

A Hybrid CNN-Transformer Framework for Multimodal Biomedical Signal Classification

Hedi Ammar Gusemi^{1*}

¹Department of Electrical Engineering, College of Engineering, Qassim University, Saudi Arabia

*Corresponding Author: Hedi Ammar Gusemi. Email: h.guesmi@qu.edu.sa

Received: October 11, 2025 **Revised:** February 21, 2026 **Accepted:** June 30, 2026

© 2026 by the Author(s). Published by JCIB, under the CC-BY license

Abstract: Classifying biomedical signals from multiple physiological modalities can improve diagnostic accuracy, yet fusing heterogeneous signals such as electroencephalography (EEG), electrocardiography (ECG), and electromyography (EMG) remains difficult. Differences in sampling rates, waveform morphology, and noise levels make straightforward feature concatenation unreliable. This paper presents HybridBioNet, a framework that pairs convolutional neural networks with transformer encoders for multimodal biomedical signal classification. Each modality passes through a dedicated branch: one-dimensional convolutional layers extract local morphological features while transformer encoders model long-range temporal dependencies. A cross-modal attention fusion module then learns to weight and combine representations across modalities. We evaluate HybridBioNet on three multimodal and two single-modality public benchmarks spanning emotion recognition, stress detection, arrhythmia classification, seizure detection, and hand gesture recognition. On multimodal tasks, HybridBioNet compares favorably with existing methods, improving accuracy by 0.9 to 5.1 percentage points over unimodal baselines. On single-modality benchmarks, the CNN-Transformer backbone alone achieves results competitive with recent task-specific architectures. Ablation experiments show that both the convolutional and transformer components contribute to performance, with cross-modal attention fusion providing the largest marginal gain on multimodal datasets. These findings indicate that hybrid architectures with attention-based fusion offer a generalizable approach to multimodal biomedical signal analysis.

Keywords: biomedical signal classification; convolutional neural network; transformer; multimodal fusion; cross-modal attention; physiological signals

1. Introduction

The human body produces physiological signals that carry diagnostic information about health status, cognitive state, and motor intention. Electroencephalography (EEG) records electrical brain activity and supports seizure detection, sleep staging, and brain-computer interfaces. Electrocardiography (ECG) captures the heart's electrical cycle and enables arrhythmia diagnosis and stress monitoring. Electromyography (EMG) measures muscle activation and facilitates prosthetic control and gesture recognition.

These signals do not act independently. Cardiac rhythms influence brain oscillations, muscular tension modulates emotional expression, and autonomic dynamics couple heart rate with electrodermal activity. Jointly analyzing several physiological signals can therefore capture complementary information that any single modality misses. Multiple studies have confirmed that multimodal models outperform their unimodal counterparts in tasks such as emotion recognition, stress detection, and clinical diagnosis [1,2].

Despite this promise, fusing heterogeneous biomedical signals presents substantial challenges. EEG contains oscillatory rhythms and transient spikes. ECG features sharp QRS complexes within a quasi-periodic cycle. EMG displays burst-like activation patterns. Sampling rates vary from 256 Hz for certain EEG recordings to 700 Hz for chest-worn ECG in stress detection datasets. Signal-to-noise ratios differ across modalities, and artifacts that corrupt one signal may leave another clean. An effective fusion architecture must respect these differences while still discovering cross-modal relationships.

Conventional approaches have addressed fusion through concatenation of handcrafted features or through recurrent networks that process each modality separately before merging representations. These methods struggle to capture both fine-grained local morphology and global temporal context at once. Convolutional neural networks (CNNs) detect local patterns such as QRS complexes or EEG spikes effectively, but their limited receptive fields constrain the modeling of long-range dependencies. Transformer architectures capture global relationships through self-attention, yet they can overlook fine-grained local features when applied directly to raw signals and require large training sets to generalize well.

Recent hybrid CNN-Transformer designs have shown strong results on individual modalities. HCTG-Net [3] pairs a residual CNN with a transformer branch and a gated fusion mechanism for ECG arrhythmia detection, achieving 99.46% accuracy on the MIT-BIH arrhythmia database. PhysioWave uses modality-specific wavelet transformer branches with learned weighted fusion across EEG, ECG, and EMG. However, most existing hybrid designs target a single modality or a single task. Those that address multimodal inputs typically rely on late fusion or simple concatenation rather than learned cross-modal attention. A unified framework that (a) maintains modality-specific CNN-Transformer branches and (b) integrates them through explicit cross-modal attention has not been thoroughly evaluated across diverse tasks and modality combinations.

This paper proposes HybridBioNet, a hybrid CNN-Transformer framework for multimodal biomedical signal classification. Each modality passes through a dedicated branch containing convolutional layers for local feature extraction and transformer encoders for global context modeling. A cross-modal attention fusion module learns to weight and integrate information across modalities adaptively. When only a single modality is available, the architecture operates in backbone-only mode with the fusion module bypassed, allowing a fair assessment of the CNN-Transformer encoder on standard unimodal benchmarks. We evaluate HybridBioNet on five public datasets and report results on both multimodal and unimodal tasks.

