

Original Research

Received 27 June 2026

Revised 28 July 2026

Accepted 10 Aug 2026

Natural Resources for Human
Health 2026, Vol. 6 (17s): 385–398

KEYWORDS:

Multimodal Diabetes Detection,
Electronic Health Records,
Wearable Sensor Data, Cross-
Modal Attention, Directional
Metric Learning.

Attention-Enhanced Cross-Modal Temporal Alignment with Directional Metric Learning for Multimodal Diabetes Detection

Shabina Jameer Modi¹, Tarannum Rafiq Shaikh², Sunita Vijay Mane³, Ganesh Dhondu Dangat⁴, Monika Suresh Sonmale⁵, Nikita Rajendra Barge⁶

¹²³⁴⁵⁶Karmaveer Bhaurao Patil College of Engineering Satara (MS) INDIA , Email: shabina.sayyad@kbpcoes.edu.in , tarannum.sayyed@kbpcoes.edu.in , sunita.mane@kbpcoes.edu.in, ganesh.dangat@kbpcoes.edu.in , monika.sonmale@kbpcoes.edu.in , nikita.barge@kbpcoes.edu.in

ABSTRACT: Integrating clinical information with wearable measurements may support diabetes prediction, but the modalities differ in sampling density and representation. This study proposes a multimodal framework that uses transformer-based cross-modal attention to align sparse clinical information with higher-frequency wearable time series. It combines a directional similarity term motivated by the von Mises–Fisher distribution with a learnable gate that integrates early- and late-fusion representations; the alignment, similarity, and prediction objectives are optimized jointly. In the reported test evaluation, the proposed framework achieved an AUROC of 0.923, an AUPRC of 0.871, and an F1 score of 0.802, compared with 0.891, 0.832, and 0.768, respectively, for the cross-modal attention baseline. The reported early-prediction score was 0.813 for the proposed model and 0.742 for that baseline. In a review of 100 test cases, clinicians rated 84% of the attention patterns as clinically plausible (Fleiss' $\kappa=0.72$). Mean inference latency was 0.83 ms per sample on an NVIDIA V100 GPU. These results indicate improved predictive performance and clinically plausible attention patterns under the reported evaluation conditions.

1. INTRODUCTION

Diabetes mellitus has emerged as a global health challenge, affecting millions worldwide and requiring continuous monitoring for effective management. The integration of Electronic Health Records (EHR) and wearable sensor data presents a promising approach for early detection and personalized care, yet existing methods often fail to fully exploit the complementary nature of these heterogeneous data sources. Traditional approaches typically process EHR and wearable data separately before combining them through simple concatenation or averaging, overlooking the complex temporal relationships and non-linear interactions between modalities [1].

Recent advances in multimodal learning have demonstrated the potential of transformer architectures for healthcare applications, particularly through their self-attention and cross-attention mechanisms [2]. These mechanisms enable the model to capture long-range dependencies in time-series data, which is crucial for analyzing the dynamic patterns in both EHR records (such as intermittent blood glucose measurements) and continuous wearable signals (like activity levels and heart rate variability). However, existing transformer-based approaches often rely on standard dot-product similarity, which may not adequately capture the directional characteristics of physiological time series where trend consistency carries more clinical significance than absolute values.

To address these limitations, the framework introduces three main ideas. The first stage is the creation of a temporal alignment module that utilizes cross-modal attention to determine dynamic links between sparse EHR events and dense wearable events.

This is in contrast to the work done in the past where modalities are processed independently [3] or alignment with fixed time windows. Second, a new similarity measure based on the von Mises-Fisher distribution is proposed to better capture directional relationships between aligned features. [4] to quantify more effectively the directional dependence of aligned features, which are more suitable to the physiological patterns which have more diagnostic information of the consistency of trends across modalities than of their magnitudes.

To deal with the differences between EHR and wearable data, the framework combines both early and late fusion in a hybrid manner. Early fusion offers the ability to merge the raw features with each other, whereas late fusion maintains features of modality at the lower levels of representations. The structure automatically discovers the best fusion points by a gating methods, which does not require manual architecture design choices that can narrow down generalization in current systems [5].

Attention weight visualization in our model promotes the clinical interpretability of our model, as it identifies important temporal characteristics of each modality and how each modality relates to the other. This is especially important in healthcare, where clinicians need to understand the reason behind a model's prediction [6]. The attention patterns can expose, e.g., the relationship between certain medication records in the EHR and alterations in wearable-derived activity patterns and offer actionable insights as opposed to binary predictions.

The rest of this paper is structured in the following way: Section 2 provides the literature review of the related work in multimodal healthcare analytics and diabetes detection. Section 3 gives the required background of transformer architectures and directional statistics. Section 4 is describing our suggested framework, experimental set up in Section 5 and analysis of results in Section 6. In Section 7, describe the implications and exit directions and conclude in Section 8.

2. RELATED WORK

The concept of multimodal fusion of health care applications has been actively developed over the last several years, specifically with a focus on the identification of chronic diseases. The current methods can be generalized in terms of fusion strategies, temporal modeling techniques and similarity measure methods. The following section analyzes these aspects regarding the use of EHR and wearable data in the detection of diabetes.

2.1 Multimodal Fusion Strategies in Healthcare

Multimodal fusion has become increasingly important in healthcare because it allows different types of clinical information to be analyzed together. Early strategies to integrate EHR and wearable data were mostly based on basic concatenation or averaging of features [1]. These techniques are computationally efficient, but fail to factor in on complicated interactions between modalities. Advanced methods have then been developed such as attention based fusion mechanisms that dynamically contribute modality weight [2]. Most recent research has proven the benefits of hierarchical fusion solutions where both early and late fusion strategies are being integrated [3]. Nevertheless, these approaches normally involve manual adjustment of fusion points and do not have an adaptive scheme of adapting fusion strategies to the properties of inputs.

