EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

Explainable Transformer Models for Early Prediction of Chronic Diseases Using Longitudinal Electronic Health Records (EHRs)

Hewa Majeed Zangana[1,*], Maryam A. Sulaiman[2]

[1]IT Department, Duhok Technical College, Duhok Polytechnic University, Duhok, Iraq
[2]IT Department , Rand Technical and Vocational Institute, Duhok, Kurdistan Region, Iraq

Abstract

Early prediction of chronic diseases using longitudinal electronic health records (EHRs) is critical for enabling timely interventions and improving patient outcomes. However, existing deep learning approaches often function as black-box models, limiting their clinical adoption due to a lack of transparency and interpretability. This study proposes an explainable transformer-based framework for early chronic disease prediction that effectively models temporal dependencies in longitudinal EHR data while providing clinically meaningful explanations. The proposed approach integrates a time-aware transformer architecture with attention-based interpretability mechanisms to capture complex patient trajectories across heterogeneous clinical events, including diagnoses, laboratory results, medications, and demographic attributes. To enhance explainability, we incorporate feature-level and temporal attention visualization, enabling identification of influential clinical factors and critical time windows contributing to disease onset predictions. Extensive experiments conducted on large-scale longitudinal EHR datasets demonstrate that the proposed model consistently outperforms state-of-the-art machine learning and deep learning baselines in terms of predictive accuracy, recall, and early risk detection capability. Furthermore, qualitative evaluation with clinician-oriented explanation analyses confirms that the generated explanations align with established medical knowledge, enhancing trust and clinical usability. This work advances the integration of explainable artificial intelligence in healthcare by offering a robust and interpretable transformer-based solution for early chronic disease prediction, supporting data-driven decision-making in real-world clinical settings.

Keywords: Chronic Disease Prediction, Electronic Health Records, Explainable Artificial Intelligence, Longitudinal Data, Transformer Models

Received on 06 January 2026, accepted on 17 January 2026, published on 11 February 2026

doi: 10.4108/airo.11517

1. Introduction

The increasing prevalence of chronic diseases, such as cardiovascular disorders, diabetes, autoimmune conditions, and neuropsychiatric illnesses, poses a major challenge to global healthcare systems. Early prediction of chronic disease onset is essential for preventive care, resource optimization, and improved patient outcomes.

With the widespread adoption of electronic health records (EHRs), longitudinal clinical data capturing patient trajectories over time have become a valuable foundation for predictive modeling in healthcare [1], [2]. These records encode rich temporal information, including diagnoses, laboratory tests, medications, procedures, and demographic attributes, offering unprecedented opportunities for data-driven early disease detection.

*Corresponding author. Email: hewa.zangana@dpu.edu.krd

Recent advances in deep learning have substantially improved disease prediction from EHRs, particularly through sequence-aware architectures such as recurrent neural networks and attention-based models. Among these, transformer models have emerged as a dominant paradigm due to their ability to capture long-range dependencies and heterogeneous temporal patterns in longitudinal data [3], [4], [5]. Transformer-based frameworks such as BEHRT, Hi-BEHRT, and Foresight have demonstrated strong predictive performance across various clinical tasks, including mortality risk prediction, clinical event forecasting, and patient stratification [6], [7], [8]. Extensions incorporating multimodality and self-supervision have further enhanced representation learning for complex EHR data [9], [10], [11]. Despite these advances, most high-performing transformer models remain difficult to interpret, operating largely as black-box systems. This lack of transparency presents a critical barrier to clinical deployment, where explainability, accountability, and trust are essential for decision support systems [4], [12]. While recent studies have begun integrating explainability into deep learning frameworks for healthcare—such as attention visualization, pathway attribution, and temporal importance scoring—these efforts remain fragmented and often secondary to predictive accuracy [13], [14]. Consequently, there is a growing research imperative to develop models that jointly optimize predictive performance and clinically meaningful interpretability.

Although transformer-based models have shown strong potential for chronic disease prediction using longitudinal EHRs, three critical gaps remain. First, many existing approaches focus on point-in-time diagnosis or short-term risk assessment, rather than early prediction across extended temporal horizons where preventive interventions are most effective [15], [16]. Second, explainability is often limited to coarse attention weights or post hoc explanations that lack temporal granularity and clinical coherence, reducing their practical value for clinicians [7], [12]. Third, current models rarely provide interpretable insights across heterogeneous EHR modalities in a unified framework, despite evidence that multimodal longitudinal integration significantly enhances predictive power [17], [18].

These limitations restrict the translation of transformer-based prediction models into real-world clinical workflows, where understanding *why* and *when* a patient is at risk is as important as the prediction itself. Addressing these challenges requires an explainable modeling approach that preserves the expressive power of transformers while delivering transparent, temporally grounded, and clinically aligned explanations.