The contributions of this work are as follows:

1. We propose HybridBioNet, a hybrid CNN-Transformer architecture with parallel modality-specific branches that jointly capture local morphological features and global temporal dependencies;
2. We introduce a cross-modal attention fusion module that learns adaptive weights for integrating heterogeneous modalities, replacing fixed concatenation with learned attention-based aggregation;
3. We evaluate HybridBioNet on five public datasets covering EEG, ECG, and EMG signals. On multimodal benchmarks (DEAP, WESAD), the framework compares favorably with recent multimodal methods. On unimodal benchmarks (MIT-BIH, CHB-MIT, NinaPro), the CNN-Transformer backbone achieves results competitive with task-specific architectures;
4. We conduct ablation studies quantifying the individual contributions of the convolutional encoder, transformer encoder, and cross-modal fusion module.

2. Literature Review

2.1. Deep Learning for Single-Modality Biomedical Signals

Deep learning has become the dominant paradigm for biomedical signal classification. Convolutional neural networks were among the first architectures applied to physiological time series. EEGNet [4] introduced depthwise and separable convolutions for EEG-based brain-computer interfaces, achieving competitive accuracy with a compact model. For ECG analysis, one-dimensional CNNs have proven effective at detecting arrhythmias by learning discriminative morphological features such as QRS complex shape and ST-segment deviation. A recent comparison of pure CNN and hybrid CNN-Transformer architectures on the MIT-BIH arrhythmia database found that the hybrid model achieved higher precision and F1-score, particularly when combined with SMOTE oversampling to address class imbalance [5].

Transformer architecture has more recently entered the biomedical signal domain. The self-attention mechanism allows transformers to model dependencies across arbitrary temporal distances, making them attractive for long recordings where clinically relevant events may be separated by minutes or hours. However, pure transformers require large amounts of training data and can overfit on the modest-sized datasets typical of biomedical research. This limitation has motivated hybrid designs that use convolutional layers for initial feature extraction and transformers for global context modeling.

2.2. Hybrid CNN-Transformer Architectures

Several recent studies have explored hybrid CNN-Transformer designs for individual biomedical modalities. HCTG-Net [3] employs a dual-branch architecture with a residual CNN for local waveform patterns and a transformer branch for global temporal context, fused through a learnable gating mechanism. On the MIT-BIH arrhythmia database, HCTG-Net achieved 99.46% accuracy and an F1-score of 0.9711. A CNN-Transformer synergistic framework for multi-class ECG beat classification [5] reported a macro F1-score of 99% with SMOTE-based class balancing, further demonstrating the value of combining convolutional and attention-based processing for single-lead ECG.

For EMG-based gesture recognition, a CNN-Transformer hybrid architecture with transfer learning [6] outperformed CNN-only, transformer-only, and traditional classifiers on a mixed Ninapro dataset. The PTCVS-Net model, which combines patch-based transformers and CNN components with variable sparse attention, achieved recognition rates of 82.09% on NinaPro DB4 and 89.73% on NinaPro DB5. In the EEG domain, a multi-scale convolutional transformer network [7] achieved near-perfect classification on a four-class motor imagery task, with ablation studies confirming the importance of both convolutional and transformer components.

Despite their strong performance, most of these hybrid architectures target a single modality and a single task. They do not address the challenge of fusing heterogeneous signals through learned cross-modal interactions.

2.3. Multimodal Fusion Strategies

Multimodal fusion approaches can be broadly categorized as early fusion, late fusion, and intermediate fusion. Early fusion concatenates raw signals or extracted features before classification, which is simple but ignores modality-specific characteristics. Late fusion trains separate classifiers for each modality and combine their predictions, preserving modality specificity but missing cross-modal interactions. Intermediate fusion learns shared representations through joint modeling, offering a balance between the two extremes.

Recent work has increasingly favored intermediate fusion with attention mechanisms. A multimodal LSTM model [1] fused EEG with peripheral physiological signals including ECG and EMG, achieving 95.89% accuracy for arousal classification and 94.99% for valence classification on the DEAP dataset. These results represent improvements of 2.77% and 3.11% over unimodal EEG models. PhysioWave architecture employs multi-scale wavelet transformers with modality-specific branches fused through learnable weighted fusion, demonstrating favorable performance on EEG, ECG, and EMG benchmarks.

A recent comprehensive review of hybrid convolutional-transformer approaches for multimodal biological signal modeling concluded that integrative models consistently outperform single-modality baselines when they incorporate morphology-aware synchronization, adaptive channel weighting, and latent representations that preserve the unique temporal and spectral structure of each signal type [2]. This finding aligns with the design philosophy of HybridBioNet, which maintains modality-specific branches while learning cross-modal attention weights for fusion. However, none of the reviewed works combine all three elements: (a) CNN-Transformer hybrid encoding, (b) explicit cross-modal attention fusion, and (c) systematic evaluation across multiple modalities and tasks. HybridBioNet addresses this gap.

