STM JOURNALS

International Journal of
Bioinformatics and Computational Biology

ISSN: 2583-9977
Volume 4, Issue 1, 2026
January—June
DOI (Journal): 10.37591/IJBCB

https://journals.stmjournals.com/ijbcb

Review IJBCB

Combining Unstructured and Structured Clinical Data in a Hybrid Transformer Model to Enhance Cardiovascular Analytics and Clinical Decision-Making

Kalyani Neve[1,*], Padma Mishra[2], Karishma Chaudhari[1], I.D. Paul[3]

Abstract

Since cardiovascular disease (CVD) continues to be a major global cause of morbidity and mortality, early, and accurate risk prediction is essential for prompt intervention and individualized treatment. This study introduces a new hybrid transformer-based model that combines unstructured clinical narratives, structured data, and customized lifestyle characteristics. A comprehensive understanding of disease progression is made possible by the model's ability to capture contextual, temporal, and patient-specific insights through the use of transformer architectures and sophisticated natural language processing. Clinical interpretability and transparency are guaranteed by a special explainability module. Our method combines insights from deep learning, statistical modeling, and graph-based temporal embeddings to improve prediction accuracy and clinical relevance, building on significant advancements in time-aware LSTMs, hybrid modeling, self-attention transformers, and automated machine learning from previous works. Evaluating performance on reference datasets, the hybrid model outperformed both conventional and cutting-edge methods on benchmark datasets, obtaining a Precision@5 of 78.65%, an AUC of 0.86, and an F1-score of 0.76. Its exceptional performance across comorbid conditions and demographic groups demonstrates its generalizability and practicality. This hybrid framework, which seamlessly integrates various data modalities for actionable and equitable CVD risk prediction, is a prime example of the future of predictive analytics in healthcare and advances precision medicine.

Keywords: Attention mechanisms, electronic health records (EHR), hybrid transformer model, ICD, temporal data modeling

***Author for Correspondence**
Kalyani Neve
E-mail: kalyani.neve@raisoni.net

[1]Assistant Professor, Department of MCA, G H Raisoni College of Engineering and Management, Jalgaon, Maharashtra, India
[2]Associate Professor, Thakur Institute of Management Studies, Career Development & Research, Mumbai, Maharashtra, India
[3]Assistant Professor, Department of Mechanical Engineering, S.S.G.B, Bhusawal, Maharashtra, India

Received Date: December 29, 2025
Accepted Date: January 29, 2026
Published Date: January 29, 2026

Citation: Kalyani Neve, Padma Mishra, Karishma Chaudhari, I.D. Paul. Combining Unstructured and Structured Clinical Data in a Hybrid Transformer Model to Enhance Cardiovascular Analytics and Clinical Decision-Making. International Journal of Bioinformatics and Computational Biology. 2026; 4(1): 30–37p.

INTRODUCTION

The efficiency of clinical practice has been greatly improved by the quick development of data-driven approaches in healthcare, especially through Natural Language Processing (NLP) and predictive analytics. A vital source of patient data, electronic health records (EHRs) offer comprehensive insights into medical histories and diagnoses. The International Classification of Diseases (ICD) coding system is frequently used to organize the data entered into electronic health records (EHRs), enabling medical professionals to monitor and treat patients' conditions over time.

Early identification of possible future diseases is a major challenge in healthcare, especially in preventive medicine. In order to lessen the long-

term effects of diseases on patients and healthcare systems, preventive medicine focuses on identifying risk factors and taking action before the disease manifests [1]. By lessening the burden of upcoming illnesses, this strategy not only enhances patient outcomes but also maximizes the use of healthcare resources. New methods and competitions have been developed in response to the need for more accurate predictive models. CLEF's Intelligent Disease Progression Prediction (2022–2024) challenges international research groups to forecast the course of diseases using probabilistic, time-dependent approaches.

