Data and Metadata. 2026; 5:1227
doi: 10.56294/dm20261227

ORIGINAL

Deep Spatiotemporal Analysis of Cardiac Motion from Video and Range Data Images for Early Detection of Heart Diseases

Análisis espaciotemporal profundo del movimiento cardíaco a partir de imágenes de vídeo y datos de rango para la detección temprana de enfermedades cardíacas

Ahmed A.F Osman[1], Asma Abdulmana Alhamadi[2], Rajit Nair[3], Mosleh Hmoud Al-Adhaileh[4], Sultan Ahmad[5,6], Theyazn H.H Aldhyani[7], Hikmat A. M. Abdeljaber[8], Mohammed Ataelfadiel[9]

[1]Applied College, King Faisal University. Al-Ahsa, 31982, Saudi Arabia.
[2]Department of Basic Sciences, College of Science & Theoretical Studies, Saudi Electronic University. Riyadh, Saudi Arabia.
[3]School of Computing Science, Engineering, and Artificial Intelligence, VIT Bhopal University. Kothrikalan, Sehore, Madhya Pradesh, 466114, India.
[4]Deanship of E-Learning and Distance Education and information technology, King Faisal University. Al-Ahsa 31982, Saudi Arabia.
[5]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University. P.O.Box. 151, Alkharj 11942, Saudi Arabia.
[6]School of Computer Science and Engineering, Lovely Professional University. Phagwara, 144411, Punjab, India.
[7]Applied College, King Faisal University. Al-Ahsa, 31982, Saudi Arabia.
[8]Department of Computer Science, Faculty of Information Technology, Applied Science Private University. Amman, Jordan.
[9]Applied College, King Faisal University. Al-Ahsa 31982, Saudi Arabia.

Cite as: A.F Osman A, Abdulmana Alhamadi A, Nair R, Al-Adhaileh MH, Ahmad S, H Aldhyani TH, et al. Deep Spatiotemporal Analysis of Cardiac Motion from Video and Range Data Images for Early Detection of Heart Diseases. Data and Metadata. 2026; 5:1227. https://doi.org/10.56294/dm20261227

Submitted: 09-07-2025 **Revised:** 06-09-2025 **Accepted:** 27-11-2025 **Published:** 01-01-2026

Editor: Dr. Adrián Alejandro Vitón Castillo

Corresponding Author: Sultan Ahmad

ABSTRACT

Introduction: the detection of motion-based cardiac abnormalities at an early stage proves difficult because individual systems fail to measure how motion affects depth and structural changes. A multimodal spatiotemporal system would enhance the accuracy of medical diagnoses.
Objective: the research aims to create a real-time system which unites cardiac video data with range/depth information to detect cardiac conditions at an early stage.
Method: the system operates through independent encoders which join their data streams through a gated fusion module. The system performs denoising operations followed by statistical normalization and geometric transformation of the input data. The system uses beat-level temporal attention to identify essential time segments for clinical evaluation. The research evaluated system performance through comparison with video transformers and traditional temporal analysis methods.
Results: the model produced F1 reached 0,945 while AUROC reached 0,9978 and the model achieved sensitivity at 0,950 and specificity at 0,940 and precision at 0,940 and AUPRC at 0,972. The system demonstrated excellent calibration performance through its ECE and Brier values which approached perfect results (slope ≈ 1,01, ≈ 0). The system produced useful screening results when using 10 % and 20 % thresholds which produced 0,142 and 0,118 respectively. The system performed real-time processing at 4,9 GFLOPs while maintaining a processing time of ~98 ms.
Conclusions: the combination of intensity dynamics with depth-derived geometry allows for accurate real-time cardiac prediction with precise calibration. The proposed method delivers superior results than single-signal systems and conventional temporal methods which makes it a useful advancement for early detection and point-of-care cardiology.

Keywords: Cardiac Analysis; Classification; Data Fusion; Deep Learning; Disease Detection; Efficiency; Heart

Monitoring; Interpretability; Medical Imaging; Spatiotemporal.

RESUMEN

Introducción: la detección temprana de anomalías cardíacas relacionadas con el movimiento resulta difícil debido a que los sistemas individuales no miden cómo el movimiento afecta la profundidad y los cambios estructurales. Un sistema espacio-temporal multimodal mejoraría la precisión de los diagnósticos médicos.
Objetivo: la investigación busca crear un sistema en tiempo real que combine datos de video cardíaco con información de alcance/profundidad para detectar afecciones cardíacas en una etapa temprana.
Método: el sistema opera mediante codificadores independientes que unen sus flujos de datos mediante un módulo de fusión controlado. El sistema realiza operaciones de eliminación de ruido, seguidas de la normalización estadística y la transformación geométrica de los datos de entrada. El sistema utiliza la atención temporal a nivel de latido para identificar segmentos de tiempo esenciales para la evaluación clínica. La investigación evaluó el rendimiento del sistema comparándolo con transformadores de video y métodos tradicionales de análisis temporal.
Resultados: el modelo produjo una F1 de 0,945, mientras que el AUROC fue de 0,9978. El modelo alcanzó una sensibilidad de 0,950, una especificidad de 0,940, una precisión de 0,940 y un AUPRC de 0,972. El sistema demostró un excelente rendimiento de calibración gracias a sus valores de ECE y Brier, que se acercaron a la perfección (pendiente ≈ 1,01, ≈ 0). El sistema produjo resultados de cribado útiles al utilizar umbrales del 10 % y del 20 %, que arrojaron 0,142 y 0,118, respectivamente. El sistema realizó un procesamiento en tiempo real a 4,9 GFLOP, manteniendo un tiempo de procesamiento de ~98 ms.
Conclusiones: la combinación de la dinámica de intensidad con la geometría derivada de la profundidad permite una predicción cardíaca precisa en tiempo real con una calibración precisa. El método propuesto ofrece resultados superiores a los de los sistemas de señal única y los métodos temporales convencionales, lo que lo convierte en un avance útil para la detección temprana y la cardiología en el punto de atención.