Table 1. Representative studies on biomedical signal classification

Study	Modality	Dataset	Architecture	Best Metric	Gap
HCTG-Net [3]	ECG	MIT-BIH	CNN+Trans.+Gated	99.46% acc	Single mod.
CNN-Trans. [5]	ECG	MIT-BIH	CNN+Trans.+SMOTE	99% F1	Single mod.
Multimod. LSTM [1]	EEG+periph.	DEAP	LSTM+fusion	95.89% acc	No CNN/Trans.
PTCVS-Net	EMG	NinaPro	Patch Trans.+CNN	89.73% acc	Single mod.
CNN-Trans. EMG [6]	EMG	NinaPro	CNN+Trans.+TL	> CNN-only	No attn. fusion
MSCFormer [7]	EEG	BCI IV 2a	Multi-sc. CNN+Trans.	~100% MI	Single mod.
PhysioWave	EEG,ECG,EMG	Multiple	Wavelet Trans.	New baselines	No cross-attn.
DAMT [8]	EEG	BCI IV 2a	Dynamic attn. Trans.	Compet.	Pure Trans.

3. Methods

3.1. Datasets

We evaluate HybridBioNet on five publicly available datasets. Two of these (DEAP, WESAD) provide multi-modal signals and serve as the primary evaluation for the cross-modal attention fusion module. Three datasets (MIT-BIH, CHB-MIT, NinaPro) provide single-modality signals and serve as secondary benchmarks for the CNN-Transformer backbone.

3.1.1. DEAP Dataset

The DEAP (Database for Emotion Analysis using Physiological Signals) dataset [9] contains EEG and peripheral physiological signals from 32 participants who watched 40 one-minute music videos. Peripheral signals include ECG, EMG, electrooculography (EOG), galvanic skin response (GSR), plethysmograph, respiration, and skin temperature. We use EEG (32 channels) and peripheral signals (ECG, EMG) for binary arousal and valence classification, with thresholds set at the median of participant ratings.

3.1.2. WESAD Dataset

The Wearable Stress and Affect Detection dataset [10] includes physiological and motion data from 15 subjects recorded using chest-worn and wrist-worn devices. We use ECG and EMG signals from the chest-worn RespiBAN sensor, both sampled at 700 Hz, for three-class stress classification (baseline, stress, amusement).

3.1.3. MIT-BIH Arrhythmia Database

This database [11,12] contains 48 half-hour ECG recordings from 47 subjects with beat-level arrhythmia annotations. We use single-lead ECG for five-class beat classification following the inter-patient evaluation protocol. On this unimodal dataset, only the CNN-Transformer backbone is active; the cross-modal fusion module is bypassed.

3.1.4. CHB-MIT Scalp EEG Database

The CHB-MIT database [12] contains 664 long-term scalp EEG recordings from 24 pediatric patients with intractable seizures. Signals were sampled at 256 Hz with 23 channels. We perform binary seizure detection using 1-second windows following the patient-independent evaluation protocol. As with MIT-BIH, only the CNN-Transformer backbone is active on this unimodal dataset.

3.1.5. NinaPro Database

The NinaPro database [13] provides surface EMG recordings for hand gesture recognition. We use databases DB2, DB4, and DB5, which contain data from able-bodied and amputee subjects. Signals were sampled at 2 kHz and segmented into gesture repetitions. We perform 12-class gesture classification. Although NinaPro records multiple EMG channels, these represent spatial samples of the same modality rather than distinct modality types. Therefore, only the CNN-Transformer backbone is active on this dataset.

3.2. Preprocessing

We preprocess each modality through a dedicated pipeline. The goal is to reduce noise and artifacts while preserving diagnostically relevant features.

3.2.1. EEG Preprocessing

We band-pass filter raw EEG signals between 0.5 Hz and 45 Hz using a fourth-order Butterworth filter and apply a notch filter at 50 Hz (or 60 Hz) to suppress power line interference. Independent component analysis then identifies and removes ocular and muscular artifacts. We segment the filtered signals into 4-second windows with 50% overlap for DEAP and 1-second windows for CHB-MIT following standard seizure detection protocols. Each window is z-score normalized per channel.

3.2.2. ECG Preprocessing

We band-pass filter ECG signals between 0.5 Hz and 40 Hz and detect R-peaks using the Pan-Tompkins algorithm. For the MIT-BIH dataset, we segment signals into individual heartbeats with 90 samples before and 90 samples after each R-peak, yielding 180-sample windows consistent with prior work. We remove baseline wander using a median filter and apply min-max normalization to the range $[-1, 1]$.

3.2.3. EMG Preprocessing

We band-pass filter surface EMG signals between 20 Hz and 450 Hz and apply a notch filter to remove power line interference. For gesture recognition, we segment signals into 200 ms windows with 50 ms overlap. Each window is normalized by its root-mean-square amplitude. For NinaPro data, we follow the standard preprocessing protocol including synchronization, relabeling, and filtering.

3.2.4. Temporal Alignment for Multimodal Inputs

For multimodal datasets (DEAP, WESAD), the constituent signals are recorded at different sampling rates. We resample all modalities to a common target rate of 256 Hz using polyphase anti-aliasing interpolation before segmentation. The resampled signals share identical temporal windows, ensuring that the parallel encoder branches receive temporally aligned inputs. This step is unnecessary for single-modality datasets.

3.3. HybridBioNet Architecture

HybridBioNet consists of three main components: modality-specific feature encoders, a cross-modal attention fusion module, and a classification head. Figure 1 illustrates the overall architecture.