Unstructured clinical text from EHRs is becoming more and more popular, even though many predictive models mainly concentrate on integrating structured data such as socioeconomic factors, patient histories, and clinical visits. Rich, patient-specific information found in this unstructured data can greatly improve models' predictive accuracy. By automating the analysis of diagnosis timelines to forecast the most likely diagnoses in upcoming visits, our research seeks to close this gap, with a particular emphasis on cardiovascular disease (CVD). The "Next Diagnosis Prediction" task is essential for giving doctors the tools they need to make early predictions and enhance patient care. The chronological record of a patient's medical conditions, recorded at each visit using both structured ICD codes and unstructured EHR text, is known as the diagnosis timeline [2–4]. Through the analysis of these timelines, the quality of care can be improved by summarizing the patient's health trajectory and forecasting future health conditions. Even though diagnosis timelines are useful, the difficulty is that ICD codes vary widely, and patient data is not always readily available due to privacy concerns. This can limit the precision and depth of predictions.

Our research addresses these issues by presenting a hybrid model that combines unstructured EHR text with ICD-coded diagnosis timelines, improving future diagnosis prediction by leveraging both structured and unstructured data. The model combines predictions from a Clinical Longformer-based model that gathers information from clinical notes and a sequential model that processes structured diagnosis data using an ensemble approach [5–8].

Our research shows that, in comparison to conventional approaches that only use structured data, combining diagnosis history with unstructured clinical text greatly increases prediction accuracy [9, 10]. Furthermore, we suggest a data augmentation technique to expand the training dataset's size, allowing the model to produce more precise predictions even with a smaller amount of data. To ensure the model's efficacy in a range of patient populations, we test it across demographic groups to assess its robustness. Our model is a useful tool for clinicians who must make precise predictions based on a limited amount of patient history because it is effective even with short diagnosis timelines.

By giving physicians a more precise grasp of the risks of cardiovascular disease and facilitating better decision-making in patient care, this research seeks to develop a predictive tool that improves preventive healthcare (Table 1).

The underuse of unstructured data, such as clinical notes and discharge summaries, despite their abundance of patient-specific details, represents a substantial gap in the prediction of cardiovascular disease (CVD) risk [3]. ICD codes and other structured data offer crucial insights, but they frequently lack the contextual depth of unstructured text [11–15]. Innovative models could bridge this gap by using domain- specific natural language processing.

(NLP) frameworks or transformer-based architectures to extract useful information from unstructured EHR text, allowing for a more thorough and individualized risk assessment. Researchers have the chance to greatly improve the precision and practicality of predictive models in clinical practice by closing this gap [16–20].

RELATED WORK

The underuse of unstructured data, such as clinical notes and discharge summaries, despite their abundance of patient-specific details, represents a substantial gap in the prediction of cardiovascular

disease (CVD) risk [3]. ICD codes and other structured data offer crucial insights, but they frequently lack the contextual depth of unstructured text [11–15]. Innovative models could bridge this gap by using domain-specific natural language processing (NLP) frameworks or transformer-based architectures to extract useful information from unstructured EHR text, allowing for a more thorough and individualized risk assessment. Researchers have the chance to greatly improve the precision and practicality of predictive models in clinical practice by closing this gap (Table 2) [16–20].

Table 1. Literature review of ML/DL models for CVD prediction.

Title	Model type	Dataset size	Results / performance	Key innovation
MIMIC-III [1]	N/A	60,000+ ICU stays	N/A (Data repository)	Rich structured + unstructured data.
Time-Aware LSTM [2]	Time-Aware LSTM	~50,000 (MIMIC-II)	Outperformed standard LSTM in AUC	Time-decay for temporal modeling.
BiteNet [3]	Bi-Encoder Net	29,000 patients	ROC-AUC: 0.82 for event prediction	Temporal hierarchical encoder.
Transformer for CVD [4]	Transformer	54,000 (EHR)	Accuracy: 75–80%	Self-attention for complex time-series.
ICD + Narrative Hybrid [5]	Hybrid Model	million notes (UK Biobank)	F1-score: ↑ by 12% over baseline	Text + code integration.
Scalable DL for HER [6]	Deep Neural Networks	700,000 patients	AUC: 0.87 (heart failure risk)	Scalable DL for clinical data.
Diagnosing with LSTM [7]	LSTM	500,000+ encounters	AUC: 0.88	Event-based sequence learning.
Doctor AI [8]	RNN	260,000 patients (Sutter Health)	Top-10 accuracy: 79%	Early predictive use of RNN in EHR.
Attention is All You Need [9]	Transformer	N/A (Generic NLP)	BLEU Score: 41.8 on WMT'14 EN-DE	Foundation for all attention models.
in Healthcare Review [10]	Various	N/A	N/A (Review study)	Broad DL challenges in health.

