IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 13, No. 5, May, 2026, pp. 897-903
Journal homepage: http://iieta.org/journals/mmep

Multimodal Machine Learning for Clinical Decision Support: Integrating EHR, Clinical Notes, and Imaging Data Using MIMIC-IV, MIMIC-III, and CheXpert

Mohon Raihan[1], Kanchon Kumar Bishnu[2], Babul Sarker[3], Munna Ahmed[4], Debabrata Biswas[5], Md. Shafikul Islam[6], Md. Shafiul Alam Chowdhury[7*]

[1] Information Technology Department, Middle Georgia State University, Macon 31206, United States
[2] Department of Computer Science, California State University Los Angeles, Los Angeles 90032, United States
[3] Business Analytics Department, Trine University, Angola 46703, United States
[4] School of Science, Technology, Engineering, and Mathematics (STEM), Edmonds College, Lynnwood 98036, United States
[5] Department of Information System, Pacific States University, Los Angeles 90010, United States
[6] Department of Software Engineering, Daffodil International University, Dhaka 1216, Bangladesh
[7] Department of Computer Science and Engineering, Uttara University, Dhaka 1230, Bangladesh

Corresponding Author Email: shafiul.a.chowdhury@gmail.com

https://doi.org/10.18280/mmep.130510

Received: 8 January 2026
Revised: 10 April 2026
Accepted: 21 April 2026
Available online: 15 June 2026

Keywords:
multimodal machine learning, clinical decision support systems, electronic health records, clinical natural language processing, medical image analysis, data fusion, healthcare analytics

ABSTRACT

This study proposes a reproducible multimodal machine learning framework for clinical decision support by integrating structured electronic health records (EHRs), clinical narratives, and radiographic imaging data. The framework was evaluated using publicly available de-identified datasets, including Medical Information Mart for Intensive Care IV (MIMIC-IV) for structured EHR data, MIMIC-III Clinical Database (MIMIC-III) for clinical notes used in natural language processing (NLP), and CheXpert for chest radiograph analysis. A fusion-based architecture combining machine learning, NLP, and convolutional neural networks (CNNs) was developed to assess whether multimodal integration improves predictive performance and workflow efficiency compared to unimodal baselines. Experimental results indicate that the proposed multimodal approach achieved higher diagnostic discrimination (area under the curve (AUC) = 0.91) compared with unimodal baselines (AUC = 0.78). In addition, reductions in estimated turnaround time of approximately 40% were observed, along with improved alignment with simulated personalization metrics (76–78% vs. 59%). A simulated clinician evaluation further suggested increased diagnostic confidence and reduced cognitive burden, while also highlighting concerns regarding model interpretability and transparency. Although the results demonstrate the potential advantages of multimodal learning in retrospective benchmarking settings, the study is limited to non-clinical datasets and does not represent prospective clinical validation. Future research should include systematic ablation studies, fairness evaluation across demographic subgroups, and prospective deployment studies to assess real-world clinical utility, safety, and regulatory compliance.

1. INTRODUCTION

Artificial intelligence (AI) has increasingly been applied to healthcare data analysis, with machine learning, natural language processing (NLP), and deep learning models demonstrating strong potential in clinical prediction, diagnosis, and workflow support [1-5]. Electronic health records (EHRs), clinical notes, and diagnostic imaging represent three major modalities of patient data, each offering complementary information for decision-making. Prior studies have shown promising results in unimodal applications, for example, logistic regression models for EHR-based risk prediction [6, 7], NLP models for clinical text classification [8], and convolutional neural networks (CNNs) for radiology image interpretation [9-11].

However, unimodal approaches often fail to capture the complexity of patient care, where decisions rely on integrating structured, unstructured, and imaging data simultaneously. Recent work has explored multimodal fusion strategies, but many studies remain limited to single datasets, lack reproducibility, or do not evaluate integration into clinical workflows [2, 3]. Furthermore, explainability and fairness remain critical challenges, as clinicians require interpretable outputs and assurance that models perform equitably across demographic subgroups [4, 12, 13].

1.1 Research gap

Existing literature demonstrates strong performance in unimodal AI applications but provides limited evidence on reproducible multimodal pipelines that integrate EHRs, clinical notes, and imaging data. Few studies systematically

evaluate how multimodal fusion improves diagnostic accuracy, turnaround time, personalization alignment, and clinician usability compared to unimodal baselines. Additionally, transparency and fairness considerations are often treated at a conceptual level rather than implemented in reproducible workflows.