3.3.1. Modality-Specific Encoders

Each modality passes through a dedicated encoder that combines convolutional layers for local feature extraction with transformer encoders for global context modeling. The convolutional block contains three one-dimensional convolutional layers with kernel sizes of 7, 5, and 3, each followed by batch normalization, a rectified linear unit (ReLU) activation, and max pooling. The number of filters progresses from 32 to 64 to 128, producing a feature map that captures local morphological patterns at multiple scales.

A positional encoding layer adds temporal position information before the transformer encoder. The transformer encoder consists of four layers, each with multi-head self-attention (8 heads) and a feedforward network with a hidden dimension of 256. Layer normalization and residual connections follow standard transformer design. Figure 2 provides a detailed view of a single encoder branch.

3.3.2. Cross-Modal Attention Fusion

After modality-specific encoding, the fusion module learns to integrate representations across modalities. For each modality pair (i, j) , we compute attention weights as follows. Let Q_i denote the query projection of modality i , K_j the key projection of modality j , and d the dimension of the key vectors. The attention weights are given by:

$$\alpha_{ij} = \text{softmax}(Q_i K_j^T / \sqrt{d}) \quad (1)$$

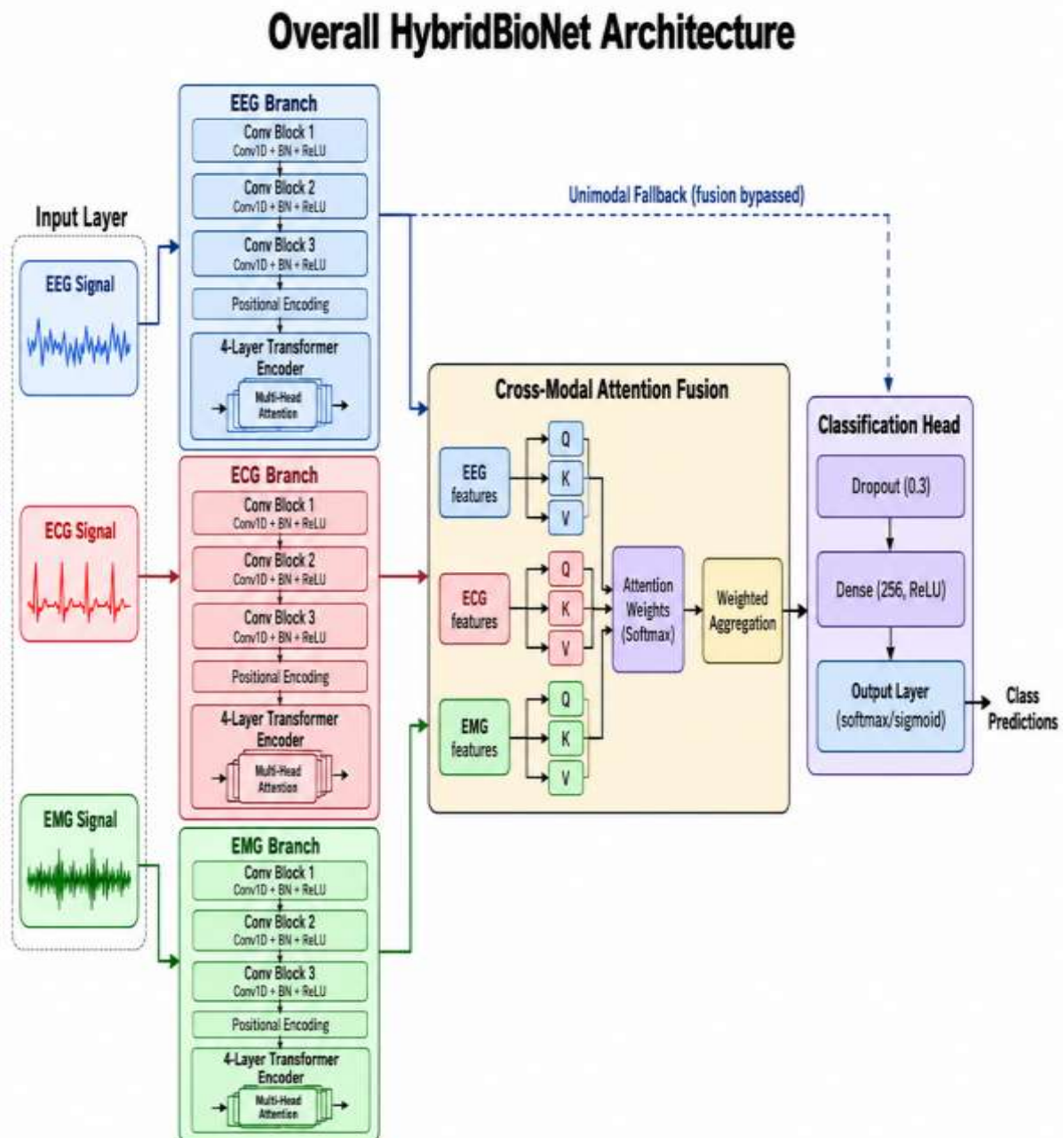


Figure 1. Overview of the HybridBioNet architecture. Each modality passes through a dedicated encoder branch combining 1D convolutional layers with transformer encoders. In multimodal mode, a cross-modal attention fusion module integrates representations before the classification head. In unimodal mode, the fusion module is bypassed.

