the transformer from black-box unimodal systems commonly used in imaging or NLP tasks [28]. The approach facilitates collaborative validation between AI outputs and expert judgment in ICU workflows [29]. Ultimately, attention-driven explanations enhance the framework's utility for educational and clinical decision-support purposes [24].

Fusion Strategies

Early vs late vs intermediate fusion

Early fusion concatenates raw or low-level features from all modalities at the input stage, which risks overwhelming the model with unaligned heterogeneous data and increases sensitivity to missing inputs [7]. Late fusion processes each modality independently before combining high-level decisions, yet this strategy often misses subtle cross-modal interactions essential for satisfying the multifaceted Berlin definition [20]. Intermediate fusion, positioned between these extremes, balances preservation of modality-specific information with joint learning and is therefore favored in complex healthcare applications [19]. Trade-offs in computational efficiency and representational power guide

the selection, with intermediate methods showing conceptual superiority for integrating imaging, text, and numerical streams [28].

The preference for intermediate fusion stems from its ability to leverage transformer layers for adaptive combination without premature loss of granular details [22]. In ARDS contexts, this avoids the pitfalls of early noise accumulation or late decision silos observed in other multimodal medical tasks [29]. Conceptual evaluations of these strategies underscore the need for fusion tailored to the complementary nature of chest X-rays, notes, and labs [7]. The framework adopts an intermediate path to optimize information synthesis [19].

Table 3 presents a theoretical comparison of multimodal fusion strategies, emphasizing the conceptual advantages of intermediate and attention-based approaches for ARDS detection.

Table 3. Theoretical Trade-Offs Across Multimodal Fusion Strategies in Critical Care AI Systems

Fusion Strategy	Stage of Integration	Strengths	Limitations
Early Fusion	Input-level	Full feature availability	No amplification of misalignment
Late Fusion	Decision-level	Modular and interpretable	Misses complex interactions
Intermediate Fusion (Proposed)	Representation-level	Balanced integration and interaction	High computational cost
Attention-Based Fusion	Dynamic within transformer	Adaptive weighting of modalities	Complex optimization
Concatenation-Based Fusion	Static feature merging	Simple implementation	Limited interaction modeling

Gated Fusion	Weighted modality control	Handles modality importance	Req tunin super
--------------	---------------------------	-----------------------------	-----------------

Proposed fusion approach

The proposed fusion approach concatenates tokenized representations from the modality-specific encoders into a unified sequence before applying joint self-attention within the cross-modal transformer layers [8]. This token-level integration allows the model to perform self-attention across the entire multimodal input, capturing higher-order dependencies such as the alignment of laboratory trends with radiographic opacities and narrative descriptions [10]. Residual connections ensure that original modality signals remain accessible throughout the fusion process [19]. The design culminates in a compact fused embedding that encodes the holistic patient state for ARDS assessment [9].

By relying on joint self-attention rather than explicit gating mechanisms, the approach maintains end-to-end differentiability and conceptual simplicity [25]. This strategy aligns with the framework's attention-driven principles, promoting seamless exchange of complementary information [23]. The resulting fusion supports robust classification even when modalities exhibit partial misalignment in timing or quality [7]. Overall, the method advances multimodal learning by treating disparate data streams as a cohesive clinical narrative [8].

Training Considerations

Data requirements

Training the multimodal transformer necessitates large-scale paired datasets comprising contemporaneous chest X-rays, clinical notes, and laboratory values from ICU encounters to mirror real-world ARDS scenarios [15]. Data augmentation techniques, such as image transformations for radiographs and synonym replacement for notes, can expand effective dataset size while preserving clinical fidelity [2]. The framework assumes access to labeled outcomes based on Berlin criteria to enable supervised learning across modalities [7]. Similar multimodal phenotypic studies in other complex syndromes illustrate the value of diverse, high-quality paired data for robust generalization [29].

Curated repositories with temporal alignment between inputs are essential to support the cross-modal attention mechanisms [19]. Strategies for balancing class distributions reflecting ARDS prevalence further enhance training stability [3]. The conceptual design accounts for variability in data collection practices across institutions to promote broader applicability [20]. These requirements form the foundation for developing a reliable ARDS detection tool [2].