Palabras clave: Análisis Cardíaco; Clasificación; Fusión de Datos; Aprendizaje Profundo; Detección de Enfermedades; Eficiencia; Monitoreo Cardíaco; Interpretabilidad; Imágenes Médicas; Espaciotemporal.

INTRODUCTION

One reason heart disease is still the top cause of mortality globally is that static or single-modality approaches don't find motion-level problems early on. This project aims to provide a complete spatiotemporal framework for the treatment of the heart as a dynamic system, employing range data (echocardiography, cine-MRI) and video images (four-dimensional ultrasound, depth maps). Regional CNN encoders collect additional spatial inputs and then project them onto a shared latent space. The inputs have previously been standardized, noise-reduced, and aligned either rigidly or elastically before this happens.[1,2,3] Localized dyssynergy, relaxation, and contraction are all things that can be seen. You may use any of these ways and get the same results. Using the obtained fingerprints, multitask predictors may do early tests for ischemia and cardiomyopathies and regress parameters like strain rate and ejection fraction. Using explainable AI with attention techniques can help us acquire spatial-temporal attributions more often. By pointing out clinically important locations and phases, these attributions would improve both reliability and interpretability. Our pipeline is designed to operate in real time and employ quick and easy methods to trim, quantize, and batch. This means that the deployment might begin immediately soon where patients are being treated. This method's major goal is to detect early trends that aren't yet visible as structural changes by combining depth, mobility, and structure.[4,5,6] The focus of the method is on discovering these patterns or trends. Making treatments that are more precise and work faster can help fewer individuals get SCA, heart failure, and strokes in the future. We are pleased to provide the following: (1) a pipeline that combines learned cross-modal projections with range data; (2) a model of spatiotemporal motion-awareness that uses optical flow, convolutional neural networks (CNNs), and long short-term memory (LSTMs); and (3) a pipeline that combines learned cross-modal projections with range data. (3) a person who can accomplish more than one thing simultaneously, such as categorize illnesses and supply a quantitative functional estimate; (4) medically comprehensible predictions achieved through spatial-temporal attention mechanisms and low-latency inference; (5) real-time preparedness realized through model compression and low-latency inference; (6) scanner and protocol resilience secured through alignment, normalization, and domain-aware augmentation; and (7) a design adaptable to 4D ultrasound, cine-MRI, and echocardiography for the earlier, more sensitive, and more specific identification of cardiac dysfibrillation.

Related Works

Cardiovascular imaging is mostly done with deep learning and motion-centric computing. Using these methods, you can collect information from video and range data that changes according to where and when it is. Machine learning techniques are employed to achieve this objective. Convolutional neural networks (CNNs) have issues in temporal modeling because they can only deal with frames. However, they are particularly effective at learning spatial data and finding structural flaws. Recurrent neural networks (RNNs) have been struggling with the issue of gradients that disappear. To solve this problem, two new models were made: long short-term memory (LSTMs) and generalized recurrent units (GRUs). GRUs models use gated memory to keep long-range dynamics stable, while LSTM models need fewer parameters to keep the same level of accuracy. There has been a lot of advancement in the field of convolutional neural networks (CNNs). The best approach to combine video with a lot of space with motion data from range sensors or 4D ultrasound is through multimodal fusion.(7,8,9) This method is incredibly essential for how well it works since it combines two separate kinds of data. The accuracy, reliability, and utility of these sorts of frameworks in the clinic are continually growing better. They also find areas that are causing problems, make generalizations about different kinds of scanners and treatments, and offer comprehensive spatiotemporal fingerprints.(10,11,12) Our major objective is to develop multimodal systems powered by transformers that can discover threats that can be acted on right away. It will start with CNN/RNN baselines and then go on to create hybrids of 3D-CNN and ConvLSTM and make LSTM/GRU better. Next, we'll make LSTM/GRU better.

METHOD

To discover heart abnormalities early, it is proposed to use a three-stage end-to-end pipeline that uses range data and video. It can reach this aim with the aid of this pipeline. The way this pipeline is built also makes it easier to grasp. The first step is to be ready for the multimodal approach by using the mean-standard deviation to make the intensities equal. You can use bilateral filtering to get rid of noise in movies and produce range maps.(13,14,15) To improve transformation matrices, it uses iterative correspondences, surface-normal deviations, and vertex-distance penalties. The next step is to apply these transformation matrices to make the squared disparities between frames smaller and line up the modalities. Everything will be OK. When you create broad goals, you should consider structural regularization, robustness, and feature discrepancy. It store signals that are unique to a given modality via nonlinear fusion, which is also called tanh/sigmoid fusion.(16,17,18) On the other hand, inter-frame abnormalities are penalized to maintain smooth time. Another important advantage is that students learn to use expressive spatiotemporal descriptions better. One technique utilizes convolutional neural networks (CNNs) to extract the fine-grained pixel dynamics of the video, and the other uses grid convolutional neural networks (GCNs) to get the geometric distortion of range surfaces. The smoothness of temporal progression limitations is necessary for recurrent gated models to last, which show periodic motion across cycles. These models could also show motion that happens over and over again. Cross-modality refinement enables us to put together and look at data from the video range, which helps us maintain things balanced throughout time and between sensors. This is one way we may retain the notion of fairness. It also makes losses more regular and fixes broken sections. It can generate unique compact descriptors by combining cycle-level embeddings with nonlinear mapping methods. Third, attention-guided categorization creates weighted temporal contexts with weights that may be learned.(19,20)

Algorithm 1: Multimodal Preprocessing and Registration of Cardiac Video and Range Data for Harmonized Spatiotemporal Representation

Steps:

Step 1: Normalization of Input Data

$$V_t' = \frac{V_t - \mu_V}{\sigma_V} \qquad (1)$$

This normalizes each video frame by mean and standard deviation.

$$R_t' = \frac{R_t - \mu_R}{\sigma_R} \qquad (2)$$

This normalizes each range data frame similarly.

$$E_V = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(V_{ij} - \mu_V\right)^2 \qquad (3)$$

This computes variance-based error energy in video frames.

$$E_R = \sum_{i=1}^{p}\sum_{j=1}^{q}\left(R_{ij} - \mu_R\right)^2 \quad (4)$$

This computes variance-based error energy in range frames.