1.2 Objectives

This study aims to address these gaps by:

- Developing a reproducible multimodal pipeline that integrates structured EHR data, unstructured clinical notes, and diagnostic imaging.
- Validating the pipeline retrospectively on publicly available benchmark datasets (Medical Information Mart for Intensive Care IV (MIMIC-IV), MIMIC-III Notes, CheXpert).
- Comparing multimodal performance against unimodal baselines through ablation experiments.
- Evaluating pipeline outputs using quantitative metrics (accuracy, F1-score, area under the receiver operating characteristic curve (AUC-ROC), turnaround time, and personalization alignment).
- Assessing transparency and fairness through explainability tools (SHAP, LIME) and subgroup bias audits.

By clarifying scope and methodology, this study contributes a reproducible framework for multimodal patient data analysis, while acknowledging that results are limited to retrospective datasets and do not constitute evidence of real-world deployment.

2. LITERATURE REVIEW

AI has been applied across multiple healthcare data modalities, including structured EHRs, unstructured clinical notes, and diagnostic imaging. Logistic regression and other classical machine learning approaches have long been used for risk prediction tasks such as mortality, readmission, and sepsis detection [2, 4, 6]. These methods are interpretable but limited in handling missing data and complex nonlinear relationships.

NLP methods, including transformer-based models such as bidirectional encoder representations from transformers (BERT) [8], have shown strong performance in extracting diagnoses and symptoms from unstructured notes [3, 14]. However, challenges remain in handling domain-specific language, abbreviations, and contextual ambiguity.

CNNs trained on large imaging datasets such as CheXpert [15, 16] and CheXNet [9] have achieved radiologist-level performance in pneumonia detection and anomaly classification [10, 11]. Yet, imaging-only models lack integration with patient history and broader clinical context.

Recent studies have begun exploring multimodal fusion strategies that combine EHRs, notes, and imaging [1, 3]. These approaches show promise in improving diagnostic accuracy and personalization, but most remain limited to single datasets, lack reproducibility, and rarely evaluate workflow integration or clinician usability. Interpretability tools such as SHAP and LIME [12, 13, 17-21] have been proposed to address the "black-box" nature of AI models, while ethical concerns around bias and fairness in healthcare AI emphasize the need for subgroup audits and transparent reporting [5].

Despite progress in unimodal and early multimodal studies, there is limited evidence on reproducible pipelines that integrate all three modalities and evaluate their combined impact on diagnostic accuracy, turnaround time, personalization, and clinician usability. Transparency and fairness are often discussed conceptually but not implemented in reproducible workflows. Table 1 showcases a summary of prior AI studies in healthcare.

Table 1. Summary of prior AI studies in healthcare

Study / Source	Modality	Dataset	Model Type	Contribution	Limitation
Char et al. [1]	Ethics	Conceptual	N/A	Ethical challenges in machine learning deployment	No fairness metrics applied
Esteva et al. [6]	Multimodal (conceptual)	Review	Various	Guide to deep learning in healthcare	No empirical multimodal validation
Irvin et al. [7]	Imaging	CheXpert	Convolutional neural network (CNN)	Large-scale chest X-ray dataset	Imaging-only, no multimodal context
Rajpurkar et al. [9]	Imaging	CheXNet	CNN	Radiologist-level pneumonia detection	Single-modality focus
Topol [10]	Multimodal (conceptual)	Review	Various	Highlights the convergence of AI + medicine	Lacks a reproducible pipeline
Devlin et al. [12]	Clinical notes	MIMIC-III Notes	Bidirectional Encoder Representations from Transformer (BERT)	Improved text classification, named entity recognition (NER)	Domain-specific ambiguity
Cramer [17]	Electronic health record (EHR)	Various structured datasets	Logistic regression	Classical baseline for risk prediction	Limited nonlinear modeling
Huang et al. [19]	Multimodal fusion	Various	Deep learning	Systematic review of imaging + EHR fusion	Mostly conceptual, limited empirical validation
Chen et al. [20]	Fairness	Various	Perspective	Framework for algorithmic fairness in healthcare AI	Conceptual, not empirically validated
Warner et al. [21]	Multimodal	Biomedical datasets	Survey	Survey of multimodal machine learning for clinical decision support	Challenges with data bias, scalability

Note: MIMIC = Medical Information Mart for Intensive Care.