Loss functions

The primary loss function employs binary cross-entropy to optimize the model's output probability for ARDS presence or severity aligned with the Berlin definition [2]. Auxiliary losses for modality alignment, such as contrastive terms that encourage correspondence between image and text embeddings, reinforce cross-modal coherence during training [19]. These additional objectives prevent any single modality from dominating the fused representation [7]. Weighted combinations of the losses allow emphasis on clinical priorities like sensitivity to hypoxemia indicators [10].

The framework conceptualizes these loss components as mutually reinforcing to achieve both accurate classification and interpretable feature interactions [9]. Regularization terms, including those for attention sparsity, can further improve generalization without introducing performance metrics [25]. Such a multi-objective setup supports the integration of complementary data sources in a balanced manner [8]. Overall, the loss design ensures the model learns clinically meaningful patterns across modalities [3].

Handling missing modalities

Handling missing modalities during training incorporates masking techniques that temporarily exclude absent inputs while maintaining the joint transformer structure [19]. Dropout applied at the modality level simulates real-world incompleteness, enabling the model to rely on available data streams without retraining [7]. Inference procedures similarly accommodate partial inputs by adjusting attention computations dynamically [20]. This robustness is critical in ICU environments where laboratory panels or imaging may not always be immediately available [3].

The conceptual approach avoids imputation biases by learning directly from masked representations, preserving the integrity of cross-modal attention [19]. Auxiliary

reconstruction objectives can encourage the model to infer missing modality contributions from the others [10]. Such strategies enhance the framework's practical deployability in heterogeneous clinical settings [2]. By design, the architecture gracefully degrades performance only when essential information is absent rather than failing outright [7].

Evaluation Strategy

Metrics for multimodal models

Evaluation of the multimodal transformer relies on standard classification metrics including area under the receiver operating characteristic curve to assess discrimination for ARDS detection [2]. Sensitivity and specificity provide insights into the model's ability to identify true positive cases while minimizing false alarms, particularly important for the Berlin criteria's oxygenation and imaging components [7]. Conceptual comparisons against unimodal baselines highlight the added value of integrating chest X-rays, notes, and laboratories [20]. These metrics emphasize clinical utility through balanced performance across ARDS severity levels [3].

Additional focus on calibration ensures that predicted probabilities align with observed diagnostic outcomes in critical care [19]. The framework advocates for holistic assessment that accounts for explainability alongside accuracy [24]. Such metrics collectively validate the synergistic benefits of the proposed architecture [8]. This evaluation lens supports iterative conceptual refinement of the multimodal design [2].

Validation protocols

Validation protocols incorporate k-fold cross-validation on diverse ICU cohorts to confirm internal consistency of the multimodal representations [3]. External validation on independent datasets from varied institutions tests generalizability beyond the training distribution [2]. Ablation studies systematically remove individual modalities to quantify their unique contributions to ARDS detection [8]. These protocols ensure the framework's robustness across different electronic health record systems [20].

Temporal validation splits further simulate prospective deployment by respecting the chronological order of patient data [19]. The conceptual strategy includes sensitivity analyses for varying levels of data completeness to mirror

clinical realities [15]. Such comprehensive validation strengthens confidence in the transformer's applicability for objective ARDS assessment [3]. Overall, the protocols align with best practices for trustworthy artificial intelligence in healthcare [2].

Conclusion

The multimodal transformer architecture integrates chest X-ray images, clinical notes, and laboratory values through modality-specific encoders, cross-modal attention layers, and intermediate fusion to enable comprehensive ARDS detection aligned with the Berlin definition. This conceptual framework synthesizes complementary data streams into a unified representation that addresses the limitations of traditional diagnostic approaches. By leveraging transformer principles, the design captures intricate relationships across imaging, textual, and numerical modalities. The proposed structure offers a scalable blueprint for advancing artificial intelligence applications in critical care diagnostics.

Key advantages include the seamless integration of complementary clinical information and the inherent explainability provided by attention mechanisms. Cross-modal interactions reduce diagnostic subjectivity while preserving the interpretability needed for clinician acceptance. The framework's attention-driven fusion surpasses unimodal methods by learning holistic patient representations. These strengths position the architecture as a promising conceptual advance for objective syndrome detection in intensive care.

Limitations encompass the requirement for large paired multimodal datasets, substantial computational resources for transformer training, and the ongoing challenge of acquiring high-quality labeled examples based on Berlin criteria. Data availability in diverse ICU settings may constrain immediate scalability, and handling extreme missingness patterns requires further conceptual exploration. Computational costs associated with cross-modal attention could limit deployment in resource-constrained environments without optimization. Despite these considerations, the framework remains adaptable through targeted refinements in future developments.