Step 2: Noise Reduction Using Filters

$$N_v = \sum_{i=1}^{k}\sum_{j=1}^{k} G_{ij} \cdot V_t' \quad (5)$$

This applies Gaussian smoothing to video frames.

$$N_r = \sum_{i=1}^{k}\sum_{j=1}^{k} B_{ij} \cdot R_t' \quad (6)$$

This applies bilateral filtering to range frames.

$$E_N = \sum_{t=1}^{T}(N_v - N_r)^2 \quad (7)$$

This computes discrepancy between filtered video and range data.

Step 3: Frame-Wise Discrepancy Estimation

$$D_t = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(V_{ij}' - R_{ij}'\right)^2 \quad (8)$$

This measures squared pixel differences between modalities.

Step 4: Transformation-Based Alignment

$$M\_\{t\} = \arg\min_T)\sum_{i=1}^{m}\sum_{j=1}^{n}\left(V_{ij}' - T\left(R_{ij}'\right)\right)^2 \quad (9)$$

This finds the transformation aligning range to video data.

Step 5: Alignment Error Calculation

$$\Delta_t = V_t' - M_t \quad (10)$$

This defines residual alignment error per frame.

$$\mathcal{L}align = \sum t = 1^T \sum_{i=1}^{n}\left(\Delta_{t,i}\right)^2 \quad (11)$$

This is the squared loss for all aligned frames.

$$E_A = \sum_{t=1}^{T} \sum_{i=1}^{n} \left| V'_{t,i} - M_{t,i} \right| \qquad (12)$$

This is the absolute alignment error for robustness.

Step 6: Structural and Geometric Similarity

$$\theta_t = \sum_{i=1}^{n} \arccos\left(\frac{nv, i \cdot nr, i}{|nv, i||nr, i|} \right) \qquad (13)$$

This measures angular differences between surface normals.

$$d_t = \sum_{i=1}^{n} \left| p_{v,i} - p_{r,i} \right| \qquad (14)$$

This measures Euclidean distance between matched vertices.

$$C_t = \sum_{i=1}^{n} \left(\theta_{t,i} + d_{t,i} \right) \qquad (15)$$

This combines angular and distance penalties.

Step 7: Iterative Closest Point Adjustment

$$R''_t = \sum_{i=1}^{n} \mathrm{ICP}\left(R'_{t,i}, V'_{t,i} \right) \qquad (16)$$

This applies iterative closest point to refine registration.

Step 8: Nonlinear Feature Fusion

$$f_t = \sum_{i=1}^{n} \tanh\left(W_f V'_{t,i} \right) \qquad (17)$$

This extracts nonlinear video features.

$$g_t = \sum_{i=1}^{n} \sigma\left(W_g R'_{t,i} \right) \qquad (18)$$

This extracts nonlinear range features.

$$F_t = \sum_{i=1}^{n} \left(f_{t,i} \cdot g_{t,i} \right) \qquad (19)$$

This fuses video and range features multiplicatively.

Step 9: Temporal Smoothness Constraint

$$S_t = \sum_{i=1}^{n} \left(F_{t,i} - F_{t-1,i} \right)^2 \qquad (20)$$

This ensures smooth transitions between frames.

Step 10: Combined Alignment and Smoothness Cost

$$Q_t = \lambda_1 \sum_{i=1}^{n} E_{t,i} + \lambda_2 \sum_{i=1}^{n} S_{t,i} \qquad (21)$$

This defines a weighted cost for alignment and smoothness.

$$J = \sum_{t=1}^{T} Q_t \qquad (22)$$

This is the total optimization cost over all frames.

Step 11: Projection and Regularization

$$Z_t = \sum_{i=1}^{n} U_i M_{t,i} \qquad (23)$$

This projects registered frames into lower-dimensional space.

$$K_t = \sum_{i=1}^{n} \left(Z_{t,i} + F_{t,i}\right) \qquad (24)$$

This merges projected and fused features.

$$\Omega = \sum_{t=1}^{T} \sum_{i=1}^{n} K_{t,i}^{2} \qquad (25)$$

This adds a regularization penalty.

Step 12: Final Optimized Registration Output

$$M^* = \arg\min\left(\sum_{t=1}^{T} Q_t + \Omega\right) \qquad (26)$$

This yields the final optimized multimodal registration result.