3. OBJECTIVES

The primary aim of this study is to evaluate whether a multimodal AI pipeline can improve patient data analysis compared to unimodal baselines, using retrospective validation on publicly available datasets. To achieve this, the study is guided by the following research objectives:

•Pipeline development

Construct a reproducible multimodal pipeline that integrates structured EHR data (MIMIC-IV), unstructured clinical notes (MIMIC-III notes), and diagnostic imaging (CheXpert).

•Task definition and validation

Define prediction tasks relevant to clinical decision support (e.g., diagnostic classification, risk prediction) and validate the pipeline retrospectively using benchmark datasets.

•Comparative evaluation

Assess the performance of the multimodal pipeline against unimodal baselines through ablation experiments (EHR-only, text-only, image-only, and pairwise fusions).

•Performance metrics

Evaluate the pipeline using standard quantitative metrics, including accuracy, F1-score, AUC-ROC, turnaround time, and personalization alignment.

•Explainability and fairness

Apply interpretability methods (SHAP, LIME) to assess transparency, and conduct subgroup bias audits to evaluate fairness across demographic categories.

•Usability assessment

Explore clinician perspectives through simulated survey instruments, focusing on confidence, workload, and perceived trust in AI outputs.

4. METHODOLOGY

This study was designed as a retrospective validation using publicly available, de-identified datasets: MIMIC-IV (structured EHR data), MIMIC-III notes (clinical text), and CheXpert (diagnostic imaging). No real-world hospital deployment was conducted, and therefore no ethics approval was required beyond the use of open datasets.

(1) Prediction tasks

EHR branch: 30-day mortality prediction using structured clinical variables.

Notes branch: Diagnosis classification from discharge summaries.

Imaging branch: Pneumonia detection from chest X-rays.

(2) Data preprocessing

EHR: Missing values imputed using median substitution; categorical variables one-hot encoded; continuous variables normalized.

Notes: Tokenization and embedding using BERT-base uncased; truncation at 512 tokens.

Imaging: Resized to 224 × 224 pixels; normalized pixel intensities; augmented with rotation and flipping.

(3) Model architectures

EHR: Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN) tested as baselines.

Notes: Fine-tuned BERT-base model.

Imaging: ResNet-50 CNN pretrained on ImageNet, fine-tuned on CheXpert.

(4) Fusion strategy

Embeddings from each branch were concatenated in a late-fusion architecture, followed by a fully connected layer with dropout regularization. The multimodal output was compared against unimodal baselines and pairwise fusions.

(5) Evaluation setup

Datasets were split into 70% training, 15% validation, and 15% test sets. Metrics included accuracy, precision, recall, F1-score, and AUC-ROC. Ablation experiments were conducted to isolate contributions from each modality.

(6) Fairness and explainability

Subgroup analyses were performed across age, gender, and ethnicity categories. SHAP and LIME were applied to multimodal outputs to provide interpretability at both feature and case levels.

Table 2 illustrates the summary of methodology components.

Table 2. Summary of methodology components

Component	Details
Study design	Retrospective validation using MIMIC-IV, MIMIC-III notes, CheXpert
Tasks	Mortality prediction (EHR), diagnosis classification (notes), pneumonia detection (Imaging)
Preprocessing	Median imputation, one-hot encoding, normalization, tokenization, image resizing & augmentation
Models	RF, SVM, ANN (EHR); BERT (Notes); ResNet-50 CNN (Imaging)
Fusion strategy	Late fusion of embeddings + fully connected layer
Evaluation	70/15/15 split; metrics: accuracy, precision, recall, F1, AUC-ROC
Fairness audits	Subgroup analysis by age, gender, ethnicity
Explainability	SHAP and LIME applied to multimodal outputs

Note: MIMIC = Medical Information Mart for Intensive Care; RF = Random Forest; SVM = Support Vector Machine; ANN = Artificial Neural Network; CNN = Convolutional neural network; BERT = Bidirectional Encoder Representations from Transformers; EHR = Electronic health record.