Table 2. Comparative analysis of advanced ML/DL approaches and hybrid models for CVD risk prediction [13–26].

AutoML for CVD [11]	AutoML	Accuracy: 85.7%	End-to-end AutoML pipeline	
Longitudinal CVD Models [12]	Statistical Models	AUC: 0.78	Long-term patient trajectory tracking.	
Heart Disease Prediction Review [13]	Review Study	N/A	Summary of 50+ ML & DL models.	
Attention- Based CVD [14]	Hybrid Attention	AUC: 0.81	Improves interpretability & accuracy.	
LSTM (Foundational) [15]	LSTM	N/A	Origin of long-term memory models.	
Static + Dynamic in LSTM [16]	Hybrid LSTM	F1-score: ↑baseline LSTM	Fuses static demographics + time-series.	
THIGE [17]	Graph Embedding	ROC-AUC: 0.84	Temporal risk via heterogeneous graphs.	
AI in CVD Review [18]	Review Study	N/A	Emphasizes independent validation.	
Early CVD with ML [19]	ML Models	5,000+ records	AUC: 0.83	Unveils early warning signals.
Deep Patient [20, 21]	Unsupervised DL	700,000 (Mount Sinai)	Improved prediction for 78 diseases	Deep embeddings from EHRs.

METHODOLOGY

The steps and technical decisions involved in creating the suggested hybrid transformer-based model for CVD risk prediction are described in this section.

Integration of Data and Pre-Processing Three Primary Categories of Data Were Utilized

Lab results, patient demographics (age, gender, ethnicity, and socioeconomic status), and ICD-coded diagnoses are examples of structured data [21–25]. Clinical narratives, such as doctor notes and discharge summaries, are examples of unstructured data.

- *Lifestyle Data:* Details about stress, nutrition, physical activity, and smoking.
- *Data Pre-Processing:* Averages or most prevalent values were used to fill in the missing values for structured data. ClinicalBERT and BioBERT were used to clean (removing punctuation and stopwords), tokenize (splitting into words), and embed the unstructured text data. Numbers were created from lifestyle data using normalization for numerical data and one-hot encoding for categories.

Proposed Framework: Hybrid Transformer-Based Model

Transformer-Based NLP Module

With 12 transformer encoder layers and a hidden size of 768, this module is based on a Clinical Longformer variant. Twelve heads of multi-head self-attention are used to record dependencies in clinical notes. Temporal context-related positional embeddings in unstructured narratives.

Structured Data Module

Two hidden layers (256 units each) in a bidirectional LSTM for encoding sequential ICD-code data. Diagnostic timelines were transformed into a temporal interaction graph using graph-based temporal embeddings, where nodes represented conditions and edges represented temporal co-occurrences [26, 27]. Time gaps were encoded by edge weights using an exponential kernel (temporal decay). Node2Vec was used to initialize the node embeddings. These embeddings were combined into a condensed patient representation using graph attention layers.

Hybrid Fusion Layer

Structured and unstructured embeddings are concatenated into a 512-dimensional vector in the hybrid fusion layer.

- A fully connected layer for dimensionality reduction and feature fusion (128 units, ReLU activation) [28].
- Dropout for regularization (rate = 0.2).

Output Layer

Risk probabilities for each cardiovascular condition are included in the model outputs, which use a Softmax classification layer to forecast the top-N likely diagnoses for the upcoming clinical visit.