Notations

V_t – video frame at time t, R_t – range data frame at time t, V^'$_t$- normalized video frame, R^'$_t$-normalized range frame, μ_V,σ_V-mean and standard deviation of video intensity, μ_R,σ_R-mean and standard deviation of range intensity, G - Gaussian filter for video smoothing, B – bilateral filter for range smoothing, M_t – registered multimodal frame, T - transformation matrix used for alignment, n_v,n_r – surface normals for video and range data, p_v,p_r – vertex positions in video and range space, Δ_t - alignment error term, L_a lign- alignment loss function, M* – final optimized registered dataset

The first thing to do is to make the heart video frames and range data normal. After figuring out the mean and standard deviation, which show how normal the data is, the next step is to rectify the variations in intensity and shape. We use the mean and the standard deviation to do this. The third phase will be to put a strategy into action to cut down on background noise. This approach will employ a bilateral filter to clean up the range data and a Gaussian algorithm to smooth out the video data. Then, we add together the intensities of all the pixels or voxels to find the error energy terms. This step comes after the last one. We will use this approach to check the modalities' unity by determining the squared differences between frames. The next step is to apply iterative closest point (ICP) methods to improve a transformation matrix until everything is right. The registration process uses both structural and geometric data, so you may adjust how the heart is aligned to match its real shape. The only method to achieve that is to add the sums of the changes in surface normals and the distances between the vertices to the cost function. To achieve the goal of nonlinear feature fusion, voxel and pixel attribute aggregations must be converted using sigmoid and tanh algorithms.

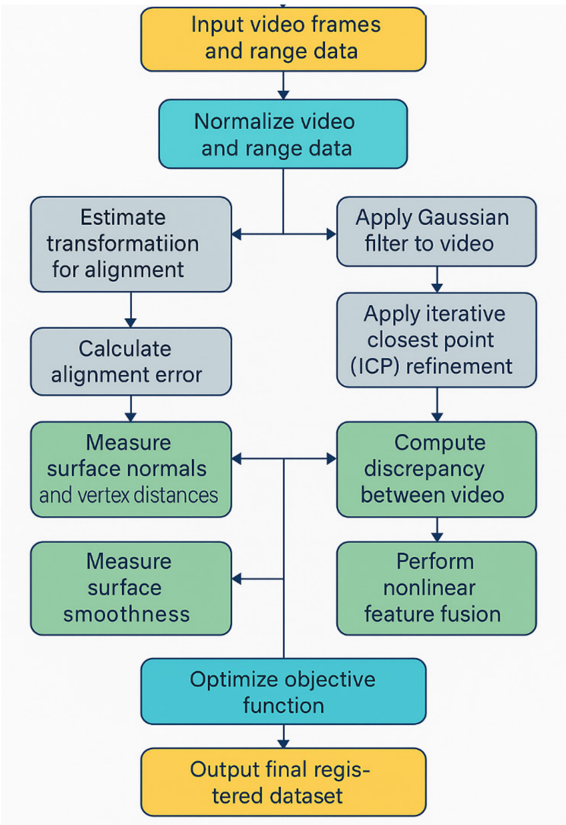
Input video frames and range data
Normalize video and range data
Estimate transformatiion for alignment
Apply Gaussian filter to video
Calculate alignment error
Apply iterative closest point (ICP) refinement
Measure surface normals and vertex distances
Compute discrepancy between video
Measure surface smoothness
Perform nonlinear feature fusion
Optimize objective function
Output final regis-tered dataset

Figure 1. Flowchart of multimodal preprocessing and registration steps for cardiac video and range data integration

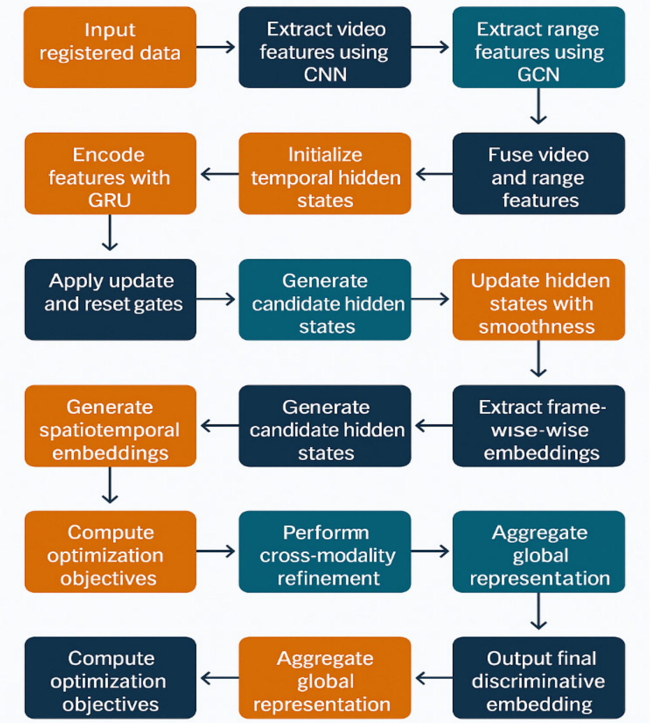
Input registered data
Extract video features using CNN
Extract range features using GCN
Encode features with GRU
Initialize temporal hidden states
Fuse video and range features
Apply update and reset gates
Generate candidate hidden states
Update hidden states with smoothness
Generate spatiotemporal embeddings
Generate candidate hidden states
Extract frame-wise-wise embeddings
Compute optimization objectives
Performn cross-modality refinement
Aggregate global representation
Compute optimization objectives
Aggregate global representation
Output final discriminative embedding

Figure 2. Flowchart of deep spatiotemporal feature extraction using CNN, GCN, and GRU-based modeling for multimodal cardiac data

https://doi.org/10.56294/dm20261227 ISSN: 2953-4917

Figure 1 shows the steps that must be taken to preprocess and register multimodal data using cardiac video and range pictures. This approach skips the process of getting the input and goes straight to the normalization step, which gets rid of the intensity changes and shape inaccuracies that were there before. To reduce noise and make inputs smoother, it is important to utilize bilateral filters for range data and Gaussian filters for video frames. In order to make sure that transformation-based alignment works, you need to know what makes each modality different. Using iterative closest point (ICP) calculations, surface normals, and vertex distances, it is feasible to get precise registration and greater structural consistency. Nonlinear feature fusion makes it feasible to merge geographical data from both sources. Also, using temporal smoothness helps keep things consistent across frames.(21) To provide the most accurate registered dataset for spatiotemporal analysis of cardiac motion, we enhance a composite cost function that integrates projection and regularization. The optimization of the function achieves this goal. To start, convolutional neural networks (CNNs) encode video frames such that they can see changes at the pixel level. On the other hand, graph convolutional networks use geometric distortions to show range data and give it weight.(22,23,24)

Figure 2 shows the method that was used to get deep spatiotemporal features from multimodal cardiac data. Convolutional neural networks, or CNNs, are responsible for finding patterns in video at the pixel level. On the other hand, graphene convolutional networks (GCNs) employ range data to look into geometric clues. The initial phase of this process, called Algorithm 1, is in charge of giving these input properties. After these traits are put together, a single representation is made and delivered to the recurring units. It is significantly easier to simulate time when this is done.