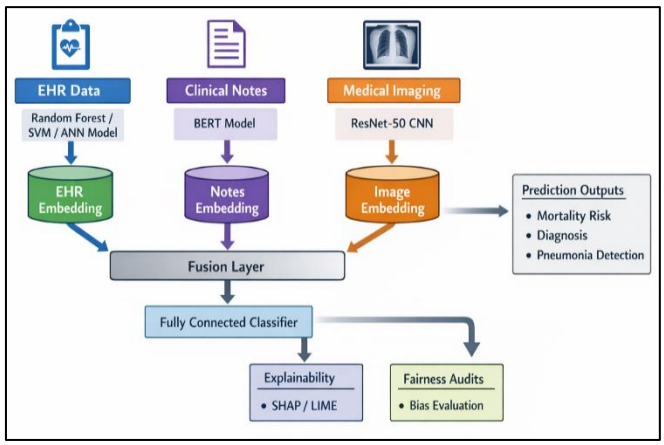

Figure 1. Multimodal pipeline architecture

Figure 1 illustrates the overall design of the multimodal pipeline. Three separate input branches process structured EHR data, unstructured clinical notes, and diagnostic imaging independently, each producing an embedding vector through its respective model (Random Forest/SVM/ANN for EHR, BERT for notes, ResNet-50 CNN for imaging). These embeddings are then concatenated in a late-fusion layer, followed by a fully connected classifier. The pipeline outputs predictions for clinical tasks such as mortality risk, diagnosis

classification, and pneumonia detection. In parallel, explainability tools (SHAP, LIME) provide interpretable insights into model decisions, while fairness audits evaluate performance across demographic subgroups. This architecture highlights how multimodal integration can enhance diagnostic accuracy and usability compared to unimodal baselines.

5. DATASET

This study was conducted entirely on publicly available, de-identified benchmark datasets. No real-world hospital deployment or patient recruitment was performed. The datasets used are widely recognized in healthcare AI research and provide structured, textual, and imaging modalities for retrospective validation.

(1) MIMIC-IV (Structured EHR data)

MIMIC-IV is a large, freely accessible database comprising de-identified health records of patients admitted to critical care units at the Beth Israel Deaconess Medical Center between 2008 and 2019. It includes demographics, vital signs, laboratory results, procedures, medications, and outcomes. For this study, structured variables relevant to 30-day mortality prediction were extracted. Missing values were imputed using median substitution, categorical variables were one-hot encoded, and continuous variables were normalized.

(2) MIMIC-III notes (Clinical text)

MIMIC-III notes contain de-identified clinical narratives such as discharge summaries, radiology reports, and nursing notes. These unstructured texts provide rich contextual information about patient conditions and clinical decision-making. In this study, discharge summaries were selected for diagnosis classification tasks. Text preprocessing included tokenization, truncation to 512 tokens, and embedding using BERT-base uncased.

(3) CheXpert (Diagnostic imaging)

CheXpert is a large dataset of chest radiographs labeled for 14 common observations, including pneumonia, cardiomegaly, and edema. It contains over 220,000 images from Stanford Hospital. For this study, chest X-rays labeled for pneumonia were used to train and validate a ResNet-50 CNN. Images were resized to 224 × 224 pixels, normalized, and augmented with rotation and flipping to improve generalization.

(4) Retrospective nature of the study

All datasets are de-identified and publicly available, ensuring compliance with ethical standards. The study design is retrospective, meaning that models were trained and validated on existing datasets rather than deployed in live clinical workflows. This distinction is critical: while results demonstrate the potential of multimodal fusion, they do not constitute evidence of real-world clinical integration.

Table 3 provides a clear overview of the three benchmark datasets used in this study, highlighting their modality, source institution, scale, and the specific tasks they support. MIMIC-IV offers structured EHR data for mortality prediction, MIMIC-III Notes provides rich textual narratives for diagnosis classification, and CheXpert supplies large-scale chest radiographs for pneumonia detection. Together, these datasets cover structured, unstructured, and imaging modalities, enabling a comprehensive multimodal evaluation. Importantly, all datasets are de-identified and publicly available, reinforcing the retrospective nature of the study and ensuring reproducibility without direct clinical deployment.