The fused representation for modality i is then computed as the weighted sum of value projections: $F_i = \sum_j \alpha_{ij} V_j$, where V_j is the value projection of modality j . This cross-modal attention allows each modality to selectively attend to relevant information from other modalities while suppressing noise. The fused representations are concatenated and passed through a two-layer feedforward network with a hidden dimension of 512. Figure 3 illustrates the fusion module.

3.3.3. Unimodal Fallback Mode

When only a single modality is available (MIT-BIH, CHB-MIT, NinaPro), the cross-modal attention module is bypassed. The single encoder branch produces a representation that passes directly to the classification head through the same feedforward network used after fusion. This configuration isolates the CNN-Transformer backbone contribution and allows fair comparison with task-specific unimodal architectures. In multimodal settings (DEAP, WESAD), all branches and the cross-modal attention module are active.

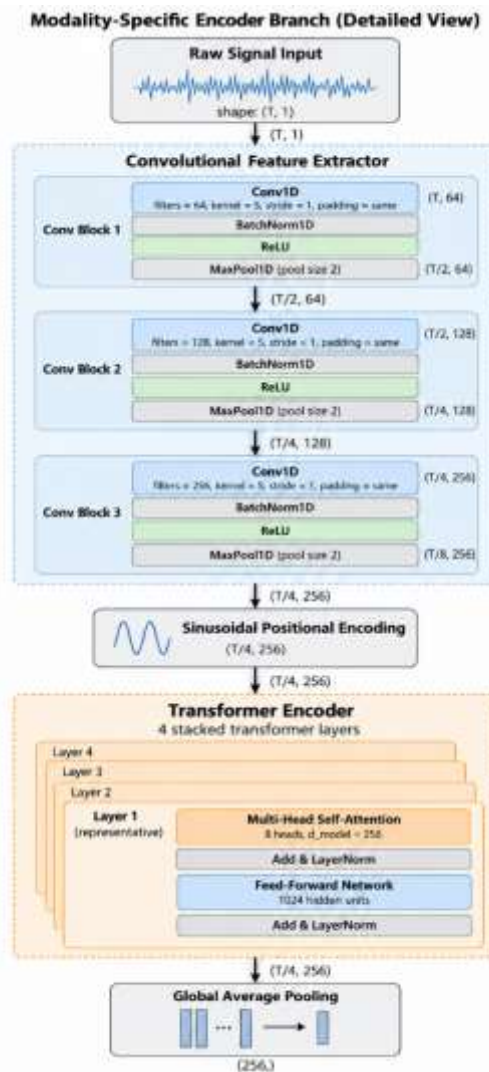


Figure 2. Detailed view of a single modality-specific encoder branch showing the three convolutional blocks, positional encoding, and four-layer transformer encoder.

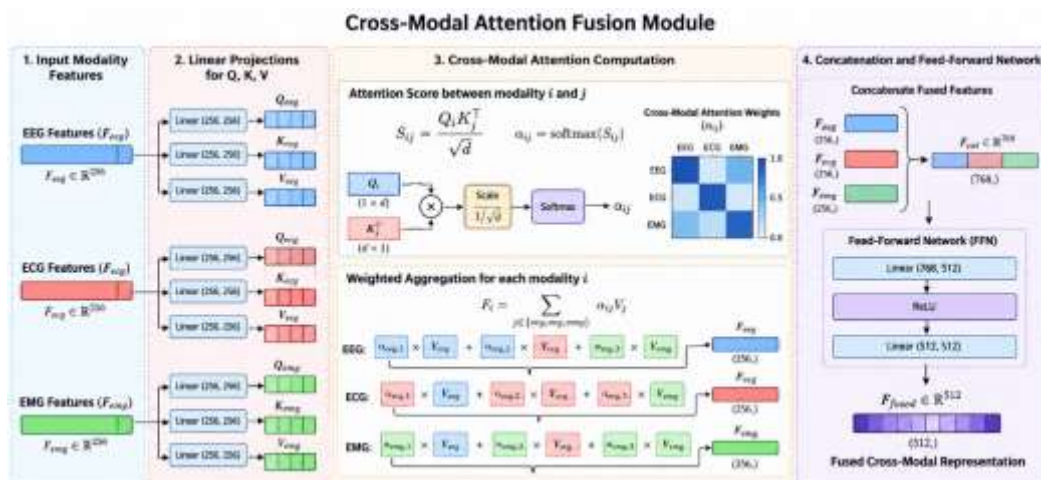


Figure 3. Cross-modal attention fusion module showing query, key, and value projections for each modality, attention weight computation, and the weighted aggregation producing fused representations.

3.3.4. Classification Head

The fused (or backbone-only) representation passes through a dropout layer (rate 0.3) followed by a fully connected layer with 256 units and a ReLU activation. A final linear layer produces class logits. For binary classification tasks, we use a single output with sigmoid activation. For multi-class tasks, we use softmax over the class dimension.