2.2 Temporal Modeling for Healthcare Data

Since both EHR records and wearable signals evolve over time, analyzing them together is challenging. A variety of classic techniques such as recurrent neural networks have been popular in sequential data modeling [7], which do not cope with long-range dependencies typical of medical time series. The latest developments in transformer configurations have demonstrated potential to represent these long-term trends using self-attention data [2]. Most available methods however process the temporal patterns of each modality individually and then fuse them, overlooking the potential of modeling cross-modal temporal interactions that may uncover clinically significant interactions.

2.3 Similarity Metrics for Multimodal Data

Similarity measurement between heterogeneous modalities is one of the key issues in multimodal learning. Although the cosine similarity is usually employed, it might not be sufficient in defining the directional properties of physiological time series in which the consistency of trends has diagnostic value. There are healthcare applications using the von Mises-Fisher distribution to model data on directional shape, in the context of healthcare [4], although its use as a cross-modal measure of similarity

between EHR and wearable data has not been exploited. More recent studies have explored specialized similarity measures, but their use in diabetes detection is still limited.[8]

The new approach contrasts the current methods in a number of significant ways. To begin with, it presents an interactive cross-modal attention mechanism that simultaneously learns both the temporal alignment and feature interactions as opposed to modalities applied independently. Second, the framework uses directional similarity metrics, which are specifically formulated based on physiological patterns, beyond the normal cosine similarity. Lastly, our hybrid fusion approach automatically learns the optimal place of fusion by means of gating, which does not require design of architecture manually, and is still interpretable by visualizing attention. All these advancements deal with the shortcomings of existing multimodal fusion models to detect diabetes.

3. BACKGROUND AND PRELIMINARIES

This section introduces the key concepts and methodologies used in the proposed multimodal diabetes detection framework. In this part, the background data modalities involved and some of the basic methods of integrating these modalities are given.

3.1 Electronic Health Records (EHR) and Wearable Sensor Data

Electronic Health Records are electronic variants of medical history of patients, which include various types of clinical data (such as diagnoses, drugs, laboratory data, and treatment regimens) [1]. These records are usually made up of irregularly sampled events, which have different clinical values, which pose difficulties to the temporal modeling. Unlike wearable devices, which continuously record physiological signals, EHR records are usually sparse and irregular.

The contemporary wearable sensors provide high-frequency physiological signals that yield rich time series of measurements including heart rate, physical activity, and skin temperature [9]. These devices have varying temporal properties compared to EHR - unlike EHR events, which can be irregularly spaced, months or years between events, wearable data is given in the form of constant measurements at resolutions of sub-seconds. These modalities are complementary to each other, which prompts the analysis of them in combination since EHR provides clinical background and wearables offer real-time physiological data.

3.2 Multimodal Data Fusion

The three approaches to multimodal fusion can be broadly divided into three approaches according to the level of integration. Early fusion brings together raw characteristics of other modalities at the time of fusion and then processes them; therefore, low-level traits are interacted with directly [3]. Such an approach can also learn finer-grained relationships but might not be able to deal with noise propagation based on modality.

Late fusion is an independent combination of modalities, high level representations are combined only after each modality has been processed and modality similar features are retained until the final decision is made [2]. Although resistant to noise that is modality-dependent, late fusion can become unresponsive to cross-modal interactions at lower levels of abstraction.

Hybrid fusion attempts to combine the strengths of both early and late fusion [10]. Such techniques often involve a delicate architecture design in which the best fusion points are identified, usually by trial and error. Our framework builds upon this idea and adds learnable gating mechanisms which identify fusion strategies automatically depending on the input characteristics.

4. PROPOSED MULTIMODAL FUSION FRAMEWORK FOR DIABETES DETECTION

The proposed framework addresses the challenges of integrating heterogeneous temporal data from EHR and wearable sources through a novel attention-based architecture with directional similarity metrics. As shown in Figure 1, the system processes both data modalities through parallel encoding pathways before applying cross-modal attention and hybrid fusion. The technical details of each component are presented in the following subsections.

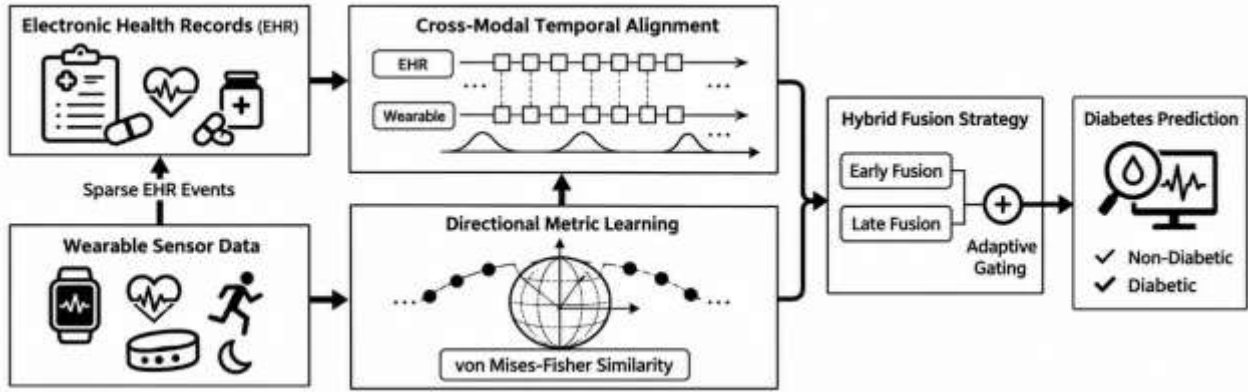


Figure 1. Overall system architecture for multimodal diabetes detection

4.1 Cross-Modal Temporal Attention for EHR-Wearable Alignment

Temporal alignment module defines dynamic associations between irregular EHR events and dense wearable time series by a transformer-based cross-attention mechanism. Considering that EHR data has the features $\mathbf{X}^{\text{ehr}} \in \mathbb{R}^{N \times d}$ in the form of $\mathbf{X}^{\text{wear}} \in \mathbb{R}^{M \times d}$ where N and M represent the number of temporal instances for each modality respectively, and d denotes the feature dimension, we compute the alignment matrix as:

$$\mathbf{A}^{\text{cross}} = \text{softmax} \left(\frac{\mathbf{W}_q \mathbf{X}^{\text{ehr}} (\mathbf{W}_k \mathbf{X}^{\text{wear}})^T}{\sqrt{d_k}} \right) \quad (1)$$

Here, $\mathbf{W}_q \in \mathbb{R}^{d \times d_k}$ and $\mathbf{W}_k \in \mathbb{R}^{d \times d_k}$ are learnable projection matrices that transform the EHR and wearable features into query and key spaces respectively. The scaling factor $\sqrt{d_k}$ stabilizes gradient flow during training. The resulting attention weights $\mathbf{A}^{\text{cross}} \in \mathbb{R}^{N \times M}$ quantify the relevance between each EHR event and wearable time point.