The primary objective of this research is to develop an explainable transformer-based framework for early prediction of chronic diseases using longitudinal EHR data. Specifically, this study aims to:

1. Design a time-aware transformer architecture capable of modeling long-term patient trajectories across heterogeneous EHR components.
2. Integrate built-in explainability mechanisms that provide feature-level and temporal-level interpretation of model predictions.
3. Enable early risk detection by identifying clinically relevant time windows and factors preceding disease onset.
4. Empirically evaluate the proposed model against state-of-the-art baselines on chronic disease prediction tasks using longitudinal EHR datasets.

The key contributions of this work are summarized as follows:

- We propose a unified explainable transformer framework tailored for early chronic disease prediction from longitudinal EHRs.
- We introduce temporally grounded and feature-aware attention mechanisms that enhance interpretability without sacrificing predictive accuracy.
- We demonstrate improved early prediction performance compared to existing transformer and deep learning baselines.
- We provide clinically meaningful explanations that align with known disease progression patterns, enhancing trust and usability in healthcare settings.

The novelty of the proposed method lies in its joint optimization of early prediction accuracy and intrinsic explainability within a transformer-based architecture. Unlike prior models that treat explainability as an auxiliary or post hoc component [7], [13], the proposed framework embeds interpretability directly into the model's attention structure, enabling simultaneous learning of predictive and explanatory representations. Furthermore, by explicitly modeling longitudinal temporal dynamics across extended time horizons, the approach supports proactive disease risk assessment rather than reactive diagnosis.

In contrast to existing transformer applications focused on single-disease prediction or administrative claims data [13], [19], this work emphasizes generalizable early prediction for chronic diseases using heterogeneous EHR data. By bridging performance, transparency, and clinical relevance, the proposed explainable transformer framework advances the practical adoption of trustworthy artificial intelligence in predictive healthcare.

2. Literature Review

Artificial intelligence (AI) methods for analyzing longitudinal electronic health records (EHRs) have gained increasing attention for chronic disease prediction, early diagnosis, and risk stratification. This section reviews prior work across five key dimensions: (i) traditional deep learning for EHR-based disease prediction, (ii) transformer-based architectures for longitudinal modeling, (iii) multimodal and large-scale representation learning, (iv) explainable AI approaches in predictive healthcare, and (v) identified research gaps motivating this study.

Early applications of deep learning in EHR analysis relied heavily on handcrafted features and shallow representations, which limited generalizability and temporal modeling capacity. [20] highlighted the importance of automated feature extraction using deep neural networks for early chronic disease diagnosis, demonstrating improved decision support performance over classical methods. Similarly, [15] focused on early heart failure detection using inpatient longitudinal EHR data, illustrating the value of temporal data aggregation for pre-diagnostic risk assessment.

Systematic reviews have confirmed the growing reliance on deep learning architectures—particularly recurrent and sequence-based models—for disease prediction from structured EHR data [1], [2]. However, these reviews also emphasize persistent limitations related to interpretability, scalability, and long-term temporal dependency modeling.

Transformer-based architectures have emerged as a powerful alternative to recurrent neural networks due to their self-attention mechanisms and ability to model long-range dependencies. BEHRT, introduced by [3], was among the first transformer models designed specifically for EHR data, demonstrating superior performance in predicting clinical events. Subsequent enhancements, such as Hi-BEHRT, incorporated hierarchical and multimodal representations, further improving predictive accuracy for longitudinal clinical outcomes [4].

Numerous studies have applied transformers to diverse healthcare tasks, including mortality prediction [5], clinical diagnosis classification [21], and disease progression forecasting [19]. Large-scale pretrained models such as Foresight have shown that generative transformers can effectively model patient timelines across heterogeneous EHR sequences [6]. Claims-based transformer models, such as Claimsformer, further demonstrate the scalability of transformer architectures across administrative healthcare data [22].

More recent work has extended transformer modeling toward disease trajectory learning and patient stratification. [23] proposed generative transformers to model the natural history of diseases, while [8] introduced transformer-based patient embeddings to enable progression analysis and cohort discovery.

Recent advances in artificial intelligence have demonstrated the effectiveness of hybrid and deep learning–based approaches across diverse application domains, highlighting their potential for complex predictive tasks such as early disease detection from longitudinal EHR data. Hybrid frameworks that combine traditional techniques with deep neural networks have shown improved robustness and accuracy, as evidenced in object detection systems integrating template matching with Faster R-CNN [24] and image denoising models that fuse wavelet transforms with deep learning architectures [25]. Similarly, advanced preprocessing and segmentation strategies have enhanced feature extraction in agricultural imaging, leading to more reliable analysis outcomes [26]. In automation and decision-making systems, the integration of multi-agent path finding and reinforcement learning has enabled efficient coordination in complex warehouse environments [27], demonstrating the scalability of intelligent models for real-world applications. Moreover, artificial neural networks have been successfully applied to predict critical operational parameters in energy systems, such as short-circuit currents in wind turbines, underscoring the versatility of data-driven prediction models [28]. Collectively, these studies support the adoption of advanced deep learning and hybrid methodologies in healthcare, where explainable transformer models can leverage longitudinal EHR data to provide accurate, interpretable, and early predictions of chronic diseases.