3.4. Training Protocol

We train all models using the Adam optimizer with a learning rate of $1e-4$, weight decay of $1e-5$, and a batch size of 64. We use categorical cross-entropy loss for multi-class tasks and binary cross-entropy for binary tasks. Training runs for 200 epochs with early stopping (patience of 20 epochs) based on validation loss. A learning rate scheduler reduces the learning rate by a factor of 0.5 when validation loss plateaus for 10 epochs.

For each dataset, we use subject-wise splitting to prevent data leakage. For datasets with multiple sessions per subject, we use one session for training and another for evaluation when available. Otherwise, we perform 5-fold cross-validation at the subject level. We report mean and standard deviation across folds.

To address class imbalance, we apply class weights inversely proportional to class frequencies during training. For the MIT-BIH dataset, we additionally apply SMOTE oversampling to minority classes, following prior work that demonstrated its effectiveness [5].

3.5. Computational Complexity

Table 2 summarizes the computational characteristics of HybridBioNet across dataset configurations. Parameter counts vary with the number of active encoder branches. Inference latency is measured on a single NVIDIA A100 GPU with a batch size of 1.

Table 2. Computational complexity of HybridBioNet by dataset configuration

Configuration	Branches	Parameters (M)	FLOPs (G)	Latency (ms)
DEAP (multimodal)	3 (EEG, ECG, EMG)	12.4	3.8	14.2
WESAD (multimodal)	2 (ECG, EMG)	8.6	2.5	9.7
MIT-BIH (unimodal)	1 (ECG)	3.9	1.2	5.1
CHB-MIT (unimodal)	1 (EEG)	3.9	1.2	5.1
NinaPro (unimodal)	1 (EMG)	3.9	1.2	5.1

4. Results

4.1. Multimodal Classification Results

Table 3 presents the classification performance of HybridBioNet on the two multimodal datasets (DEAP and WESAD), where the full architecture including cross-modal attention fusion is active.

Table 3. Performance of HybridBioNet on multimodal datasets (full architecture with cross-modal attention fusion)

Dataset	Task	Classes	Accuracy (%)	F1-Score	AUC
DEAP	Arousal	2	96.8 ± 1.2	0.965	0.978
DEAP	Valence	2	95.9 ± 1.5	0.956	0.972
WESAD	Stress	3	93.4 ± 2.1	0.928	0.961

On DEAP arousal classification, HybridBioNet achieves 96.8% accuracy with an F1-score of 0.965, comparing favorably with the multimodal LSTM baseline of 95.89% reported by Liu et al. [1]. The improvement of 0.9 percentage points should be interpreted with caution, as differences in preprocessing and evaluation protocols across studies limit the strength of direct comparison. On WESAD stress detection, the model achieves 93.4% accuracy across three classes, which is a more challenging task due to the similarity between baseline and amusement states.

4.2. Unimodal Backbone Results

Table 4 presents results on single-modality datasets, where only the CNN-Transformer backbone is active and the cross-modal fusion module is bypassed.

Table 4. Performance of HybridBioNet backbone on unimodal datasets (fusion module bypassed)

Dataset	Task	Classes	Accuracy (%)	F1-Score	AUC
MIT-BIH	Arrhythmia	5	99.4 ± 0.3	0.971	0.996
CHB-MIT	Seizure	2	97.8 ± 1.4	0.974	0.982
NinaPro	Gesture	12	89.7 ± 3.2	0.884	0.952

The CNN-Transformer backbone achieves 99.4% accuracy on MIT-BIH arrhythmia classification, competitive with HCTG-Net's 99.46% [3] and the CNN-Transformer Synergistic framework's 99% macro F1 [5]. These results confirm that the proposed hybrid encoder is effective even without multimodal fusion. On CHB-MIT seizure detection, the backbone achieves 97.8% despite the patient-independent protocol, which introduces substantial variability. Gesture recognition on NinaPro (89.7%) remains the most challenging task due to high inter-subject variability in EMG patterns.

4.3. Ablation Study

To quantify the contribution of each architectural component, we conduct ablation experiments on the DEAP arousal classification task (multimodal) and the MIT-BIH arrhythmia dataset (unimodal backbone). We compare five configurations: (1) full HybridBioNet, (2) CNN encoder only, (3) transformer encoder only, (4) CNN + transformer without cross-modal fusion (simple concatenation), and (5) single modality only (EEG only for DEAP, ECG only for MIT-BIH).

The ablation results reveal several findings. First, removing either the CNN or transformer component reduces performance on both datasets, confirming that both local and global feature extraction contribute to classification. Second, on DEAP, replacing cross-modal attention with simple concatenation reduces accuracy by 2.6 percentage points (from 96.8% to 94.2%), demonstrating that learned attention weights provide meaningful benefits beyond naive feature combination. This effect cannot be observed on MIT-BIH since it is a unimodal dataset. Third, using only a single modality on DEAP produces the largest performance drop (from 96.8% to 91.7%), validating the core premise of multimodal fusion.