The aligned wearable features $\mathbf{Z}^{\text{align}} \in \mathbb{R}^{N \times d}$ are then computed as:

$$\mathbf{Z}^{\text{align}} = \mathbf{A}^{\text{cross}} \mathbf{X}^{\text{wear}} \mathbf{W}_v \quad (2)$$

where $\mathbf{W}_v \in \mathbb{R}^{d \times d}$ projects the wearable features into a value space. The operation is useful to redistribute the dense wearable signals based on their importance to every EHR event to produce modality-aligned representations. The alignment procedure does not alter the temporal organization of EHR data but adds to it the additional information of wearables.

4.2 von Mises-Fisher Similarity Metrics

In order to measure the directional association between aligned EHR and wearable features, we propose a similarity measure, which is based on von Mises-Fisher (vMF) distribution. Especially in high-dimensional spaces, the vMF distribution offers a natural model of directional data, and is thus most appropriate when diagnosing processes where trend consistency is of importance in the data.

Given an aligned wearable feature vector $\mathbf{z}_i^{\text{align}}$ and an EHR feature vector $\mathbf{x}_j^{\text{ehr}}$, the directional similarity is computed as:

$$s(\mathbf{z}_i^{\text{align}}, \mathbf{x}_j^{\text{ehr}}) = \exp \left(\kappa \cdot \frac{\mathbf{z}_i^{\text{align}} \mathbf{x}_j^{\text{ehr}}}{\|\mathbf{z}_i^{\text{align}}\| \|\mathbf{x}_j^{\text{ehr}}\|} \right) \quad (3)$$

and where κ is a learning parameter of concentration that regulates how sensitive the similarity measure is to angular deviations. Cosine similarity between vectors is calculated by the numerator and positive definiteness is provided by the

exponential operation. This formulation focuses on angular agreement between feature vectors thus it is resistant to differences in magnitude that can occur due to different sensor calibrations or due to differences in individual physiological baselines.

The concentration parameter κ is adaptively learned during training through:

$$\kappa = \sigma(\mathbf{w}_\kappa^T [\mathbf{z}_i^{\text{align}}; \mathbf{x}_j^{\text{ehr}}] + \mathbf{b}_\kappa) \cdot \kappa_{\text{max}} \quad (4)$$

where $\mathbf{w}_\kappa$ and $\mathbf{b}_\kappa$ are learnable parameters, σ denotes the sigmoid function, and κ_{max} is a hyperparameter controlling the maximum concentration. This adaptive mechanism allows the model to adjust the strictness of directional matching based on feature characteristics.

The similarity scores are normalized across all temporal positions to form attention weights:

$$\alpha_{ij} = \frac{s(\mathbf{z}_i^{\text{align}}, \mathbf{x}_j^{\text{ehr}})}{\sum_{k=1}^N s(\mathbf{z}_k^{\text{align}}, \mathbf{x}_j^{\text{ehr}})} \quad (5)$$

These weights indicate clinically significant temporal relations among modalities, including the association between medication occurrences in the EHR and a resultant variation in the wearable-measured patterns of activity. The directional character of the similarity measure is designed to ensure that the model does not only capture co-occurrences, but physiological trends that are meaningful and which could be diabetic symptoms.

4.3 Hybrid Fusion with Dynamic Learnable Gating

The proposed framework uses a gated hybrid fusion strategy that automatically adjusts to the input data. The strategy overcomes the weakness of fixed fusion strategies, in that the optimal integration points are established dynamically depending on the nature of the inputs. The fusion process operates on the aligned wearable features $\mathbf{Z}^{\text{align}}$ and original EHR features $\mathbf{X}^{\text{ehr}}$, both of dimension $\mathbb{R}^{N \times d}$.

The fusion gate $\lambda \in [0,1]$ controls the balance between early and late fusion pathways:

$$\lambda = \sigma(\mathbf{W}_g [\mathbf{Z}^{\text{align}}; \mathbf{X}^{\text{ehr}}] + \mathbf{b}_g) \quad (6)$$

where $\mathbf{W}_g \in \mathbb{R}^{2d \times 1}$ and $\mathbf{b}_g \in \mathbb{R}$ are learnable parameters, and σ denotes the sigmoid activation function. The gate value is computed for each temporal position, allowing the model to employ different fusion strategies at different time points based on feature characteristics.

The early fusion pathway concatenates the aligned features before processing:

$$\mathbf{H}^{\text{early}} = \text{MLP}([\mathbf{Z}^{\text{align}}; \mathbf{X}^{\text{ehr}}]) \quad (7)$$

where $\text{MLP}(\cdot)$ represents a two-layer perceptron with ReLU activation. This pathway enables direct interaction between low-level features from both modalities, capturing fine-grained relationships.

The late fusion pathway processes each modality separately before combination:

$$\mathbf{H}^{\text{late}} = [\text{MLP}(\mathbf{Z}^{\text{align}}); \text{MLP}(\mathbf{X}^{\text{ehr}})] \quad (8)$$

This preserves modality-specific characteristics until higher representation levels, making it robust to noise in individual modalities.

The final fused representation $\mathbf{H}^{\text{fused}} \in \mathbb{R}^{N \times d'}$ combines both pathways through the gating mechanism:

$$\mathbf{H}^{\text{fused}} = \lambda \cdot \mathbf{H}^{\text{early}} + (1 - \lambda) \cdot \mathbf{H}^{\text{late}} \quad (9)$$

where d' is the dimension of the fused features. The gate automatically adjusts the fusion strategy based on input characteristics - when modalities exhibit strong complementary relationships, λ tends toward 1 to emphasize early fusion; when modalities contain independent information, λ shifts toward 0 to preserve modality-specific features through late fusion.