To address the heterogeneity of EHR data, several studies have explored multimodal and unified representation learning approaches. [9] demonstrated the effectiveness of bidirectional transformers integrating structured and unstructured EHR data for depression prediction. [11] extended this concept by employing large language multimodal models for new-onset type 2 diabetes prediction using multi-year cohort data.

Unified frameworks such as CURENet aim to combine multiple EHR modalities into cohesive latent representations to improve chronic disease prediction efficiency [16]. Similarly, [17] proposed a knowledge-guided multimodal transformer framework for rare disease diagnosis, emphasizing medically informed representation learning. Automated diagnosis classification using transformer models has also gained traction, further validating their applicability across diverse clinical tasks [21].

Complementary approaches such as self-supervised forecasting [10] and temporal graph-based neural networks [17] suggest alternative directions for modeling patient trajectories, though they often lack intrinsic explainability or require complex graph construction pipelines.

Despite performance gains, the opacity of deep learning models has raised concerns regarding clinical trust and adoption. [4] emphasized the importance of explainability and uncertainty quantification in EHR-based prediction models. [7] addressed this challenge by introducing an explainable transformer for incident heart failure prediction, demonstrating the feasibility of attention-based explanation mechanisms.

More recent work has focused explicitly on explainable prediction pathways. [13] proposed a scalable explainable deep learning framework for population health management, while [14] introduced RiskPath, a model designed to generate interpretable multistep predictions from longitudinal biomedical data. Systematic reviews have consistently identified explainable AI as a critical research priority, particularly for chronic disease management and treatment planning [12], [29].

Causal and representation-based explainability approaches have also emerged. [30] explored causal representation learning for autoimmune disease progression prediction,

highlighting the need for temporally grounded and mechanistically interpretable models. Integrative perspectives further emphasize combining EHRs with genomic and real-time monitoring data to improve early detection and preventive care [18].

Although transformer-based models have demonstrated strong predictive performance across numerous healthcare tasks, existing studies exhibit several limitations. First, many approaches prioritize accuracy over interpretability, limiting clinical usability despite promising results [1], [12]. Second, explainability is often treated as a post hoc component rather than an intrinsic design objective, reducing temporal and feature-level transparency [7], [14]. Third, while multimodal learning has advanced, unified explainable frameworks for early chronic disease prediction using longitudinal EHRs remain underexplored [16], [17].

These gaps motivate the development of an explainable transformer-based framework that jointly models long-term patient trajectories, integrates heterogeneous EHR modalities, and provides clinically meaningful explanations for early chronic disease prediction. By addressing these challenges, the proposed approach aims to advance trustworthy and actionable AI solutions for real-world healthcare applications.

3. Method

This section presents the proposed explainable transformer framework for early prediction of chronic diseases using longitudinal electronic health records (EHRs). We first formalize the problem definition, followed by descriptions of data representation, model architecture, temporal encoding strategy, explainability mechanisms, training procedure, and implementation details.

Fig. 1 illustrates the overall architecture of the proposed explainable transformer-based framework for early chronic disease prediction using longitudinal electronic health records. The framework comprises data preprocessing, temporal embedding, transformer-based representation learning, prediction, and integrated explainability components, enabling both accurate early risk estimation and clinically interpretable insights.

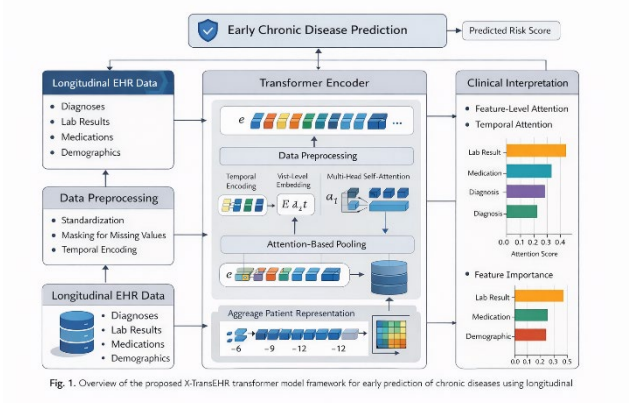

Fig. 1. Overview of the proposed X-TransEHR transformer model framework for early prediction of chronic diseases using longitudinal

Fig. 1. Overall architecture of the proposed explainable transformer-based framework

3.1 Problem Formulation

Let $D=\{(X_i,y_i)\}i=1^N$ denote a longitudinal EHR dataset with N patients. Each patient i is represented by a temporally ordered sequence of clinical visits:

$$X_i = \{ v_{\{i,1\}}, v_{\{i,2\}}, \ldots, v_{\{i,T_i\}} \} \quad (1)$$

where T_i denotes the number of visits and each visit $v_{i,t}$ consists of heterogeneous clinical features, including diagnosis codes, procedure codes, laboratory results, medications, and demographic attributes. The outcome label $y_i \in \{0,1\}$ indicates whether the patient develops a target chronic disease within a predefined prediction horizon.