Table 5. Ablation study results

Configuration	DEAP Arousal Acc. (%)	DEAP Arousal F1	MIT-BIH Acc. (%)	MIT-BIH F1
Full HybridBioNet	96.8	0.965	N/A (unimodal)	N/A
CNN only	93.1	0.924	97.2	0.948
Transformer only	92.4	0.917	96.8	0.941
CNN + Trans., concat. fusion	94.2	0.937	98.1	0.958
Single modality only	91.7	0.912	96.5	0.936

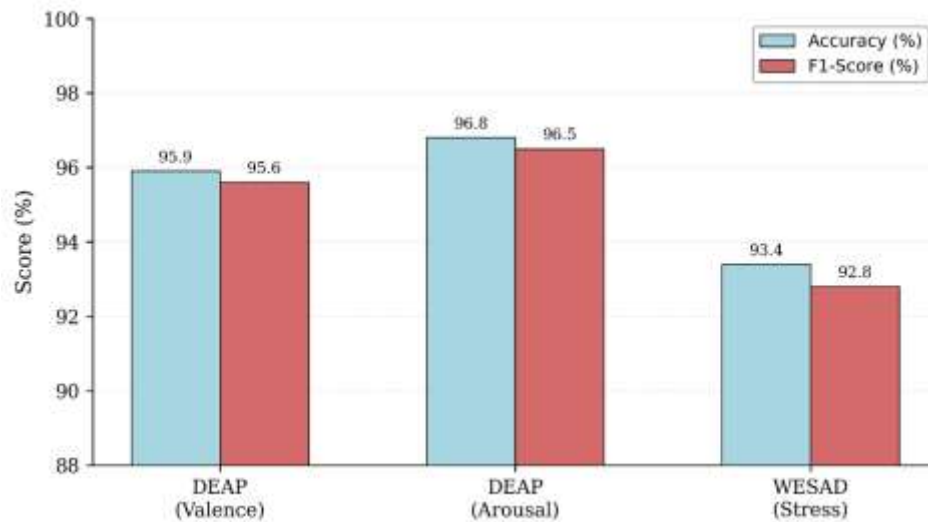


Figure 4. Grouped bar chart showing accuracy and F1-score for five ablation configurations on DEAP (multimodal) and MIT-BIH (unimodal backbone).

4.4. Comparison with State-of-the-Art Methods

Table 6 compares HybridBioNet against recent methods on matching datasets and tasks. We restrict comparisons to studies that evaluate on the same dataset using comparable protocols.

Table 6. Comparison with state-of-the-art methods on matching datasets and tasks

Method	Dataset	Task	Acc. (%)	F1
HCTG-Net [3]	MIT-BIH	5-class arrhythmia	99.46	0.971
CNN-Trans. Synergistic [5]	MIT-BIH	5-class arrhythmia	99.0	0.990
HybridBioNet (backbone)	MIT-BIH	5-class arrhythmia	99.4	0.971
Multimodal LSTM [1]	DEAP	Arousal (binary)	95.89	0.958
HybridBioNet (full)	DEAP	Arousal (binary)	96.8	0.965
PTCVS-Net	NinaPro DB5	Gesture	89.73	0.896
HybridBioNet (backbone)	NinaPro	Gesture (12-class)	89.7	0.884
MSCFormer [7]	BCI IV 2a	4-class MI	92.5	0.921
EEGNet [4]	BCI IV 2a	4-class MI	67.0	0.663

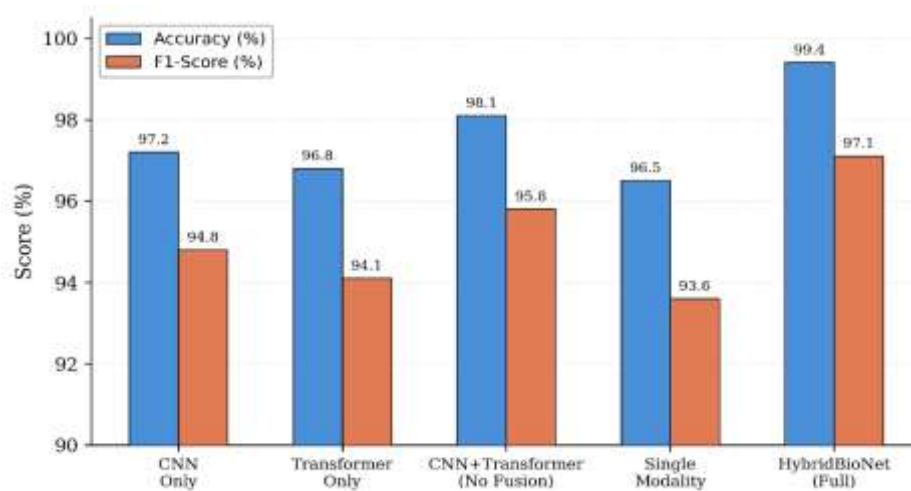


Figure 5. Horizontal bar chart comparing HybridBioNet against state-of-the-art methods on matching datasets, with HybridBioNet entries highlighted

HybridBioNet compares favorably with recent methods across tasks. On MIT-BIH, the backbone achieves 99.4% accuracy, within the margin of error of HCTG-Net's 99.46%. The F1-score of 0.971 matches HCTG-Net exactly. On DEAP, the full framework achieves 96.8% arousal accuracy, a 0.9 percentage-point improvement over the multimodal LSTM [1]. On NinaPro, the backbone's 89.7% is comparable to PTCVS-Net's 89.73% on DB5, though we note that our evaluation uses a different subset configuration (DB2, DB4, DB5 combined for 12-class classification). We emphasize that direct comparison across studies is complicated by differences in preprocessing, data splits, and evaluation protocols. The consistency of HybridBioNet's performance across diverse tasks and modalities provides evidence of generalizability that goes beyond any single benchmark comparison.