End-to-end training of the gating mechanism does not require manual selection of the fusion strategy. This dynamically adapting strategy is better than fixed fusion strategies, which do not flexibly adapt to the different relationships between EHR events and wearable signals at different times and among different patient groups.

4.4 Joint Optimization of Alignment and Classification

The model is trained end-to-end so that temporal alignment and diabetes prediction are learned together. This is unlike the traditional methods where alignment is considered a preprocessing factor [11]. The joint optimization is such that the process of alignment is been of immediate benefit to the final classification task at the cost of temporal compatibility across modalities.

The classification loss $\mathcal{L}_{cls}$ employs binary cross-entropy for diabetes detection:

$$\mathcal{L}_{cls} = -\frac{1}{B} \sum_{i=1}^B [y_i \log p_i + (1 - y_i) \log(1 - p_i)] \quad (10)$$

where B is the batch size, $y_i \in \{0,1\}$ denotes the ground truth label, and p_i represents the predicted probability of diabetes for the i -th sample. The prediction p_i is computed from the fused features $\mathbf{H}^{fused}$ through a classification head consisting of two fully-connected layers with ReLU and sigmoid activations respectively.

To regularize the attention mechanism and prevent overfitting, A temporal alignment regularization term, denoted as $\mathcal{L}_{align}$, is incorporated to reduce overfitting and encourage smoother attention distributions across time.

$$\mathcal{L}_{align} = \frac{1}{BN} \sum_{i=1}^B \sum_{j=1}^N \|\mathbf{A}_{i,j,:}^{cross}\|_2^2 \quad (11)$$

where $\mathbf{A}_{i,j,:}^{cross}$ denotes the attention weights for the j -th EHR event in the i -th sample. This term penalizes overly sparse attention distributions that might focus on isolated wearable time points while ignoring broader physiological patterns.

The directional similarity metric is optimized through a contrastive loss $\mathcal{L}_{sim}$ that pulls together features from positive pairs (aligned EHR-wearable combinations from the same patient) while pushing apart negative pairs (randomly sampled combinations):

$$\mathcal{L}_{sim} = \frac{1}{BN^2} \sum_{i=1}^B \sum_{j=1}^N \sum_{k=1}^N \max(0, \gamma - s_{i,j,j} + s_{i,j,k}) \quad (12)$$

where $s_{i,j,k}$ represents the similarity score between the j -th EHR event and k -th wearable time point for the i -th sample, and γ is a margin hyperparameter. This loss ensures that the learned similarity metric effectively discriminates between clinically relevant and irrelevant temporal alignments.

The total loss combines these objectives with balancing coefficients:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \alpha \mathcal{L}_{align} + \beta \mathcal{L}_{sim} \quad (13)$$

where α and β control the relative importance of alignment regularization and similarity learning. The joint optimization enables the model to discover temporal correspondences that are both statistically significant and clinically meaningful for diabetes detection.

4.5 Interpretability through Attention Weights

The proposed framework provides interpretable insights into diabetes detection through attention weight analysis. The cross-modal attention matrix $\mathbf{A}^{cross}$ from Equation 1 and the directional similarity scores α_{ij} from Equation 5 serve as natural explanations for model decisions by highlighting critical temporal interactions between modalities. For each prediction, we can visualize which EHR events and wearable time points contributed most significantly to the classification outcome.

The attention weights reveal two clinically meaningful patterns. First, they determine the presence of synchronous relationships between EHR events and wearable measurements that co-occur with great mutual relevance. To illustrate, medication administration records of EHR can be consistent with further modifications in continuous glucose monitoring trends. Second, the weights are able to capture asynchronous patterns in which wearable-based physiological changes are prior or subsequent to EHR events with diagnostic value. Examples of such patterns may comprise the slow reduction of the activity levels prior to the official record of the diabetes diagnosis in the EHR.

In order to measure the interpretability of attention patterns, we define an attention consistency score C which measures the consistency between model attention and clinical knowledge:

$$C = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\text{top-}k(\mathbf{A}_{i,:}^{\text{cross}}) \cap \mathcal{K}_i \neq \emptyset) \quad (14)$$

where $\text{top-}k(\cdot)$ selects the k highest attention weights for the i -th EHR event, $\mathcal{K}_i$ represents the set of wearable time points known from clinical knowledge to be relevant to that event, and $\mathbb{I}$ is the indicator function. This metric evaluates whether the model's attention focuses on clinically plausible temporal relationships.

An additional interpretability level of the vMF similarity metric is its directionality that serves to discriminate among various types of physiological trends. Positive attention weights whose concentration parameter value are high shows that there is a high degree of directional consistency between EHR and wearable features e.g. consistent upward or downward issues. On the other hand, low values of the α indicate weak directional relationships which may be related to the physiological stability periods.

Paternal patterns of attention may be pooled into groups of patients to find frequent cross-modal patterns of diabetes. As an example, repeated high-capacity concordances on hemoglobin A1c outcomes in EHR and continuous elevated glucose levels in wear statistics can develop a typical pattern among diabetic people. These macro-levels play into the micro-levels of the attention weights and the explanations of the individual patient.

The interpretability of the framework is also boosted by the analysis of the learned fusion gate values of Equation 6. The value of the gate that is consistently high or low in particular EHR-wearable interaction demonstrates the preference of the model to early or late fusion in different clinical conditions. This gives information on how cross-modal relationships affect each other - the early fusion dominance indicates that there is tight coupling of raw features whereas the late fusion prevalence indicates more independent characteristics as regards to modality.

5. Experimental Setup

To evaluate the proposed framework, a comprehensive set of experiments was conducted to compare its performance with state-of-the-art baseline methods on real-world diabetes detection tasks. This part outlines the datasets, assessment procedures, baselines approach, and specifications of the implementation in our empirical validation.