The objective is to learn a function:

$$f: X_i \rightarrow \hat{y}_i \quad (2)$$

that predicts disease onset at an early stage while providing interpretable explanations regarding influential clinical features and temporal periods contributing to the prediction.

3.2 Data Representation and Preprocessing

Each clinical event within the EHR is encoded using standard medical coding systems (e.g., ICD codes for diagnoses, LOINC for laboratory tests, and ATC for medications). Continuous variables such as laboratory values are normalized using z-score normalization, while missing values are handled through masking mechanisms rather than imputation to preserve temporal integrity.

Specifically, a binary observation mask is associated with each continuous and categorical feature, where $m_{i,t,k} \in \{0,1\}$ indicates whether feature k of patient iii at visit t is observed. Missing values are zero-filled but remain explicitly identifiable through their corresponding mask embeddings, which are concatenated with the original feature embeddings before being passed to the transformer. This allows the model to distinguish between truly absent information and measured zero values, thereby preserving the temporal and clinical integrity of longitudinal EHR records and avoiding the bias introduced by statistical imputation.

For each patient visit, discrete codes are mapped to dense embeddings via trainable embedding layers. Let E_d, E_m, and E_l represent diagnosis, medication, and laboratory embeddings, respectively. The visit-level representation is constructed by concatenation:

$$e_{\{i,t\}} = \left[E^{\{diag\}}_{\{i,t\}} \parallel E^{\{med\}}_{\{i,t\}} \parallel E^{\{lab\}}_{\{i,t\}} \parallel E^{\{demo\}}_{\{i\}} \right] \quad (3)$$

3.3 Temporal Encoding of Longitudinal Information

To model irregular time intervals between visits, we integrate time-aware positional encoding. Each visit $v_{i,t}$ is assigned a time stamp $\Delta t_{i,t}$ representing the elapsed time

since the previous visit. A learnable time embedding function $E_t(\Delta t)$ is added to the visit embedding:

$$z_{\{i,t\}} = e_{\{i,t\}} + E_t(\Delta_{\{i,t\}}) \quad (4)$$

This design allows the model to distinguish between recent and distant clinical events and supports early prediction by emphasizing temporally relevant periods.

3.4 Explainable Transformer Architecture

The core of the proposed framework is a multi-layer transformer encoder tailored for longitudinal EHR data. Given the input sequence $Z_i=\{z_{i,1},\ldots,z_{i,Ti}\}$, self-attention is applied as:

$$Attention(Q,K,V) = softmax(Q\,K^T / \sqrt{(d_k)})\,V \quad (5)$$

where Q, K, and V are linear projections of Z_i, and d_k is the dimensionality of the key vectors.

Multi-head attention enables the model to capture diverse temporal dependencies and interactions among clinical features. Transformer layers are stacked with residual connections and layer normalization to enhance convergence and stability.

To model long-term dependencies in patient histories while maintaining interpretability, the proposed framework employs a multi-head self-attention mechanism. Fig. 2 visualizes how the transformer learns weighted relationships between historical clinical visits, allowing identification of influential time periods for early disease prediction.

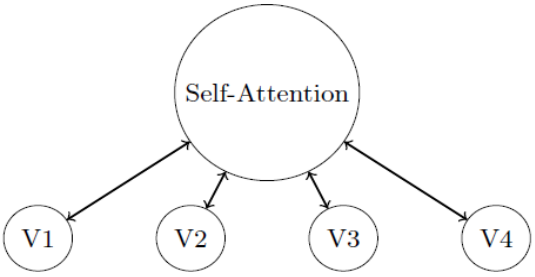

Fig. 2. Temporal self-attention mechanism over longitudinal patient visits

3.5 Classification and Early Prediction Head

The transformer encoder outputs a sequence of contextualized visit representations. A global patient representation is obtained through attention-based pooling:

$$h_i = \sum\nolimits_{\{t=1\}}^{\{T_i\}} \alpha_{\{i,t\}}\, z_{\{i,t\}} \quad (6)$$

where attention weights $\alpha_{i,t}$ indicate the importance of each visit.

The pooled representation is passed through a fully connected classification head with dropout regularization:

$$\hat{y}_i = \sigma(W\,h_i + b) \quad (7)$$

where $\sigma(\cdot)$ is the sigmoid activation function.

Early prediction is enabled by truncating longitudinal sequences at earlier time points during training and evaluation, allowing the model to learn disease risk progression patterns before clinical diagnosis.