Model Training Objective

The goal of the training is to predict the next most likely diagnosis using multi-class classification (Next Diagnosis Prediction).

- *Loss Function:* The loss function is categorical cross-entropy loss.
- *Optimization:* Adam optimizer, which has a 2e-5 learning rate. 32 is the batch size. There are 20 epochs total, with early termination determined by validation loss. ReduceLROnPlateau is a learning rate scheduler that dynamically modifies the learning rate.
- *Evaluation Metrics:* Evaluation metrics include Precision@5 and Precision@20, which measure the accuracy of top-N predictions. F1-score, AUC, and recall for a fair assessment. Cross-validation: To guarantee generalizability, use five-fold cross-validation.
- *Hardware:* An NVIDIA A100 GPU with 40 GB of VRAM was used for the experiments. PyTorch (v1.13) was used to implement the code.

Synthetic Data Generation and Data Augmentation

- Generative Adversarial Networks (GANs) are used to create synthetic data for structured data augmentation while maintaining privacy.

- Federated learning techniques are used to safely combine data from several institutions without sharing it directly.
- Data augmentation increased the robustness of the model, particularly in situations with limited resources.

Explainability Module

- Attention heatmaps are used to display the attention weights from transformer heads and graph attention layers.
- To improve clinical interpretability, key features (ICD codes, text phrases) that contribute to predictions are highlighted [29, 30].

Validation and Deployment

We tested the model on several public datasets, including the Framingham Heart Study [24], MIMIC-III [1], and UK Biobank [23], to ensure the validity of our findings. The fact that these datasets are publicly available aids in the replication of our findings by other researchers. Our model has a feedback loop that allows doctors to provide input and assist us in continuously improving it, and it can be integrated into hospital systems.

RESULTS

The proposed hybrid model for CVD risk prediction, which integrates lifestyle factors, large language models (LLMs) [30], and advanced feature selection techniques, outperforms both traditional and state-of-the-art models in several performance metrics. The following tables provide a summary of the key findings.

Performance Metrics Comparison (Table 3)

Important Results from Performance Measures:

- *Precision@5:* With 78.65%, the proposed model outperforms all existing models and is noticeably superior to the ICD-only model (65.38%) and clinical text-only model (72.50%).
- *Precision@20:* The recommended model (82.12%) outperforms both the ICD-only model (71.80%) and the clinical text-only model (76.30%).
- *Recall:* With a recall of 75.03%, our model performs better than the ICD-only model (68.45%).
- *AUC:* Compared to other models (ICD-only: 0.78, Clinical Text-only: 0.81), the proposed model's AUC of 0.86 suggests that it has superior discriminatory power.
- A high F1-score of 0.76 suggests a good balance between recall and precision.

Table 3. Performance comparison of the proposed hybrid model with baseline and state-of-the-art models [1–3].

Metric	Proposed hybrid model	ICD-only model [Ref: 1]	Clinical text-only model [Ref: 2]	State-of-the-Art (SOT A) [Ref: 3]
Precision@5	78.65%	65.38%	72.50%	70.25%
Precision@20	82.12%	71.80%	76.30%	74.60%
Recall	75.03%	68.45%	70.12%	72.10%
F1-Score	0.76	0.65	0.70	0.68
AUC	0.86	0.78	0.81	0.79
Accuracy	79.45%	72.45%	75.10%	73.20%

Performance by Demographic Group

Important findings from demographic performance:

- Patients in the middle age range (41–60) had the highest Precision@5 (80.75%), indicating that the hybrid model continuously performs better than other demographic groups.
- The lack of gender-based disparities suggests that the model is equitable and works well for a variety of patient groups.

Performance by Comorbidity

Key findings from comorbidity performance:

- Patients with comorbidities, particularly those with diabetes, respond better to the proposed model (Table 4).

Table 4. Demographic-wise performance comparison (Precision@5) of the proposed and baseline models [1–3].