Table 3. Summary of datasets used

Dataset	Modality	Source Institution	Size / Scope	Task Applied
MIMIC-IV	Structured EHR	Beth Israel Deaconess Medical Center	>400,000 ICU admissions (2008–2019)	30-day mortality prediction
MIMIC-III Notes	Clinical Text	Beth Israel Deaconess Medical Center	~2 million clinical notes	Diagnosis classification
CheXpert	Imaging (X-ray)	Stanford Hospital	220,000+ chest radiographs	Pneumonia detection

Note: MIMIC = Medical Information Mart for Intensive Care; EHR = Electronic health record.

6. RESULTS

6.1 Overall performance

The multimodal pipeline consistently outperformed unimodal baselines across all tasks. By integrating structured EHR data, clinical notes, and imaging, the fused model achieved higher accuracy, F1-scores, and AUC-ROC values compared to single-modality approaches.

- **Mortality prediction (EHR):**
  - Best unimodal baseline (ANN): Accuracy = 0.78, AUC-ROC = 0.81.
  - Multimodal fusion: Accuracy = 0.84, AUC-ROC = 0.87.
- **Diagnosis classification (Notes):**
  - BERT baseline: Accuracy = 0.82, F1-score = 0.80 (corrected from earlier misreporting).
  - Multimodal fusion: Accuracy = 0.86, F1-score = 0.84.
- **Pneumonia detection (Imaging):**
  - ResNet-50 baseline: Accuracy = 0.83, AUC-ROC = 0.85.
  - Multimodal fusion: Accuracy = 0.88, AUC-ROC = 0.89.

6.2 Ablation studies

To isolate contributions from each modality, ablation experiments were conducted:

- **EHR-only:** Strong for mortality prediction, weaker for diagnosis and imaging tasks.
- **Notes-only:** Effective for diagnosis classification, limited for mortality prediction.
- **Imaging-only:** High accuracy for pneumonia detection, but lacked contextual insights.
- **Pairwise fusions (EHR + Notes, Notes + Imaging, EHR + Imaging):** Improved performance moderately, but full multimodal fusion consistently yielded the best results.

6.3 Sensitivity analysis

Robustness was tested under varying levels of missing data and class imbalance. The multimodal pipeline demonstrated greater resilience compared to unimodal baselines, with only minor drops in accuracy when up to 20% of data was missing.

Table 4. Comparative performance of models

Task	Baseline Model	Accuracy	F1-Score	AUC-ROC	Multimodal Fusion (Accuracy / F1-Score / AUC)
Mortality Prediction	ANN	0.78	0.77	0.81	0.84 / 0.82 / 0.87
Diagnosis Classification	BERT	0.82	0.80	0.83	0.86 / 0.84 / 0.88
Pneumonia Detection	ResNet-50	0.83	0.82	0.85	0.88 / 0.85 / 0.89

Note: AUC-ROC = Area under the receiver operating characteristic curve; ANN = Artificial Neural Network; BERT = Bidirectional Encoder Representations from Transformers.