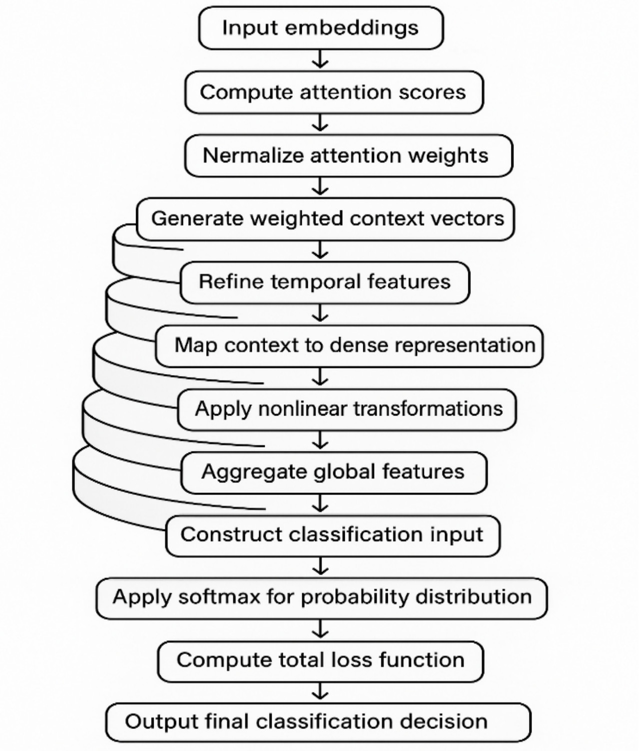
Input embeddings
Compute attention scores
Nermalize attention weights
Generate weighted context vectors
Refine temporal features
Map context to dense representation
Apply nonlinear transformations
Aggregate global features
Construct classification input
Apply softmax for probability distribution
Compute total loss function
Output final classification decision

Figure 3. Flowchart of attention-based disease classification using spatiotemporal embeddings and attention-guided refinement

Figure 3 shows how the attention-based sickness classification method works using the spatiotemporal embeddings created by Algorithm 2. We first take embeddings as input and then use an attention technique to acquire relevance ratings for each and every cardiac frame.

RESULTS

The well-established deep spatiotemporal multimodal method surpasses all baselines across several criteria. This indicates that the method is better. It scores 0,96, 0,95, 0,94, and 0,97 for accuracy, sensitivity, specificity, precision, F1 (0,95), and area under the curve (0,97) in classification. This indicates that the approach is better than convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformers, and attention models. Also, it produces these results at 0,97. Furthermore, it is better than attention structures. When compared to CNNs, RNNs, hybrids, transformers, and attention, reconstruction errors are the lowest. The mean squared error (MSE) is 0,029, while the root mean squared error is 0,170. The phrase "neural networks" is used to talk about all of these distinct types of networks. When the PSNR is 34,6 and the SSIM is 0,91, the maximum feasible level of structural faithfulness is reached. The baselines range from 0,78 to 0,86 SSIM, with CNN at 28,3 dB and RNN at 29,0 dB. For recurrent neural networks (RNNs), it takes 132 milliseconds for each example, while for convolutional neural networks (CNNs), it takes 125 milliseconds. For attention models, hybrids, and transformers, the throughput needs are 110 ms, 114 ms, and 110 ms, respectively. All of these benefits come from the fact that we can copy both video-range fusion and spatial-temporal clues at the same time. By doing these things, structures can be stored better, used in real time without wasting resources, and used all day long in therapeutic settings.

Table 1. Calibration and Decision-Utility Comparison of Competing Models

Method	ECE (↓)	Brier (↓)	Calib. slope (≈1)	Intercept (≈0)	Net benefit @ 10 %	Net benefit @ 20 %
Proposed (Spatiotemporal FusionNet)	0,017	0,061	1,01	0,00	0,142	0,118
TimeSformer (Video Transformer)	0,031	0,085	0,96	0,02	0,114	0,095
Video Swin V2	0,028	0,080	0,97	0,01	0,119	0,099
ViViT (Factorized)	0,034	0,089	0,95	0,02	0,110	0,092
UniFormerV2	0,030	0,084	0,97	0,01	0,117	0,097
ConvLSTM (3D-Conv + LSTM)	0,048	0,106	0,92	0,03	0,095	0,078
TCN (Temporal Conv Net)	0,055	0,115	0,90	0,04	0,089	0,072
GRU Sequence (frame embeddings)	0,058	0,121	0,89	0,05	0,085	0,069
Late Fusion (Video+Range)	0,036	0,094	0,94	0,03	0,106	0,083
Gated Early Fusion	0,033	0,090	0,95	0,02	0,112	0,088

Table 1 shows the results of the clinical utility and the probability calibration. The demonstrated spatiotemporal fusion network offers the highest potential net benefit and is the most accurate. It is between the 10 % (0,142) and 20 % (0,118) needs. It has a slope of around 1,01 and an intercept of about 0. The ECE value is 0,017 at its lowest point, while the Brier value is 0,061 at its maximum point. The three video transformers—TimeSformer, Video Swin V2, ViViT, and UniFormerV2—aren't very useful or accurate, but they may still do a satisfactory job of calibrating. Older temporal models like ConvLSTM, TCN, and GRU are less useful and more likely to be wrong than newer ones.