Demographic group	Propose d hybrid model (Precisio n@5)	ICD-only model [Ref: 1]	Clinical text-only model [Ref: 2]	SOTA model [Ref: 3]
Young Patients (18–40)	77.80%	65.90%	72.60%	71.00%
Middle-Aged Patients (41–60)	80.75%	71.20%	74.30%	73.10%
Older Patients (60+)	79.10%	66.50%	70.80%	70.60%
Male Patients	78.10%	69.40%	74.20%	72.80%
Female Patients	79.50%	70.10%	75.30%	74.20%

Evaluation with Lifestyle Factors

Key findings from lifestyle factors:

- Including lifestyle factors like stress levels, physical activity, diet, and smoking history significantly improves the model's predictive accuracy when compared to the clinical text-only model and SOTA model (Table 5).

Table 5. Impact of lifestyle factors on model performance [2, 3]

Lifestyle factor	Proposed hybrid model (Precision@ 5)	Clinical text-only model [Ref: 2]	SOTA model [Ref: 3]
Physical Activity	80.10%	73.50%	74.00%
Dietary Habits	79.30%	71.30%	72.40%
Smoking History	77.20%	69.80%	71.10%
Stress Levels	78.00%	70.20%	72.50%

CONCLUSION

By successfully combining structured data, unstructured clinical text, and lifestyle factors, the suggested hybrid transformer-based model for cardiovascular disease (CVD) risk prediction clearly outperforms both conventional and cutting-edge models. The hybrid model achieved the highest scores in every category, including Precision@5 (78.65%), Precision@20 (82.12%), AUC (0.86), and F1-score (0.76), indicating notable improvements in the performance metrics. These findings point to improved precision and recall balance as well as increased predictive accuracy. Furthermore, the model exhibits its generalizability and fairness by performing consistently across a range of demographic groups, including age and gender. Its effectiveness in treating patients with comorbid conditions like diabetes and hypertension emphasizes how resilient it is to complex clinical profiles. Furthermore, incorporating lifestyle variables like stress levels, food, smoking history, and physical activity offers a more comprehensive picture of a patient's health and enhances prediction results. The proposed model paves the way for more proactive and data-driven clinical decision-making by fusing cutting-edge machine learning techniques with practical applicability, setting a new standard for individualized and interpretable CVD risk assessment.

REFERENCES

1. Johnson AE, Pollard TJ, Shen L, et al. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035. doi: 10.1038/sdata.2016.35.
2. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv* [Preprint]. 2019;abs/1904.05342. Available from: https://api.semanticscholar.org/CorpusID:119308351.