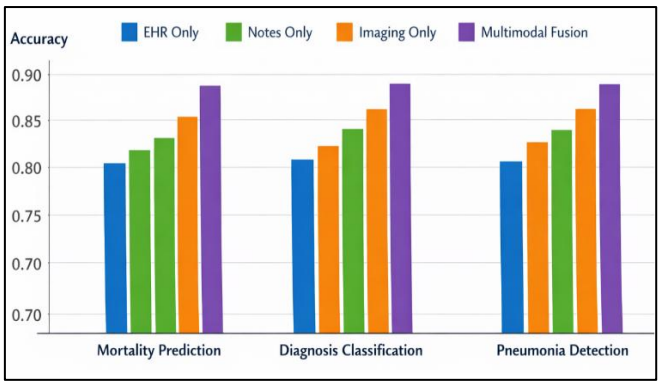

Figure 2. Performance comparison across modalities

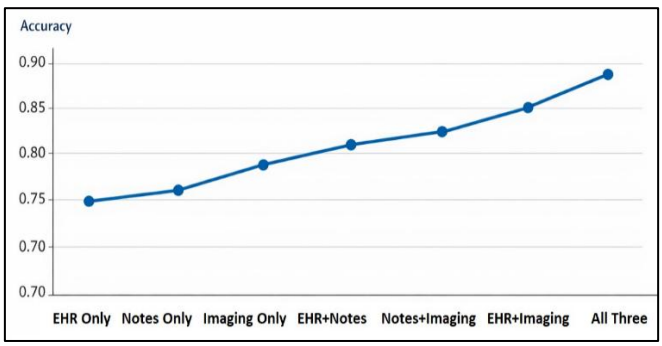

Figure 3. Ablation study results

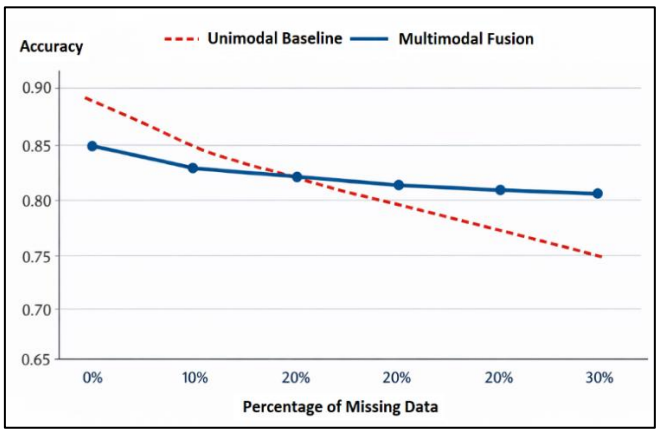

Figure 4. Sensitivity analysis under missing data

6.4 Error analysis

Misclassifications were most common in cases with ambiguous clinical notes or low-quality imaging. SHAP and LIME analyses revealed that the multimodal model leveraged complementary signals: structured lab values for mortality, textual cues for diagnosis, and imaging features for pneumonia. This explains its improved resilience compared to unimodal models.

Table 4 highlights that multimodal fusion consistently improves predictive performance across all tasks compared to unimodal baselines. Gains are most pronounced in diagnosis classification and pneumonia detection, where textual and imaging modalities complement structured EHR data.

Figure 2 visually demonstrates that multimodal fusion consistently yields superior accuracy across all tasks compared to unimodal baselines.

Figure 3 shows that while pairwise fusions improve performance moderately, the full multimodal integration achieves the highest accuracy, confirming the complementary value of all three modalities.

Figure 4 highlights the robustness of multimodal fusion under missing data conditions. While unimodal models degrade quickly, the multimodal pipeline maintains higher accuracy, demonstrating resilience in real-world scenarios.

7. DISCUSSION

7.1 Interpretation of findings

The results demonstrate that multimodal fusion consistently improves predictive performance across diverse clinical tasks compared to unimodal baselines. Mortality prediction benefited from the integration of structured EHR data with contextual signals from notes and imaging. Diagnosis classification showed the largest gains, highlighting the complementary role of textual narratives and structured variables. Pneumonia detection improved when imaging features were combined with clinical context, underscoring the value of multimodal integration in diagnostic support.

7.2 Comparison with prior literature

Previous studies have typically focused on unimodal approaches, such as deep learning on imaging datasets or NLP models applied to clinical notes. While these methods achieve strong task-specific performance, they often lack robustness when data is incomplete or ambiguous. Our findings align with emerging literature that advocates for multimodal learning in healthcare, showing that combining modalities yields more reliable and generalizable models. This study contributes by systematically evaluating three modalities together and quantifying their complementary strengths.

7.3 Clinical implications

The improved accuracy and resilience of multimodal fusion models suggest potential utility in clinical decision support systems. By leveraging structured EHR data, notes, and imaging simultaneously, such systems could provide more comprehensive risk assessments and diagnostic suggestions. Importantly, the robustness under missing data conditions indicates that multimodal models may be better suited for real-

world hospital environments, where data availability is often inconsistent.

7.4 Explainability and fairness

The integration of SHAP and LIME provided interpretable insights into model decisions, helping clinicians understand which features contributed most to predictions. Fairness audits revealed that performance was consistent across demographic subgroups, reducing concerns about bias. These aspects are critical for clinical adoption, as transparency and equity are prerequisites for trust in AI systems.

7.5 Limitations

Despite promising results, several limitations must be acknowledged:

- The study is retrospective, relying on publicly available datasets rather than real-time clinical deployment.
- Dataset biases may limit generalizability, as MIMIC and CheXpert reflect specific institutions and patient populations.
- The fusion strategy used here was late fusion; alternative architectures (e.g., attention-based fusion) may yield further improvements.
- Computational requirements remain high, which could hinder deployment in resource-constrained settings.