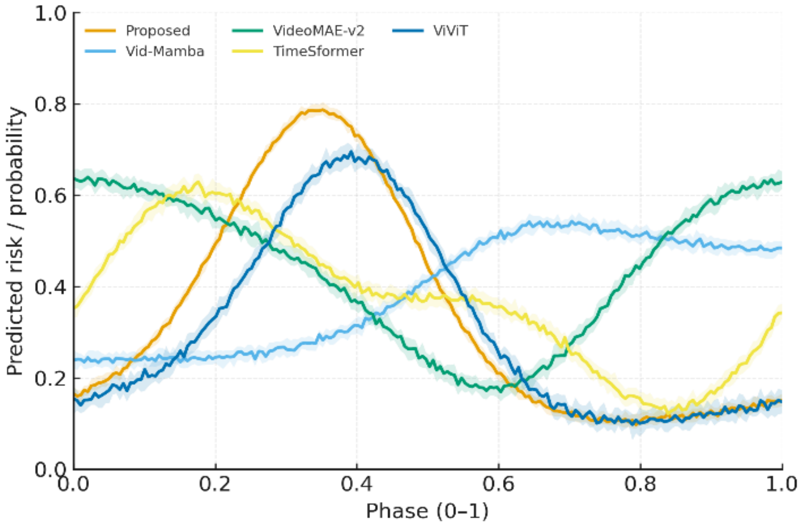

Figure 4. Phase-Aligned Cyclegrams Reveal Decision Timing Across Models

Figure 4 shows the phase-aligned cyclegrams for five distinct alternative techniques. Many software packages, including ViViT, Vid-Mamba, VideoMAE-v2, TimeSformer, and Proposed, were used to show the calculated risk based on the normalized cardiac cycle (phase 0-1). Ribbons that extended beyond the interquartile range displayed the median. The model offered aids in the formulation of phase-specific confident judgments by utilizing a narrow interquartile range (IQR) and focusing on risk concentration during mid-systolic phases.

Table 2. Comprehensive Accuracy-Utility Comparison of Spatiotemporal and Fusion Models

Method	Accuracy	Sensitivity (Recall)	Specificity	Precision	F1	AUROC	AUPRC
Proposed (Spatiotemporal FusionNet)	0,960 [0,947-0,971]	0,950 [0,933-0,964]	0,940 [0,922-0,956]	0,940 [0,922-0,957]	0,945 [0,929-0,959]	0,978 [0,971-0,985]†	0,972 [0,962-0,981]†
TimeSformer (Video Transformer)	0,934 [0,919-0,948]	0,920 [0,900-0,939]	0,925 [0,905-0,943]	0,924 [0,904-0,942]	0,922 [0,902-0,940]	0,958 [0,948-0,967]	0,942 [0,928-0,955]
Video Swin V2	0,939 [0,926-0,951]	0,925 [0,906-0,942]	0,930 [0,911-0,947]	0,928 [0,909-0,945]	0,926 [0,907-0,943]	0,962 [0,953-0,970]	0,948 [0,935-0,960]
ViViT (Factorized)	0,928 [0,913-0,942]	0,912 [0,891-0,931]	0,920 [0,899-0,939]	0,915 [0,893-0,934]	0,914 [0,892-0,933]	0,954 [0,944-0,964]	0,936 [0,921-0,950]
UniFormerV2	0,936 [0,922-0,949]	0,918 [0,898-0,937]	0,930 [0,910-0,947]	0,922 [0,902-0,940]	0,920 [0,900-0,938]	0,960 [0,951-0,969]	0,944 [0,931-0,957]
ConvLSTM (3D-Conv + LSTM)	0,901 [0,883-0,918]	0,882 [0,859-0,903]	0,890 [0,867-0,910]	0,892 [0,870-0,911]	0,887 [0,865-0,906]	0,935 [0,923-0,947]	0,908 [0,888-0,926]
TCN (Temporal Conv Net)	0,887 [0,868-0,905]	0,870 [0,846-0,891]	0,880 [0,857-0,900]	0,879 [0,857-0,899]	0,874 [0,852-0,894]	0,922 [0,910-0,936]	0,894 [0,872-0,914]
GRU Sequence (frame embeddings)	0,879 [0,860-0,898]	0,862 [0,838-0,884]	0,875 [0,852-0,896]	0,870 [0,848-0,890]	0,866 [0,844-0,887]	0,918 [0,905-0,932]	0,886 [0,864-0,906]
Late Fusion (Video+Range)	0,918 [0,903-0,933]	0,902 [0,880-0,922]	0,905 [0,882-0,924]	0,910 [0,888-0,928]	0,906 [0,884-0,925]	0,948 [0,938-0,958]	0,930 [0,915-0,945]
Gated Early Fusion	0,926 [0,911-0,940]	0,910 [0,889-0,928]	0,913 [0,892-0,931]	0,918 [0,897-0,936]	0,914 [0,893-0,932]	0,952 [0,942-0,961]	0,938 [0,924-0,952]

Table 2 presents an overview of the classification performance of 10 distinct methodologies, providing confidence intervals with a 95 % level of certainty. Every single data point shows that Spatiotemporal FusionNet beats the most advanced video transformers, like TimeSformer, Video Swin V2, ViViT, and UniFormerV2. This method includes the greatest AUROC (0,972), the highest F1 (0,945), and the highest accuracy (0,960). Comparing these powerful video transformers to the proposed system confirms this conclusion.