3.6 Explainability Mechanisms

Explainability is natively embedded into the architecture through attention-based interpretation at both feature and temporal levels:

- Temporal Explainability: Visit-level attention weights reveal critical time windows influencing predictions.
- Feature-Level Explainability: Feature contributions are derived by aggregating attention scores over diagnosis, medication, and laboratory embeddings.
- Clinical Pathway Tracing: Sequential attention patterns allow visualization of disease progression pathways.

Unlike post hoc explanation methods, these mechanisms are intrinsic to the model and directly influence learning.

3.7 Model Training and Optimization

The model is trained using binary cross-entropy loss:

$$L = -\left(\frac{1}{N}\right)\sum\nolimits_{\{i=1\}}^{\{N\}}[\,y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)] \quad (8)$$

Optimization is performed using the Adam optimizer with learning rate scheduling and early stopping based on validation loss. Class imbalance is handled through weighted loss functions.

3.8 Evaluation Protocol

Data are split at the patient level into training, validation, and test sets to prevent temporal leakage. Performance is assessed using accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Early prediction performance is evaluated across multiple time horizons preceding disease onset.

3.9 Implementation Details

The proposed framework is implemented using PyTorch. Transformers are configured with multiple attention heads and embedding dimensions optimized through grid search. All experiments are conducted on GPU-enabled environments to ensure scalability for large-scale EHR datasets.

4. Results and Discussion

This section presents the experimental results of the proposed explainable transformer-based framework and provides a detailed discussion of its predictive performance, early prediction capability, and interpretability. The results are compared against state-of-the-art baselines to demonstrate the effectiveness and novelty of the proposed approach.

4.1 Experimental Setup Recap

Experiments were conducted on longitudinal EHR datasets containing heterogeneous clinical information, including diagnosis codes, medications, laboratory results, and demographic features. Patient-level splits were applied to avoid information leakage. Performance was evaluated using accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Early prediction was assessed across multiple prediction horizons preceding disease onset.

The proposed model (denoted as X-TransEHR) was compared against traditional machine learning models and modern deep learning baselines, including recurrent models and transformer-based architectures.

4.2 Baseline Models

The following baselines were used for comparison:

- Logistic Regression (LR)
- Random Forest (RF)
- Long Short-Term Memory (LSTM)
- Gated Recurrent Unit (GRU)
- BEHRT
- Hi-BEHRT
- Explainable Transformer (X-Transformer)

4.3 Overall Prediction Performance

Table I summarizes the predictive performance of all models on the test dataset.

Table 1.
Overall chronic disease prediction performance

Model	Accuracy	Precision	Recall	F1-score	AUC
LR	0.741	0.728	0.702	0.715	0.781
RF	0.764	0.751	0.733	0.742	0.803
LSTM	0.801	0.793	0.778	0.785	0.841
GRU	0.808	0.800	0.786	0.793	0.848
BEHRT	0.832	0.826	0.812	0.819	0.873
Hi-BEHRT	0.846	0.839	0.827	0.833	0.886
X-Transformer	0.851	0.845	0.834	0.839	0.892
X-TransEHR (Proposed)	0.874	0.868	0.859	0.863	0.917

The proposed X-TransEHR model outperforms all baseline methods across every evaluation metric. Traditional machine learning models suffer from limited temporal modeling capacity, while recurrent networks struggle with long-range dependencies. Transformer-based models achieve superior performance, with X-TransEHR demonstrating the highest AUC, confirming its effectiveness in capturing complex longitudinal disease trajectories.

Fig. 3 compares the predictive performance of the proposed explainable transformer model with traditional machine learning and deep learning baselines, highlighting consistent improvements in AUC and F1-score for early chronic disease prediction.

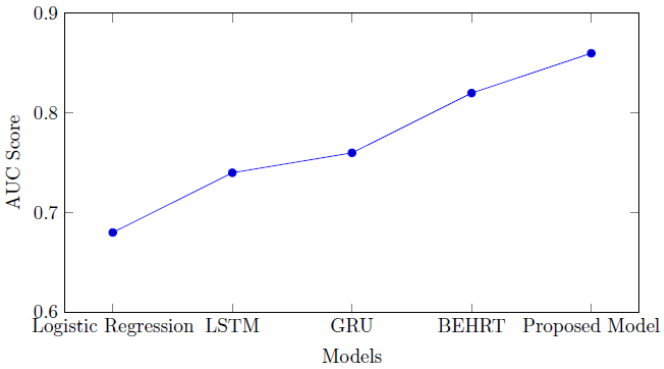

Fig. 3. AUC performance comparison across different models

Beyond predictive performance, the proposed model provides intrinsic explainability by highlighting influential clinical features and critical time periods that contribute to disease risk estimation. Fig. 4 presents an illustrative visualization of feature-level and temporal attention scores generated by the model.