Implementation of this conceptual framework on public ICU datasets such as MIMIC-CXR and eICU is encouraged to validate its potential and accelerate translation into clinical

workflows. Collaborative efforts among AI researchers, intensivists, and data scientists can refine the architecture for real-world integration. Such initiatives will contribute to more timely and accurate ARDS detection, ultimately improving patient outcomes in critical care. The proposed multimodal transformer represents a forward-looking step toward data-driven, explainable artificial intelligence in healthcare systems.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 10 May 2021 Revised: 06 Jul 2021 Accepted: 22 Aug 2021
Published online: 20 January 2022

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Sjoding MW, Hofer TP, Co I, Courey A, Cooke CR, Iwashyna TJ. Interobserver reliability of the Berlin ARDS definition and strategies to improve the reliability of ARDS diagnosis. *Chest*. 2018;153(2):361-7.
<https://doi.org/10.1016/j.chest.2017.11.037>.
- Sayed M, Riaño D, Villar J. Novel criteria to classify ARDS severity using a machine learning approach. *Crit Care*. 2021;25(1):150.
<https://doi.org/10.1186/s13054-021-03563-w>.
- Li H, Odeyemi YE, Weister TJ, Liu C, Chalmers SJ, Lal A, et al. Rule-based cohort definitions for acute respiratory distress syndrome: a computable phenotyping strategy based on the Berlin definition. *Crit Care Explor*. 2021;3(6):e0451.
<https://doi.org/10.1097/CCE.0000000000000451>.
- Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest X-ray classification. *Sci Rep*. 2019;9(1):6381.
<https://doi.org/10.1038/s41598-019-42294-8>.
- Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. In: *Proc 2nd Clin Nat Lang Process Workshop*. Stroudsburg (PA): Association for Computational Linguistics; 2019. p. 72-78.
- Hung CY, Lin CH, Chang CS, Li JL, Lee CC. Predicting gastrointestinal bleeding events from multimodal in-hospital electronic health records using deep fusion networks. In: *2019 41st Annu Int Conf IEEE Eng Med Biol Soc (EMBC)*. Piscataway (NJ): IEEE; 2019. p. 2447-50.
<https://doi.org/10.1109/EMBC.2019.8857402>.
- Mohsen F, Ali H, El Hajj N, Shah Z. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Sci Rep*. 2022;12(1):17981.
<https://doi.org/10.1038/s41598-022-22514-4>.
- Tang W, He F, Liu Y, Duan Y. MATR: multimodal medical image fusion via multiscale adaptive transformer. *IEEE Trans Image Process*. 2022;31:5134-49.
<https://doi.org/10.1109/TIP.2022.3195278>.
- Song X, Chao H, Xu X, Guo H, Xu S, Turkbey B, et al. Cross-modal attention for multi-modal image registration. *Med Image Anal*. 2022;82:102612.
<https://doi.org/10.1016/j.media.2022.102612>.
- Jana S, Dasgupta T, Dey L. Predicting medical events and ICU requirements using a multimodal multiobjective transformer

network. *Exp Biol Med* (Maywood). 2022;247(22):1988-2002.
<https://doi.org/10.1177/15353702221121775>.

Majdi MS, Salman KN, Morris MF, Merchant NC, Rodriguez JJ. Deep learning classification of chest X-ray images. In: 2020 IEEE Southwest Symp Image Anal Interpr (SSIAI). Piscataway (NJ): IEEE; 2020. p. 116-9.
<https://doi.org/10.1109/SSIAI49293.2020.9094615>.

Khan E, Rehman MZ, Ahmed F, Alfouzan FA, Alzahrani NM, Ahmad J. Chest X-ray classification for the detection of COVID-19 using deep learning techniques. *Sensors* (Basel). 2022;22(3):1211.
<https://doi.org/10.3390/s22031211>.

Shelke A, Inamdar M, Shah V, Tiwari A, Hussain A, Chafekar T, et al. Chest X-ray classification using deep learning for automated COVID-19 screening. *SN Comput Sci.* 2021;2(4):300.
<https://doi.org/10.1007/s42979-021-00695-8>.