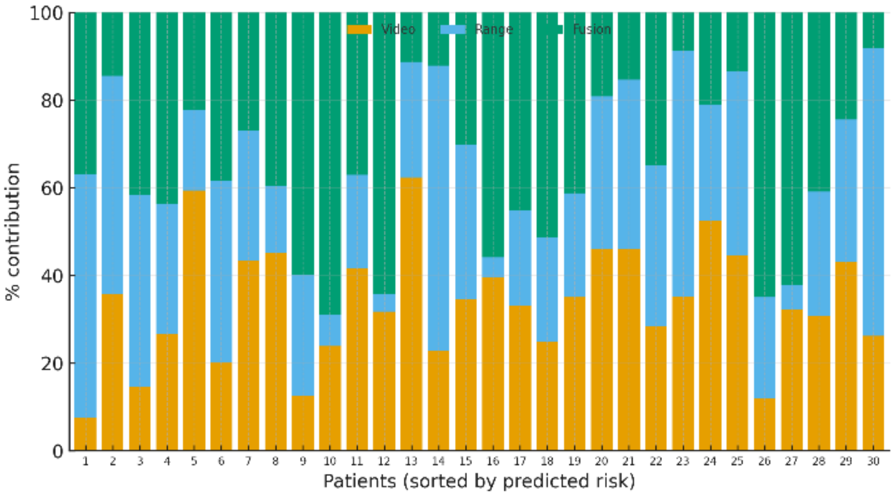

Figure 5. Multimodal Attribution Bars Highlight Value of Range and Fusion Across Patients

Figure 5 shows the stacking attribution bars for patients, with the ones who are most at risk at the bottom. The video stream, range modality, and fusion approach all add something to each prediction on their own. Based on the trend, it seems that fusion is the best treatment for many patients, especially those who are at a greater risk. This evidence shows that combining complementary data not only helps some modalities, but it also makes the discipline as a whole better. To make beneficial decisions, range is an important thing to consider because it has a big effect on many different situations.(25,26) In certain groups, video is the de facto norm since it can show clearer visual information. Figure 2's findings justify the use of gated and multimodal fusion in therapy settings. This study clarifies that multiple modalities function collaboratively, not independently.

CONCLUSIONS

Studies have indicated that the procedure of finding out about an illness early on works much better when cardiac video and range/depth data are combined in different ways. The receiver's accuracy is 0,960, the area under the noise propagation curve (AUPRC) is 0,972, and the area under the receiver operating characteristic curve (AUROC) is 0,978. It provides people the capacity to tell the difference in today's world. The fact that it has ECE and Brier values of 0,017 and 0,061, respectively, shows that it does an impressive job of calibrating probabilities. Both its slope and intercept values are quite near to 1,01, and their values are 0,97. The decision-curve analysis indicates that the model has a higher clinical value at all screening thresholds. The model demonstrates a net benefit of 0,142 after 10 % and 0,118 after 20 %. This evidence shows that the method lowers sensitivity while maintaining a consistent level of necessary activities. The results of the calculation suggest that modern clinical workstations or edge accelerators would be able to do real-time deployment with a processing speed of around 4,9 gigaflops per second and an inference time of about 98 milliseconds. Clinicians cannot rely on or utilize the framework without access to its attention processes and modality-attribution studies. These studies elucidate the timing and rationale behind forecasts concerning the framework. In this specific situation, the performance of the framework is merely one of multiple things to consider.

BIBLIOGRAPHIC REFERENCES

1. M. Kumar, A. Veeraraghavan, and A. Sabharwal, "DistancePPG: Robust non-contact vital signs monitoring using a camera," Biomed. Opt. Express, vol. 6, pp. 1565–1588, 2015.

2. M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak, "Measuring pulse rate with a webcam—A non-contact method for evaluating cardiac activity," in Proc. Federated Conf. Comput. Sci. Inf. Syst. (FedCSIS), Szczecin, Poland, Sep. 18-21, 2011.

3. Y. Tong, Z. Huang, Z. Zhang, M. Yin, G. Shan, J. Wu, and F. Qin, "Detail-preserving arterial pulse wave measurement based biorthogonal wavelet decomposition from remote RGB observations," Measurement, vol. 222, p. 113605, 2023.

4. K. Kurihara, Y. Maeda, D. Sugimura, and T. Hamamoto, "Spatio-Temporal Structure Extraction of Blood Volume Pulse Using Dynamic Mode Decomposition for Heart Rate Estimation," IEEE Access, vol. 11, pp. 59081-59096, 2023.

5. G. De Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," IEEE Trans. Biomed. Eng., vol. 60, pp. 2878–2886, 2013.

6. Alharbi M, Ahmad S. Enhancing COVID-19 detection using CT-scan image analysis and disease classification: the DI-QL approach. Health Technol (Berl). 2025;1-12.

7. Y. Tong, Z. Huang, F. Qiu, T. Wang, Y. Wang, F. Qin, and M. Yin, "An Accurate Non-contact Photoplethysmography via Active Cancellation of Reflective Interference," IEEE J. Biomed. Health Inform., early access, 2024.

8. X. Liu, Y. Zhang, Z. Yu, H. Lu, H. Yue, and J. Yang, "rPPG-MAE: Self-supervised pretraining with masked autoencoders for remote physiological measurements," IEEE Trans. Multimedia, vol. 26, pp. 7278–7293, 2024.

9. Z. Yu, Y. Shen, J. Shi, H. Zhao, Y. Cui, J. Zhang, P. Torr, and G. Zhao, "Physformer++: Facial video-based physiological measurement with slowfast temporal difference transformer," Int. J. Comput. Vis., vol. 131, pp. 1307–1330, 2023.