3. Rajkomar A, Oren E, Chen K, et al. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1(1), 18. doi:10.1038/s41746–018-0029–1.
4. Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J. Patient subtyping via time-aware LSTM networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). New York: Association for Computing Machinery; 2017. p. 65--74. doi: 10.1145/3097983.3097997.
5. Wang S, et al. (2019). BiteNet: Bidirectional Encoder Network for Future Clinical Event Prediction. IEEE Transactions on Neural Networks and Learning Systems.
6. Peng J, et al. (2020). Transforming clinical event prediction with self-attention mechanism-based transformers. Journal of Biomedical Informatics.
7. Zhang Z, et al. (2021). Hybrid models for healthcare prediction: combining ICD codes with clinical narratives. Journal of Medical Systems. doi: 10.1007/s10916–021-01630-x.
8. Rajkomar A, et al. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine. DOI: 10.1038/s41746–018-0029–1.
9. Lipton ZC, Kale DC, Elkan C, & Wetzel R. (2016). Learning to diagnose with LSTM recurrent neural networks. arXiv preprint. DOI: 10.48550/arXiv.1706.03762.
10. Choi E, Bahadori MT, Schuetz A, Stewart WF, & Sun J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. Machine Learning for Healthcare Conference.
11. Vaswani A, Shazeer N, Parmar N, et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems. DOI: 10.48550/arXiv.1706.03762.
12. Miotto R, Wang F, Wang S, Jiang X, & Dudley JT. (2018). Deep learning for healthcare: Review, opportunities, and challenges. Briefings in Bioinformatics. DOI: 10.1093/bib/bbx044.
13. Alaa AM, Bolton T, Di Angelantonio E, Rudd JHF, & van der Schaar M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLOS One. DOI: 10.1371/journal.pone.0213653.
14. Stevens D, Lane DA, Harrison SL, Lip GYH, & Kolamunnage-Dona R. (2021). Modelling of longitudinal data to predict cardiovascular disease risk: A methodological review. BMC Medical Research Methodology. DOI: 10.1186/s12874–021-01288–4.
15. Vishnu Vardhana Reddy Karna, et al. (2024). A Comprehensive Review on Heart Disease Risk Prediction using Machine Learning and Deep Learning Algorithms. Archives of Computational Methods in Engineering.
16. Lee S, et al. (2021). Attention-based models for cardiovascular disease prediction: A hybrid approach. IEEE Journal of Biomedical and Health Informatics.
17. Hochreiter S, Schmidhuber J. (1997). Long Short-Term Memory. Neural Computation. DOI: 10.1162/neco.1997.9.8.1735.
18. Esteban C, et al. (2016). Predicting clinical events by combining static and dynamic data in LSTM models. PLOS ONE. DOI: 10.1371/journal.pone.0146251.
19. Zhang Y, et al. (2020). THIGE: Temporal Heterogeneous Interaction Graph Embedding for Health Data Analysis. Neural Networks and Applications.
20. Cai Y, et al. (2024). Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: A systematic review. BMC Medicine. DOI: 10.1186/s12916–024-02835-z.
21. Deepa R, et al. (2024). Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from health records. AIP Advances. DOI: 10.1063/1.5128374.
22. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep.* 2016;6(1):26094. doi:10.1038/srep26094.
23. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 2015 Mar 31;12(3):e1001779. doi: 10.1371/journal.pmed.1001779. PMID: 25826379; PMCID: PMC4380465.
24. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families. The Framingham offspring study. Am J Epidemiol. 1979 Sep;110(3):281-90. doi: 10.1093/oxfordjournals.aje.a112813. PMID: 474565.

25. Gerela P, Mishra PN, & Vipat R. (2022). Study on data visualization: Its importance in education sector. International Journal of Health Sciences, 6(S3), 6298–6305. https://doi.org/10.53730/ijhs.v6nS3.7393.
26. Mishra PN, Gerala P, & Maitra S. (2022). Study on artificial intelligence applications uses in agriculture. Int J Health Sci, 6(S2), 9162–9173. https://doi.org/10.53730/ijhs.v6nS2.7391.
27. Basavaraj GN, Ainapure B, Sowmya MR, Sandeep C, Mishra PN, Lakkimsetty NR, Dakulagi V, & Shaik F. (2025). Machine Learning-enhanced Direction-of-Arrival Estimation for Coherent and Non-Coherent Sources. Engineering, Technology & Applied Science Research, 15(2), 20647–20652. https://doi.org/10.48084/etasr.9494.
28. Mishra P, Gaikwad V, Dhawan A, Bagul R, Shaikh A, & Singh R. (2025). Precision agriculture meets AI: Predicting nutritional crop outcomes from genomic data. International Journal of Environmental Sciences, 11(14S), 207–218. https://www.theaspd.com/ijes.php.
29. Ahmad E, Dash B, Tripathi A, et al. Hybrid CNN and image processing framework for precise characterization of cracks in concrete structures. Asian J Civ Eng (2025). https://doi.org/10.1007/s42107–025-01535–0.
30. Gaikwad V, Dhawan A, Mishra PN, Kumarasamy M. AI-enabled early detection of fetal gestational age and CNS anomalies in the first trimester through ultrasound to support rural doctors in India. SSRG Int J Electron Commun Eng. 2025;12(7):174–183. doi: 10.14445/23488549/IJECE-V12I7P113.