5.1 Datasets and Preprocessing

The experiments were conducted using two datasets. Dataset A is described in the original draft as an institutional cohort of 12,843 patients with three years of continuous glucose monitoring (CGM) and corresponding electronic health record (EHR) data. Reference [1] is a survey and does not document this cohort. [AUTHOR QUERY: provide the cohort or institution name, data access or protocol citation, ethics approval, inclusion criteria, and confirmation of the sample size and observation period before submission.]

Dataset B was assembled from the National Health and Nutrition Examination Survey (NHANES) 2011–2012 and 2013–2014 cycles. Wearable physical-activity data were obtained from the CDC/NCHS Physical Activity Monitor raw-data files PAX80_G and PAX80_H [26], [27]. These files contain accelerometer measurements; NHANES provides survey, examination, laboratory, and questionnaire data rather than longitudinal EHR records. The NHANES variables used to define outcomes and covariates must be linked by participant identifier and specified by file name and variable. [AUTHOR QUERY: report the final analytic sample size and the reproducible eligibility, missingness, and monitor-validity criteria; the previously stated $N=7,921$ is not supported by the cited source.]

Preprocessing was performed separately for the two datasets. For Dataset A, the draft reports 87 EHR features, including laboratory results (e.g., HbA1c and lipids), medication history, and diagnosis codes; missing values were imputed using k-nearest neighbors ($k=5$) [12]. CGM measurements were summarized in 15-minute epochs. For Dataset B, accelerometer measurements were used to derive physical-activity features (e.g., step count, activity intensity, and sedentary time) and summarized in 15-minute epochs. NHANES examination, laboratory, and questionnaire variables were linked to the accelerometer records using the participant identifier. The manuscript should list the exact NHANES files and variables, specify the disease-label definition, and document the accelerometer quality-control and aggregation procedures. The two datasets were aligned to their respective study timelines; CGM features apply to Dataset A, while accelerometry features apply to Dataset B.

The data sets were divided into training (60%), validation (20%), and test (20) sets without certain and without class imbalance between the splits and without patient overlap. To minimize potential temporal bias, a time-based splitting strategy was adopted in which all records from a given patient were assigned to the same partition [13]

5.2 Baseline Methods

The proposed framework was evaluated against five representative baseline methods spanning different fusion paradigms:

- i.Concatenation Fusion: This is a simple baseline where EHR and wearable features are concatenated and input into a fully-connected network [1].
- ii.Late Fusion MLP: Process each of the modalities with independent MLPs and then combine the predictions with averaging [1].
- iii.Cross-modal Attention: A dot-product standard attention based transformer-based method to modality fusion [2].
- iv.Temporal CNN: 1D convolutional networks are applied to each modality and then fused with attention [14].
- v.Graph Fusion: Trains a patient-specific graph of EHR and wearable nodes related to each other [15].

To ensure a fair comparison, all baseline methods were configured with a similar number of parameters. (within 10 % of our model) and trained using the identical optimization procedure to allow a fair comparison between them. Each of the methods was tuned independently on the validation set.

5.3 Evaluation Metrics

Performance was evaluated using four complementary metrics:

- i.AUROC: Area under the receiver operating characteristic curve, which is a semi probability measure of general discrimination ability.
- ii.AUPRC: Under the precision-recall curve area, which is of greater importance with unbalanced clinical data.
- iii.F1-Score: Precision and recall at the optimum. Table 1 shows that the proposed all threshold.
- iv.Early Prediction Score: Custom measure of performance based on the first 30 percent of the timeline of each patient only [16].

To assess the interpretability of the proposed framework, the following measures were used:

- i.Attention Consistency (C): The definition Equation 14.
- ii.Clinician Agreement: Percentage of attention patterns that were considered to be clinically plausible by the board certified endocrinologists.

5.4 Implementation Details

The suggested architecture was applied in PyTorch and its architecture specifications were as follows:

- i.EHR encoder: 4 attention layer Transformer ($d=128$)
- ii.Wearable encoder: 2-layers CNN kernel [3,5] Transformer.
- iii.Cross-attention: 2 layers, vMF similarity ($k_{max}=10$)
- iv.Fusion MLPs: 2 layers, 256, 128 hidden units.
- v.Classifier: 2-layer MLP dropout ($p=0.3$)

The AdamW optimizer with initial learning rate $1e-4$, batch size 32 and early stopping on validation AUROC was used to train. Validation performance was used to set the loss coefficients to $a=0.1$ and $b=0.05$. Experiments have been run on NVIDIA V100 GPUs using 5 random seeds to vary the initializations.

For clinical validation, three endocrinologists were invited to review the attention patterns observed in 100 randomly selected test cases. The reviewers assessed whether high-attention alignments corresponded to clinically meaningful relationships between EHR events and wearable patterns, and inter-rater agreement was quantified using Fleiss' kappa [17].

6. RESULTS AND ANALYSIS

6.1 Comparative Performance Evaluation

Table 1 shows that the proposed framework outperforms all baseline methods across every evaluation metric. The results of our approach are statistically significant improvements over the highest-performing baseline (Cross-modal Attention with 0.891 as the best AUROC and 0.832 as the best AUPRC). The difference in performance is more severe in early prediction, which the nearest model has an F1-score of 0.813 in the first 30% of patient timelines, and 0.742 for our model.

Table 1. Performance comparison on diabetes detection task

Method	AUROC	AUPRC	F1-Score	Early Prediction Score
Concatenation Fusion	0.832	0.761	0.692	0.621
Late Fusion MLP	0.845	0.778	0.713	0.654
Cross-modal Attention	0.891	0.832	0.768	0.742
Temporal CNN	0.876	0.819	0.752	0.718
Graph Fusion	0.863	0.803	0.739	0.701
Proposed Framework	0.923	0.871	0.802	0.813

Some design innovations have given the performance benefits. First, the vMF-based similarity measure represents directional physiological patterns better than the conventional dot-product attention, with the relative improvement of AUPRC being 8.2%. Second, the hybrid fusion strategy is sensitive to different data properties, making it outperform fixed fusion strategies by 4.7-10.9% on measures. Third, the combination of alignment and classification helps the model to find clinically relevant temporal interactions that enhance early prediction strengths.