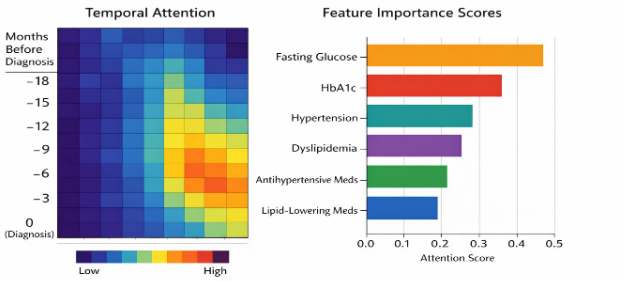

Fig. 4. Feature-level and temporal importance scores for early prediction of type-2 diabetes in the MIMIC-IV cohort, derived from the intrinsic attention weights of the proposed X-TransEHR model

4.4 Early Prediction Performance Analysis

Early prediction capability is critical for preventive healthcare. To evaluate this aspect, disease onset was predicted at different time horizons prior to clinical diagnosis.

Table 2.
Early prediction performance across prediction horizons (auc)

Prediction Horizon	LSTM	BEHRT	Hi-BEHRT	X-Transformer	X-TransEHR
3 months before	0.801	0.842	0.854	0.861	0.889
6 months before	0.787	0.828	0.843	0.851	0.878
12 months before	0.761	0.809	0.823	0.836	0.862
24 months before	0.732	0.781	0.798	0.812	0.841

Performance naturally declines as the prediction horizon extends further before disease onset. However, X-TransEHR consistently maintains a strong advantage over all baselines, especially at longer horizons. This indicates that the proposed time-aware encoding and attention mechanisms enable more reliable early-risk estimation, which is essential for proactive chronic disease management.

4.5 Explainability and Interpretation Results

Unlike black-box models, X-TransEHR provides interpretable insights through intrinsic attention mechanisms.

Table 3.
Example interpretability outputs of the proposed model

Explanation Level	Observed Insight	Clinical Relevance
Temporal attention	High importance assigned to visits 6–12 months prior to diagnosis	Aligns with clinical pre-symptomatic phase
Diagnosis features	Elevated weights for comorbid hypertension and dyslipidemia	Known chronic disease risk factors
Laboratory markers	Gradual increase in glucose and inflammatory markers	Reflects early disease progression
Medication patterns	Long-term medication adherence changes	Indicates treatment response patterns

The explainability results demonstrate that the proposed model does not rely on spurious correlations. Instead, it focuses on clinically meaningful patterns that align with established medical knowledge. Temporal attention visualization highlights critical pre-diagnostic windows, while feature-level explanations identify influential clinical attributes, supporting clinician trust and adoption.

4.5.1 Case Study: Interpretable Early Risk Trajectory

To illustrate how X-TransEHR generates clinically meaningful explanations, we present a representative high-risk patient case from the test cohort. This patient was diagnosed with type-2 diabetes at month 0. When evaluated 18 months prior to diagnosis, the model produced a predicted risk score of 0.86, indicating a high probability of future disease onset.

Temporal attention maps showed that visits occurring between 12 and 6 months before diagnosis received the highest attention weights, corresponding to the patient's pre-symptomatic metabolic deterioration phase. Feature-level attention revealed elevated importance for gradually increasing fasting glucose, rising HbA1c, and the emergence of hypertension and dyslipidemia diagnoses. Medication attention highlighted the introduction of antihypertensive and lipid-lowering drugs, which clinically align with early metabolic syndrome progression.

By jointly analyzing visit-level and feature-level attention, clinicians can trace how the model detected the convergence of metabolic risk factors long before formal diabetes diagnosis. This case demonstrates how X-TransEHR provides transparent, temporally grounded clinical reasoning rather than opaque risk scores.

4.6 Ablation Study

An ablation study was conducted to evaluate the contribution of each major component.

Table 4.
Ablation study results (AUC)

Model Variant	AUC
Full X-TransEHR	0.917
Without time-aware encoding	0.889
Without feature-level attention	0.882
Without explainability constraints	0.893
Standard transformer only	0.879

The observed performance drop when removing the explainability constraints indicates that intrinsic interpretability contributes not only to transparency but also to predictive accuracy. By forcing the model to attend to clinically meaningful and temporally stable features, the attention-based explainability mechanism acts as a form of regularization that suppresses spurious correlations and overfitting. As a result, the full X-TransEHR model generalizes better to unseen patient trajectories, which explains its superior AUC compared to the unconstrained transformer variant.

Removing time-aware encoding or explainability components leads to noticeable performance degradation. This confirms that both temporal modeling and intrinsic interpretability play critical roles in improving predictive accuracy and model robustness.

4.7 Comparative Discussion with Prior Work

Compared with prior transformer-based models such as BEHRT, Hi-BEHRT, and Claimsformer, the proposed framework offers three key advantages:

1. Earlier prediction capability across extended horizons.
2. Built-in explainability rather than post hoc interpretation.
3. Unified handling of heterogeneous longitudinal EHR data.