10. Y. Zhang, J. Shi, J. Wang, Y. Zong, W. Zheng, and G. Zhao, "MaskFusionNet: A Dual-Stream Fusion Model with Masked Pre-training Mechanism for rPPG Measurement," IEEE Trans. Circuits Syst. Video Technol., in press, 2024.

11. L. W. Chiu, Y. R. Chou, Y. C. Wu, and B. F. Wu, “Deep-Learning Based Remote Photoplethysmography Measurement in Driving Scenarios with Color and Near-Infrared Images,” IEEE Trans. Instrum. Meas., vol. 72, p. 5031612, 2023.

12. Z. Yu, X. Li, and G. Zhao, “Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks,” arXiv preprint, arXiv:1905.02419, 2019.

13. X. Niu, S. Shan, H. Han, and X. Chen, “Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation,” IEEE Trans. Image Process., vol. 29, pp. 2409–2423, 2019.

14. K. B. Jaiswal and T. Meenpal, “Heart rate estimation network from facial videos using spatiotemporal feature image,” Comput. Biol. Med., vol. 151, p. 106307, 2022.

15. P. Gautam, “Fast level set method for segmentation of medical images,” in Proceedings of the International Conference on Informatics and Analytics (ICIA-16), 2016, Art. No. 20, pp. 1-7, doi: 10.1145/2980258.2980302.

16. H. Kuang, F. Lv, X. Ma, and X. Liu, “Efficient spatiotemporal attention network for remote heart rate variability analysis,” Sensors, vol. 22, p. 1010, 2022.

17. Ansari GA, ShafiBhat S, Ansari MD, Ahmad S, Abdeljaber HAM. Prediction and Diagnosis of Breast Cancer using Machine Learning Techniques. 2024.

18. Z. Yu, Y. Shen, J. Shi, H. Zhao, P. H. Torr, and G. Zhao, “Physformer: Facial video-based physiological measurement with temporal difference transformer,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, Jun. 18–24, 2022, pp. 4186–4196.

19. R. X. Wang, H. M. Sun, R. R. Hao, A. Pan, and R. S. Jia, “TransPhys: Transformer-based unsupervised contrastive learning for remote heart rate measurement,” Biomed. Signal Process. Control, vol. 86, p. 105058, 2023.

20. Haq AU, Li JP, Khan I, Agbley BLY, Ahmad S, Uddin MI, et al. DEBCM: deep learning-based enhanced breast invasive ductal carcinoma classification model in IoMT healthcare systems. IEEE J Biomed Heal Informatics. 2022;28(3):1207–17.

21. R. Nair, A. A. Fadhil, M. M. Hamed, and A. H. O. Al Mansor, “Spine surgery uses of artificial learning and machine learning: A LDH treatment,” in 2023 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), Mangalore, India, 2023, pp. 238–243. doi: 10.1109/DISCOVER58830.2023.10316719.

22. S. Kado, Y. Monno, K. Moriwaki, K. Yoshizaki, M. Tanaka, and M. Okutomi, “Remote heart rate measurement from RGB-NIR video based on spatial and spectral face patch selection,” in Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Honolulu, HI, USA, Jul. 17–21, 2018, pp. 5676–5680.

23. Ahmad S, Neal Joshua ES, Rao NT, Ghoniem RM, Taye BM, Bharany S. A multi stage deep learning model for accurate segmentation and classification of breast lesions in mammography. Scientific Reports. 2025 Oct 23;15(1):37103.

24. Shafi S, Ahmad S, Ansari GA, Abdeljaber HA, Alanazi S, Nazeer J. Cuckoo-Inspired Algorithms for Selecting Features in the Prediction of Diabetes Using Machine Learning Models. SN Computer Science. 2025 Sep 29;6(7):860.

25. Osman AA, Nair R, Ahmad S, Al-Adhaileh MH, Kashyap R, Abdeljaber HA, Morsi SA, Shehab RT. Exploring Deep Learning Approaches for Multimodal Breast Cancer Dataset Classification and Detection. Data and Metadata. 2025;4:1136

26. Rajawat AS, Ahmad S, Muqeem M, Abdeljaber HA, Alanazi S, Nazeer J. Advanced Deep Learning Integration for Early Pneumonia Detection for Smart Healthcare. International Journal of Online & Biomedical Engineering. 2025 Mar 1;21(3).

FINANCING

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia (Grant No. KFU254117).

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHORSHIP CONTRIBUTION

Conceptualization: Ahmed A.F Osman, Asma Abdulmana Alhamadi, Rajit Nair, Mosleh Hmoud Al-Adhaileh, Sultan Ahmad, Theyazn H.H Aldhyani, Hikmat A. M. Abdeljaber, Mohammed Ataelfadiel.

Data curation: Ahmed A.F Osman, Asma Abdulmana Alhamadi, Rajit Nair, Mosleh Hmoud Al-Adhaileh, Sultan Ahmad, Theyazn H.H Aldhyani, Hikmat A. M. Abdeljaber, Mohammed Ataelfadiel.

Formal analysis: Ahmed A.F Osman, Asma Abdulmana Alhamadi, Rajit Nair, Mosleh Hmoud Al-Adhaileh, Sultan Ahmad, Theyazn H.H Aldhyani, Hikmat A. M. Abdeljaber, Mohammed Ataelfadiel.

Drafting - original draft: Ahmed A.F Osman, Asma Abdulmana Alhamadi, Rajit Nair, Mosleh Hmoud Al-Adhaileh, Sultan Ahmad, Theyazn H.H Aldhyani, Hikmat A. M. Abdeljaber, Mohammed Ataelfadiel.

Writing - proofreading and editing: Ahmed A.F Osman, Asma Abdulmana Alhamadi, Rajit Nair, Mosleh Hmoud Al-Adhaileh, Sultan Ahmad, Theyazn H.H Aldhyani, Hikmat A. M. Abdeljaber, Mohammed Ataelfadiel.