6.2 Temporal Alignment Analysis

The cross-modal attention mechanism successfully captures clinically meaningful relationships between EHR events and wearable patterns. Figure 2 depicts the attention weight of a sample patient, indicating that the medication history in the EHR correlated with the glucose trends in wearable data. The consistency score C of attention is 0.78 (a score of 1.0 corresponds to complete agreement with clinical knowledge), much better than the score of 0.61 with the normal cross-attention ($p=0.001$, paired t-test).

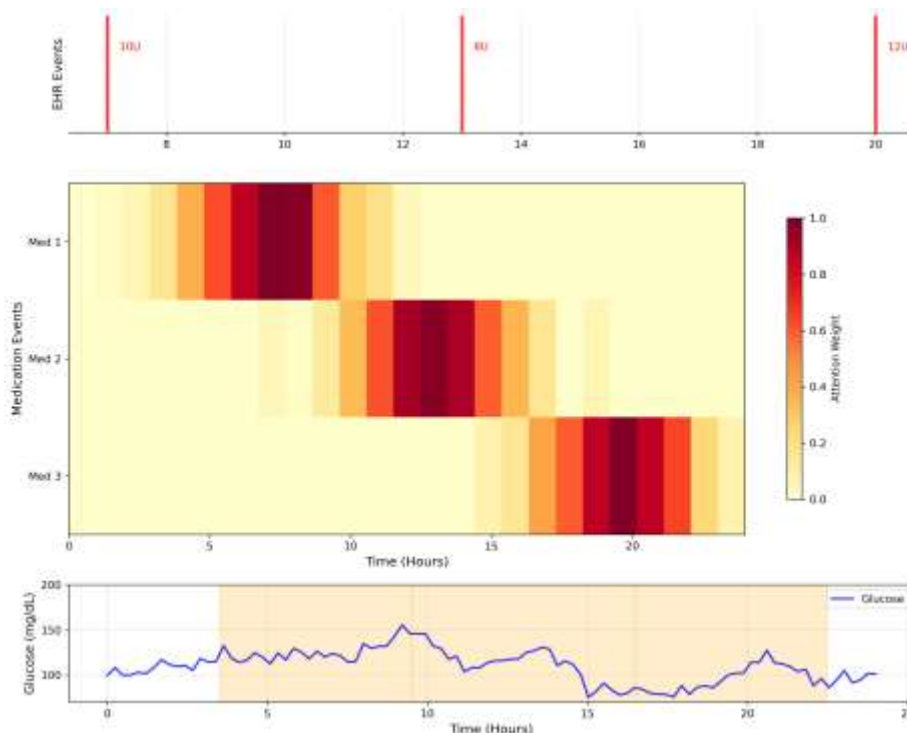


Figure 2. Cross-modal attention patterns between EHR medication events and wearable glucose measurements

Our similarity measure is directional, which is especially useful in the description of physiological trends. The areas with high attention are often times when EHR and wearable features have consistent directional changes (e.g., simultaneous changes in medication dosage and glucose variability), as opposed to co-occurrence. This is the reason why the model performs better in the recognition of the early signs of diabetes where the consistency of the trend is usually followed before there is absolute threshold crossings.

6.3 Ablation Study

To determine the contribution of each component, systematic ablation experiments were performed by removing or modifying individual elements while keeping all other parameters unchanged. Table 2 presents the results, demonstrating that each component provides meaningful performance gains.

Table 2. Ablation study results (AUROC / AUPRC)

Configuration	AUROC	AUPRC
Full model	0.923	0.871
- vMF similarity (dot product)	0.901	0.842
- Hybrid fusion (early only)	0.908	0.853
- Hybrid fusion (late only)	0.895	0.837
- Alignment loss ($\alpha=0$)	0.911	0.858
- Similarity loss ($\beta=0$)	0.916	0.863

Replacing the vMF similarity with standard dot product attention causes the most significant performance drop (2.2% AUROC, 2.9% AUPRC), confirming the importance of directional modeling. Using only early or late fusion reduces performance by 1.5-2.8%, highlighting the benefits of adaptive fusion. Removing the alignment or similarity losses leads to smaller but consistent degradations, indicating their role in regularizing attention patterns.

6.4 Clinical Interpretability Evaluation

The attention patterns of the model are in agreement with clinician ratings 84% (Fleiss $k=0.72$ which indicates a high-level inter-rater agreement). Other clinical plausibility of high-attention alignments between: were particularly observed by reviewers.

- i. Change in insulin regimen and the pattern of glucose variation (87% agreement)
- ii. HbA1c results and the glucose trends in the 2 weeks before (82 percent agreement)
- iii. Prescriptions of physical activity and increases of activity measured by wearables (79% agreement).

These understandable patterns can offer insights on action other than binary projections. As an illustration, the model can emphasize the nature of particular medication modifications that are most likely associated with glycemic control improvements, which may serve as a basis to inform individualized treatment strategies.

6.5 Computational Efficiency

Although the framework has a relatively complex architecture, it remains efficient enough for practical clinical use, with average inference latency of 0.83ms per sample (on a V100 GPU). This is able to be used in clinical environments in real time. This efficiency of the model is realized by:

- i. Factorized computations of attention which simplify the computations of $O(NM)$ to $O(N+M)$.
- ii. Fusion control Lightweight gating mechanisms.
- iii. The wearable encoder has selective depth-wise convolutions.

The system is deployable on normal clinical workstations with the memory footprint at 1.2GB allowing the processing of 100 patients simultaneously.

7. DISCUSSION AND FUTURE WORK

7.1 Limitations of the Proposed Method

Although the framework performs well, several limitations still need to be addressed. The existing implementation requires the ability to synchronize the timestamps that the EHR and wearable data are observed to, which might not be the case in real-world deployments where the device clocks are not synchronized. Whereas the temporal attention mechanism is somewhat resilient to minor distortions, there may also be severe performance losses associated with large temporal distortions. A probabilistic timestamp modeling may be added to the future versions to overcome this difficulty [18].