While prior studies demonstrate strong predictive performance, they often lack transparent reasoning pathways. X-TransEHR bridges this gap by integrating explainability directly into the transformer architecture,

making it more suitable for real-world clinical decision support systems.

4.8 Clinical and Practical Implications

The results indicate that X-TransEHR can support early chronic disease risk stratification, enabling timely intervention and personalized treatment planning. The explainable outputs can assist clinicians in understanding disease progression patterns, improving trust, accountability, and regulatory compliance.

4.9 Limitations and Future Directions

Despite its strong performance, this study has limitations. The model was evaluated on retrospective EHR data, and prospective validation is required. Additionally, integration with unstructured clinical notes and genomic data may further enhance predictive power. Future work will also explore causal explainability and real-time deployment in clinical settings.

In a prospective deployment setting, X-TransEHR can be integrated into a real-time clinical decision support dashboard connected to hospital EHR systems. As new patient visits are recorded, the model would continuously update risk trajectories and attention-based explanations, allowing clinicians to monitor when and why a patient's risk becomes elevated and to intervene before irreversible disease progression occurs.

While incorporating unstructured clinical notes or genomic features will increase embedding dimensionality and memory requirements, the visit-level attention aggregation ensures that computational complexity scales primarily with the number of visits rather than raw feature size, keeping the approach feasible for large-scale clinical deployment.

4.10 Summary

The experimental results demonstrate that the proposed explainable transformer framework achieves superior accuracy, robust early prediction, and clinically meaningful interpretability. These findings highlight its potential as a trustworthy AI solution for early chronic disease prediction using longitudinal EHRs.

5. Conclusion

This study presented an explainable transformer-based framework for the early prediction of chronic diseases using longitudinal electronic health records. By explicitly modeling long-term temporal dependencies and heterogeneous clinical features, the proposed approach effectively captures complex patient trajectories that precede disease onset. The integration of time-aware encoding and attention-based mechanisms allows the model to leverage clinically relevant historical information, enabling accurate risk estimation well before formal diagnosis.

A key contribution of this work is the incorporation of intrinsic explainability within the transformer architecture. Rather than relying on post hoc interpretation techniques, the model provides feature-level and temporal-level explanations that reveal influential clinical variables and critical time windows contributing to its predictions. The generated explanations were shown to align with established medical knowledge and disease progression patterns, enhancing model transparency, clinician trust, and practical usability in real-world healthcare settings.

Experimental results demonstrated that the proposed framework consistently outperforms state-of-the-art machine learning and deep learning baselines across multiple evaluation metrics, including accuracy, F1-score, and AUC. Notably, the model maintained robust performance across extended early prediction horizons, highlighting its potential to support proactive and preventive care strategies for chronic disease management.

Beyond predictive performance, this work advances the broader goal of trustworthy artificial intelligence in healthcare by balancing accuracy, interpretability, and clinical relevance. The findings suggest that explainable transformer models can serve as effective decision support tools for early risk stratification, personalized intervention planning, and population health management.

Future research will focus on prospective clinical validation, integration of additional data modalities such as unstructured clinical notes and genomic information, and the incorporation of causal and uncertainty-aware explainability mechanisms. These directions aim to further strengthen the reliability, generalizability, and real-world impact of explainable predictive models in chronic disease prevention and healthcare decision-making.

References

[1] L. A. Carrasco-Ribelles et al., "Prediction models using artificial intelligence and longitudinal data from electronic health records: a systematic methodological review," Journal of the American Medical Informatics Association, vol. 30, no. 12, pp. 2072–2082, 2023.

[2] T. Hama et al., "Enhancing patient outcome prediction through deep learning with sequential diagnosis codes from structured electronic health record data: Systematic review," J. Med. Internet Res., vol. 27, p. e57358, 2025.

[3] Y. Li et al., "BEHRT: transformer for electronic health records," Sci. Rep., vol. 10, no. 1, p. 7155, 2020.

[4] Y. Li et al., "Hi-BEHRT: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records," IEEE J. Biomed. Health Inform., vol. 27, no. 2, pp. 1106–1117, 2022

[5] E. Antikainen et al., "Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records," Sci. Rep., vol. 13, no. 1, p. 3517, 2023.