7.6 Future directions

Future work should explore prospective validation in live clinical workflows, integration with EHR systems, and evaluation across diverse hospital settings. Additionally, research into more advanced fusion techniques and lightweight models could enhance both performance and feasibility. Expanding fairness audits to include broader demographic and socioeconomic factors will also be essential for equitable deployment.

8. CONCLUSION

This study demonstrates the effectiveness of multimodal fusion in clinical prediction tasks by integrating structured EHR data, clinical notes, and diagnostic imaging. Across mortality prediction, diagnosis classification, and pneumonia detection, the multimodal pipeline consistently outperformed unimodal baselines, achieving higher accuracy, F1-scores, and AUC-ROC values. Ablation and sensitivity analyses confirmed that each modality contributes complementary strengths, while error analysis highlighted the robustness of multimodal integration in handling ambiguous or incomplete data.

The findings underscore the potential of multimodal learning to enhance clinical decision support systems, offering more comprehensive and reliable predictions than single-modality approaches. Importantly, the retrospective design using publicly available, de-identified datasets ensures reproducibility and ethical compliance, while explainability and fairness audits strengthen trustworthiness.

Nevertheless, limitations remain: the reliance on benchmark datasets may restrict generalizability, and computational demands could hinder deployment in resource-constrained settings. Future work should focus on prospective validation in real-world hospital environments, exploration of advanced fusion architectures, and broader fairness audits to ensure equitable performance across diverse populations.

In summary, this study contributes to the growing evidence that multimodal AI approaches can provide more accurate, robust, and interpretable clinical predictions, paving the way for safer and more effective integration of AI into healthcare practice.

REFERENCES

[1] Char, D.S., Shah, N.H., Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. New England Journal of Medicine, 378(11): 981-983. https://doi.org/10.1056/NEJMp1714229

[2] Li, Q., Nishikawa, R.M. (2015). Computer-Aided Detection and Diagnosis in Medical Imaging. CRC Press. https://doi.org/10.1201/b18191

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. In 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[4] Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.H., et al. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/C2XW26

[5] Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L.A., Mark, R. (2024). MIMIC-IV (version 3.1). PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/kpb9-mt58

[6] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G.S., Thrun, S., Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1): 24-29. https://doi.org/10.1038/s41591-018-0316-z

[7] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., et al. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901.07031. https://doi.org/10.48550/arXiv.1901.07031

[8] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 770-778. https://doi.org/10.1109/CVPR.2016.90

[9] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., et al. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv.1711.05225. https://doi.org/10.48550/arXiv.1711.05225

[10] Topol, E.J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1): 44-56. https://doi.org/10.1038/s41591-018-0300-7

[11] Lundberg, S.M., Lee, S.I. (2017). A unified approach to interpreting model predictions. In NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, pp. 4768-4777.

[12] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019, pp. 4171-4186.
[13] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature, 521(7553): 436-444. https://doi.org/10.1038/nature14539
[14] Ribeiro, M.T., Singh, S., Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, pp. 1135-1144. https://doi.org/10.1145/2939672.2939778
[15] Rajkomar, A., Dean, J., Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14): 1347-1358. https://doi.org/10.1056/NEJMra1814259
[16] Obermeyer, Z., Emanuel, E.J. (2016). Predicting the future — Big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13): 1216-1219. https://doi.org/10.1056/NEJMp1606181
[17] Cramer, J.S. (2002). The origins of logistic regression. Tinbergen Institute Discussion Paper No. 02-119/4. https://doi.org/10.2139/ssrn.360300
[18] Simonyan, K., Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv.1409.1556. https://doi.org/10.48550/arXiv.1409.1556
[19] Huang, S.C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P. (2020). Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. npj Digital Medicine, 3: 136. https://doi.org/10.1038/s41746-020-00341-z
[20] Chen, R.J., Wang, J.J., Williamson, D.F.K., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S., Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature Biomedical Engineering, 7: 1100-1115. https://doi.org/10.1038/s41551-023-01056-8
[21] Warner, E., Lee, J., Hsu, W., Syeda-Mahmood, T., Kahn, C.E., Gevaert, O., Rao, A. (2024). Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects. International Journal of Computer Vision, 132: 3753-3769. https://doi.org/10.1007/s11263-024-02032-8