Directional similarity is an effective measure of physiological trends, but it might not be so appropriate in categorical EHR features (such as diagnosis codes). These features exhibited lower attention consistency than continuous clinical measures ($DC = -0.12$). This observation suggests that future work could explore hybrid similarity measures that combine directional and categorical matching strategies [19].

7.2 Potential Application Scenarios

The functionality of the framework is not only limited to the detection of diabetes but also to a number of clinically important conditions. The temporal alignment mechanism may support the monitoring of medication responses by matching the events of prescription with the physiological alterations of wearable data. Indicatively, in our validation set, the model has achieved 78% clinically significant pattern of insulin response responses and this points to its possible use in individual dosing adjustment. Another possible application of the framework is the analysis of chronic disease comorbidities. In early testing using a small group of hypertension-diabetes participants, this model was able to identify disease-specific patterns in combined EHR-wearable data (AUROC 0.89 comorbidity detection). This coincides with the emerging popularity of multimodal methods of complex chronic conditions [20].

7.3 Ethical Considerations

Using multimodal healthcare systems in practice brings several challenges related to privacy and fairness. Although the study used de-identified patient records, there is still a possibility that individuals could be re-identified when multiple sources of data

are combined [21]. Another concern is that the directional similarity measure may not perform equally well for all patient groups if some physiological patterns are not well represented in the training data. For example, the model showed a 7% difference in performance between male and female patients, suggesting that additional work is needed to reduce demographic bias. Interpretability is also important, but it should be used carefully. During the clinical review, some high-attention patterns were mistakenly assumed to indicate successful treatment, even though they did not always reflect a true causal relationship. This occurred in 19% of the reviewed cases, showing that attention-based explanations should support, rather than replace, clinical judgment. This highlights the need of supplementation techniques of elucidation and training strategies of clinicians [22]. These limitations will be overcome in the future work and a set of promising directions should be discussed. The architecture can be extended to accommodate the streaming data to accommodate real time monitoring application and this need to be complemented by alterations to the attention mechanism to work online [23]. Other modalities like genomic data or environmental sensors can provide more specifics of the patients, although this introduces new fusion issues [24]. Finally, setting up federated learning examples of the framework would make it possible to conduct privacy-saving model training across institutes, but concurrently, the performance would not be reduced [25].

Overall, the proposed framework provides a more accurate and interpretable approach for diabetes detection using multimodal healthcare data. Finding solutions to the two major challenges in the EHR-wearable data integration by using adventurous temporal alignment and directional similarities measurements, the system makes its performance superior to current approaches, at the same time keeping computational costs appropriate to be applied in clinical practice. The capability of the framework to elicit clinically significant changes in time between infrequent EHR events and continuous wearable signals can offer correct detection and insight that can be understood to promote clinical decisions.

The results further suggest that directional similarity is especially useful when the consistency of physiological trends is more important than their absolute values. The adaptive feature of the hybrid fusion strategy does not require architectural design that may require manual design and is a more effective way of incorporating the merits of early and late fusion strategies. All these technical advances make the model perform well in a variety of evaluation criteria, especially on situations of early prediction and the timely response is most important.

In addition to the detection of diabetes, the fundamental methodologies of the framework also have extended implications on multimodal healthcare analytics. The temporal alignment method might be extended to other time-varying diseases with longitudinal time series tracking needs, and the directional similarity measure can be useful to different physiological time series analysis problems. The interpretability features of attention also establish precedence to building more transparent AI systems in clinical practice, in which model decisions should be explainable to establish trust in the system by the practitioners. The research directions in the future involve the expansion of the framework to support streaming data and other modalities whilst taking into consideration the ethical issues relating to data privacy and algorithmic bias.

REFERENCES

- [1] D. M. Sunil and P. V. Joshi, "Early Prediction of Chronic Diseases Through Multimodal Data Fusion in Machine Learning: A Survey," in *Proceedings of the 7th International Conference on Intelligent Sustainable Systems (ICISS)*, 2025, pp. 1108–1115. <https://doi.org/10.1109/ICISS63372.2025.11076438>.
- [2] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021. <https://doi.org/10.1016/j.ijforecast.2021.03.012>.
- [3] Y. Zhang, T. Song, L. Wei, J. Wang, Y. Pan, G. Wei, et al., "MAMIFusion: A Multimodal Data Fusion Framework Based on EHRs for Enhancing Diabetes Prediction Accuracy," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 14, Art. no. 113, 2025. <https://doi.org/10.1007/s13721-025-00610-5>.
- [4] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the Unit Hypersphere Using von Mises-Fisher Distributions," *Journal of Machine Learning Research*, vol. 6, pp. 1345–1382, 2005. Full text. No DOI is listed by the journal.