Hussain E, Hasan M, Rahman MA, Lee I, Tamanna T, Parvez MZ. CoroDet: a deep learning based classification for COVID-19 detection using chest X-ray images. *Chaos Solitons Fractals.* 2021;142:110495.
<https://doi.org/10.1016/j.chaos.2020.110495>.

Pawar Y, Henriksson A, Hedberg P, Naucier P. Leveraging ClinicalBERT in multimodal mortality prediction models for COVID-19. In: 2022 IEEE 35th Int Symp Comput Based Med Syst (CBMS). Piscataway (NJ): IEEE; 2022. p. 199-204.
<https://doi.org/10.1109/CBMS55023.2022.00045>.

Kalusivalingam AK, Sharma A, Patel N, Singh V. Leveraging BERT and LSTM for enhanced natural language processing in clinical data analysis. *Int J AI ML.* 2021;2(3):1-9.

Roy A, Pan S. Incorporating medical knowledge in BERT for clinical relation extraction. In: Proc 2021 Conf Empir Methods Nat Lang Process. Stroudsburg (PA): Association for Computational Linguistics; 2021. p. 5357-66.

Lamproudis A, Henriksson A, Dalianis H. Developing a clinical language model for Swedish: continued pretraining of generic BERT with in-domain data. In: Proc Int Conf Recent Adv Nat Lang Process (RANLP 2021). Stroudsburg (PA): Association for Computational Linguistics; 2021. p. 790-7.
https://doi.org/10.26615/978-954-452-072-4_089.

Zhang C, Chu X, Ma L, Zhu Y, Wang Y, Wang J, et al. M3Care: learning with missing modalities in multimodal healthcare data. In: Proc 28th ACM SIGKDD Conf Knowl Discov Data Min. New York (NY): ACM; 2022. p. 2418-28.
<https://doi.org/10.1145/3534678.3539397>.

Lopez K, Fodeh SJ, Allam A, Brandt CA, Krauthammer M. Reducing annotation burden through multimodal learning. *Front Big Data.* 2020;3:19.
<https://doi.org/10.3389/fdata.2020.00019>.

Sun Q, Fang N, Liu Z, Zhao L, Wen Y, Lin H. HybridCTrm: bridging CNN and transformer for multimodal brain image segmentation. *J Healthc Eng.* 2021;2021:7467261.

Zhang Y, Deng Y, Zhou Z, Zhang X, Jiao P, Zhao Z. Multimodal learning for fetal distress diagnosis using a multimodal medical information fusion framework. *Front Physiol.* 2022;13:1021400.
<https://doi.org/10.3389/fphys.2022.1021400>.

Zhang Y, Ou W, Shi Y, Deng J, You X, Wang A. Deep medical cross-modal attention hashing. *World Wide Web.* 2022;25(4):1519-36.
<https://doi.org/10.1007/s11280-021-00973-0>.

Shi T, Jiang H, Zheng B. C2MA-Net: cross-modal cross-attention network for acute ischemic stroke lesion segmentation based on CT perfusion scans. *IEEE Trans Biomed Eng.* 2022;69(1):108-18.
<https://doi.org/10.1109/TBME.2021.3086210>.

Song X, Zhang X, Ji J, Liu Y, Wei P. Cross-modal contrastive attention model for medical report generation. In: Proc 29th Int Conf Comput Linguistics. Stroudsburg (PA): Association for Computational Linguistics; 2022. p. 2388-97.

Zhou Z, Guo X, Yang W, Shi Y, Zhou L, Wang L, et al. Cross-modal attention-guided convolutional network for multi-modal cardiac segmentation. In: Int Workshop Mach Learn Med Imaging. Cham: Springer; 2019. p. 601-10.
https://doi.org/10.1007/978-3-030-32692-0_69.

Yan S, Wang C, Chen W, Lyu J. Swin transformer-based GAN for multi-modal medical image translation. *Front Oncol.* 2022;12:942511.
<https://doi.org/10.3389/fonc.2022.942511>.

Koivunen M, Saranto K. Nursing professionals' experiences of the facilitators and barriers to the use of telehealth applications: a systematic review of qualitative studies. *Scand J Caring Sci.* 2018;32(1):24-44.
<https://doi.org/10.1111/scs.12445>.

Markello RD, Shafiei G, Tremblay C, Postuma RB, Dagher A, Misić B. Multimodal phenotypic axes of Parkinson's disease. *npj Parkinsons Dis.* 2021;7(1):6.
<https://doi.org/10.1038/s41531-020-00144-6>.