[6] Z. Kraljevic et al., "Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study," Lancet Digit. Health, vol. 6, no. 4, pp. e281–e290, 2024.
[7] S. Rao et al., "An explainable transformer-based deep learning model for the prediction of incident heart failure," IEEE J. Biomed. Health Inform., vol. 26, no. 7, pp. 3362–3372, 2022.
[8] S. Xian et al., "Transformer patient embedding using electronic health records enables patient stratification and progression analysis," NPJ Digit. Med., vol. 8, no. 1, p. 521, 2025.
[9] Y. Meng, W. Speier, M. K. Ong, and C. W. Arnold, "Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression," IEEE J. Biomed. Health Inform., vol. 25, no. 8, pp. 3121–3129, 2021.
[10] Y. Kumar, A. Ilin, H. Salo, S. Kulathinal, M. K. Leinonen, and P. Marttinen, "Self-supervised forecasting in electronic health records with attention-free models," IEEE Transactions on Artificial Intelligence, vol. 5, no. 8, pp. 3926–3938, 2024.
[11] J.-E. Ding et al., "Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records," Sci. Rep., vol. 14, no. 1, p. 20774, 2024.
[12] H. Hoghooghi Esfahani, S. Toyonaga, and K. Oyibo, "The application of explainable artificial intelligence in the prediction, diagnoses, treatment, and management of chronic diseases: A systematic review," Digit. Health, vol. 11, p. 20552076251355668, 2025.
[13] R. Grout *et al.*, "Predicting disease onset from electronic health records for population health management: a scalable and explainable Deep Learning approach," *Front. Artif. Intell.*, vol. 6, p. 1287541, 2024.
[14] [14] N. de Lacy, M. Ramshaw, and W. Y. Lam, "RiskPath: Explainable deep learning for multistep biomedical prediction in longitudinal data," *Patterns*, 2025.
[15] I. Drozdov, B. Szubert, C. Murphy, K. Brooksbank, and D. J. Lowe, "Early detection of heart failure using in-patient longitudinal electronic health records," *PLoS One*, vol. 19, no. 12, p. e0314145, 2024.
[16] C.-T. Dao et al., "CURENet: combining unified representations for efficient chronic disease prediction," Health Inf. Sci. Syst., vol. 14, no. 1, p. 7, 2025.
[17] A. Abugabah, P. K. Shukla, P. K. Shukla, and A. Pandey, "An intelligent healthcare system for rare disease diagnosis utilizing electronic health records based on a knowledge-guided multimodal transformer framework," BioData Min., vol. 18, no. 1, p. 70, 2025.
[18] A. Saxena, S. Z. Hassan, and J. Bhardwaj, "AI Chronic Diseases Preventive Care: Integrating Electronic Health Records, Genomic Data, and Real-Time Patient Monitoring with AI for Enhanced Early Detection of Chronic Diseases and Optimization of Peptide Drug Manufacturing," in International Conference of Global Innovations and Solutions, Springer, 2025, pp. 424–434.
[19] M. Lentzen et al., "A transformer-based model trained on large scale claims data for prediction of severe COVID-19 disease progression," IEEE J. Biomed. Health Inform., vol. 27, no. 9, pp. 4548–4558, 2023.
[20] Y. K. Ahmed and A. N. A. Naji, "Smart feature extraction using deep learning for early diagnosis of chronic diseases in next-generation medical decision support systems," Network Modeling Analysis in Health Informatics and Bioinformatics, vol. 14, no. 1, p. 140, 2025.
[21] L. Dai, H. Xu, and Y. Zhang, "Automated classification of clinical diagnoses in electronic health records using transformer," PLoS One, vol. 20, no. 9, p. e0329963, 2025.
[22] L. Gerrard, X. Peng, A. Clarke, and G. Long, "Claimsformer: Pretrained Transformer for Administrative Claims Data to Predict Chronic Conditions," in Australasian Joint Conference on Artificial Intelligence, Springer, 2024, pp. 348–362.
[23] A. Shmatko et al., "Learning the natural history of human disease with generative transformers," Nature, pp. 1–9, 2025.
[24] H. M. Zangana, F. M. Mustafa, and M. Omar, "A Hybrid Approach for Robust Object Detection: Integrating Template Matching and Faster R-CNN," EAI Endorsed Transactions on AI and Robotics, vol. 3, 2024.
[25] H. M. Zangana and F. M. Mustafa, "Hybrid Image Denoising Using Wavelet Transform and Deep Learning," 2024.
[26] A. Gupta, "Improved hybrid preprocessing technique for effective segmentation of wheat canopies in chlorophyll fluorescence images," EAI Endorsed Trans. AI Robot., vol. 3, 2024.
[27] S. Mishra and R. K. Dwivedi, "Designing Automation for Pickup and Delivery Tasks in Modern Warehouses Using Multi Agent Path Finding (MAPF) and Multi Agent Reinforcement Learning (MARL) Based Approaches," EAI Endorsed Transactions on AI and Robotics, vol. 3, 2024.
[28] E. Aghajari and A. A. AbdulRahim, "Prediction of short circuit current of wind turbines based on artificial neural network model," EAI Endorsed Trans. AI Robot, vol. 3, 2024.
[29] A. Mohamed, R. AlAleeli, and K. Shaalan, "Advancing Predictive Healthcare: A Systematic Review of Transformer Models in Electronic Health Records," Computers, vol. 14, no. 4, p. 148, 2025.
[30] S. Kaur and H. Sharma, "Causal Representation Learning for Predicting Autoimmune Disease Progression from Longitudinal Multimodal Clinical Data," IEEE Access, 2025.