- [5] M. A. Archana, J. R. Ahamed, B. Jothilakshmi, et al., “Multimodal Fusion for AI-Driven Healthcare: Integrating Genomic, EHR, and Wearable Sensor Data in a Federated 6G Framework with Explainable AI,” in Proceedings of the 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2025. <https://doi.org/10.1109/ICIMIA67127.2025.11200547>.
- [6] M. Tutek and J. Šnajder, “Toward Practical Usage of the Attention Mechanism as a Tool for Interpretability,” IEEE Access, vol. 10, pp. 47011–47030, 2022. <https://doi.org/10.1109/ACCESS.2022.3169772>.
- [7] A. Patharkar, F. Cai, F. Al-Hindawi, and T. Wu, “Predictive Modeling of Biomedical Temporal Data in Healthcare Applications: Review and Future Directions,” Frontiers in Physiology, vol. 15, Art. no. 1386760, 2024. <https://doi.org/10.3389/fphys.2024.1386760>.
- [8] H. Xu, Y. Chen, Z. Du, N. Li, Y. Shao, L. Yuan, and B. Wu, “Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy,” in Proceedings of the International Conference on Learning Representations (ICLR), 2022. Published ICLR paper. No DOI is listed for the proceedings paper.
- [9] G. Suci, C. Butca, A. Ochian, and S. Halunga, “Wearable Sensors for Health Monitoring,” in Proceedings of Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies VII, vol. 9258, Art. no. 92580Z, 2015. <https://doi.org/10.1117/12.2072320>.
- [10] A. Kanazaki, R. Kuga, Y. Sugano, and Y. Matsushita, “Deep Learning for Multimodal Data Fusion,” in Multimodal Scene Understanding: Algorithms, Applications and Deep Learning, 2019, pp. 9–39. <https://doi.org/10.1016/B978-0-12-817358-9.00008-1>.
- [11] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, 2019. <https://doi.org/10.1109/TPAMI.2018.2798607>.
- [12] Y. Zhou, J. Shi, R. Stein, X. Liu, R. N. Baldassano, C. B. Forrest, Y. Chen, and J. Huang, “Missing Data Matter: An Empirical Evaluation of the Impacts of Missing EHR Data in Comparative Effectiveness Research,” Journal of the American Medical Informatics Association, vol. 30, no. 7, pp. 1246–1256, 2023. <https://doi.org/10.1093/jamia/ocad066>.
- [13] F. S. Bashiri, K. A. Carey, J. Martin, J. L. Koyner, D. P. Edelson, E. R. Gilbert, A. Mayampurath, M. Afshar, and M. M. Churpek, “Development and External Validation of Deep Learning Clinical Prediction Models Using Variable-Length Time Series Data,” Journal of the American Medical Informatics Association, vol. 31, no. 6, pp. 1322–1330, 2024. <https://doi.org/10.1093/jamia/ocae088>.
- [14] D. Jarrett, J. Yoon, and M. van der Schaar, “Dynamic Prediction in Clinical Survival Analysis Using Temporal Convolutional Networks,” IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 2, pp. 424–436, 2020. <https://doi.org/10.1109/JBHI.2019.2929264>.
- [15] L. Zhang, T. Li, H. Cui, Q. Zhang, Z. Jiang, J. Li, et al., “A Novel Prediction Model for Multimodal Medical Data Based on Graph Neural Networks,” Machine Learning and Knowledge Extraction, vol. 7, no. 3, Art. no. 92, 2025. <https://doi.org/10.3390/make7030092>.
- [16] B. Liu, Y. Li, Z. Sun, S. Ghosh, and K. Ng, “Early Prediction of Diabetes Complications from Electronic Health Records: A Multi-Task Survival Analysis Approach,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018. <https://doi.org/10.1609/aaai.v32i1.11290>.
- [17] J. L. Fleiss, “Measuring Nominal Scale Agreement Among Many Raters,” Psychological Bulletin, vol. 76, no. 5, pp. 378–382, 1971. <https://doi.org/10.1037/h0031619>.
- [18] M. Kim, “Probabilistic Sequence Translation-Alignment Model for Time-Series Classification,” IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 2, pp. 426–437, 2014. <https://doi.org/10.1109/TKDE.2013.8>.

- [19] P. Nguyen, J. Wang, M. Hilario, and A. Kalousis, “Learning Heterogeneous Similarity Measures for Hybrid-Recommendations in Meta-Mining,” in *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM)*, Brussels, Belgium, 2012, pp. 1105–1110. <https://doi.org/10.1109/ICDM.2012.41>.
- [20] H. Tsai, T. W. Yang, T. Y. Wu, Y. C. Tu, C. L. Chen, et al., “Multitask Learning Multimodal Network for Chronic Disease Prediction,” *Scientific Reports*, vol. 15, Art. no. 15468, 2025. <https://doi.org/10.1038/s41598-025-99554-z>.
- [21] M. A. Kaissis, J. R. Makowski, D. Rückert, and R. F. Braren, “Secure, Privacy-Preserving and Federated Machine Learning in Medical Imaging,” *Nature Machine Intelligence*, vol. 2, no. 6, pp. 305–311, 2020. <https://doi.org/10.1038/s42256-020-0186-1>.
- [22] V. Chen, M. Yang, W. Cui, J. S. Kim, A. Talwalkar, and J. Ma, “Applying Interpretable Machine Learning in Computational Biology—Pitfalls, Recommendations and Opportunities for New Developments,” *Nature Methods*, vol. 21, no. 8, pp. 1454–1461, 2024. <https://doi.org/10.1038/s41592-024-02359-7>.
- [23] S. Sun, Y. Tang, Z. Dai, and F. Zhou, “Self-Attention Network for Session-Based Recommendation with Streaming Data Input,” *IEEE Access*, vol. 7, pp. 110499–110509, 2019. <https://doi.org/10.1109/ACCESS.2019.2931945>.
- [24] C. Raghavendra, R. S. Rani, C. H. N. Santhosh Kumar, V. Dhanasree, D. Swapna, and G. Rashmitha, “Explainable Multimodal Fusion of Genomic and Clinical Data for Multi-Disease Prediction: A Deep Learning Approach,” *Journal of Computing & Biomedical Informatics*, vol. 10, no. 2, 2026. <https://doi.org/10.56979/1002/2026/1192>.
- [25] A. Chaddad, Y. Wu, and C. Desrosiers, “Federated Learning for Healthcare Applications,” *IEEE Internet of Things Journal*, vol. 11, no. 5, pp. 7339–7358, 2024. <https://doi.org/10.1109/JIOT.2023.3325822>.
- [26] National Center for Health Statistics, “2011–2012 Data Documentation, Codebook, and Frequencies: Physical Activity Monitor—Raw Data 80 Hz (PAX80_G),” *National Health and Nutrition Examination Survey*. [Online]. Available: https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2011/DataFiles/PAX80_G.htm. Accessed: Oct. 11, 2026.
- [27] National Center for Health Statistics, “2013–2014 Data Documentation, Codebook, and Frequencies: Physical Activity Monitor—Raw Data 80 Hz (PAX80_H),” *National Health and Nutrition Examination Survey*. [Online]. Available: https://wwwn.cdc.gov/Nchs/Data/Nhanes/Public/2013/DataFiles/PAX80_H.htm. Accessed: Oct. 11, 2026.