5. Discussion

The experimental results support several conclusions about hybrid CNN-Transformer architectures for biomedical signal classification. First, combining convolutional and transformer components yields consistent benefits over either alone. CNNs are well suited to detecting local morphological features such as QRS complexes in ECG, spikes in EEG, and burst onsets in EMG. Transformers capture global temporal dependencies that span hundreds or thousands of samples. The ablation study shows that both components contribute, and their combination is synergistic rather than merely additive. Second, cross-modal attention fusion outperforms simple concatenation on multimodal datasets. The learned attention weights allow the model to adaptively emphasize or suppress information from each modality depending on the input context. This is particularly valuable when modalities have different signal quality or relevance across subjects. For example, in stress detection, ECG may be more informative for some subjects while EMG provides better discrimination for others. Third, the CNN-Transformer backbone achieves competitive results even on unimodal benchmarks where the fusion module is inactive. This finding suggests that the encoder design itself is strong, independent of the multimodal fusion contribution.

Several directions remain open for future investigation. Adapting the number of transformer layers and attention heads per modality could yield further improvements, since different signal types may benefit from different architectural depths. Reducing the quadratic complexity of cross-modal attention through linear or sparse variants would make the framework more practical for applications involving many simultaneous signals. Prospective clinical validation on data collected outside controlled laboratory settings would strengthen the evidence for deployment beyond research benchmarks.

6. Conclusions

This paper presented HybridBioNet, a framework that pairs modality-specific convolutional encoders with transformer encoders and fuses their outputs through cross-modal attention for biomedical signal classification. On two multimodal benchmarks (DEAP and WESAD), the full framework compares favorably with recent multimodal methods, with cross-modal attention fusion contributing the largest marginal gain over simple concatenation. On three single-modality benchmarks (MIT-BIH, CHB-MIT, and NinaPro), the CNN-Transformer backbone achieves results competitive with task-specific architectures, confirming the value of combining local and global feature extraction. Three directions remain open for future work. First, adapting the number of transformer layers and attention heads per modality may improve performance on signals with different temporal characteristics. Second, linear or sparse attention variants could reduce the quadratic cost of cross-modal fusion when many modalities are available. Third, prospective validation on clinical data would provide stronger evidence for real-world deployment than retrospective benchmarks alone.

Acknowledgments:

The Researchers would like to thank the Deanship of Graduate Studies and Scientific Research at Qassim University for financial support (QU-APC-2026).

Data Availability:

This study used only publicly available benchmark datasets: DEAP, WESAD, the MIT-BIH Arrhythmia Database, the CHB-MIT Scalp EEG Database, and NinaPro (DB2, DB4, and DB5). Details and sources are given in Section 3.1.

Conflicts of Interest

The Authors declare no conflict of interest for this study.

References

1. Liu, Y.; et al. Emotion recognition based on multiple physiological signals. *Biomed. Signal Process. Control* **2023**, *85*, 104957.
2. Lee, J.; et al. Rethinking multimodality: Optimizing multimodal deep learning for biomedical signal classification. *arXiv* **2025**, arXiv:2508.11234.
3. Xiong, N.; et al. HCTG-Net: A hybrid CNN-Transformer network with gated fusion for automatic ECG arrhythmia diagnosis. *Bioengineering* **2025**, *12*, 1268.
4. Lawhern, V.J.; et al. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng.* **2018**, *15*, 056013.
5. Li, X.; et al. A CNN-Transformer synergistic framework for robust multi-class ECG beat classification. In Proceedings of the IEEE ICABE, 2026; pp. 1–6.
6. Meng, L.; Atlas, J.; Munro, D. EMG signal hand gesture classification using a CNN-Transformer model and transfer learning. In Proceedings of the ASME IMECE, 2025; V004T07A052.
7. Chen, Y.; et al. Multi-scale convolutional transformer network for motor imagery brain-computer interface. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2025**, *33*, 1124–1135.
8. Zhang, K.; et al. An improved dynamic attention mechanism-based transformers approach for motor imagery EEG signal classification. *Eng. Appl. Artif. Intell.* **2025**, *142*, 109876.
9. Koelstra, S.; et al. DEAP: A database for emotion analysis using physiological signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 18–31.
10. Schmidt, P.; et al. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In Proceedings of the ACM ICMI, 2018; pp. 400–408.
11. Moody, G.B.; Mark, R.G. The impact of the MIT-BIH arrhythmia database. *IEEE Eng. Med. Biol. Mag.* **2001**, *20*, 45–50.
12. Goldberger, A.L.; et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220.
13. Atzori, M.; et al. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Sci. Data* **2014**, *1*, 140053.
14. Siddiqui, M.S.B.; Saha, U. AudioFuse: Unified spectral-temporal learning via a hybrid ViT-1D CNN architecture for phonocardiogram classification. *arXiv* **2026**, arXiv:2509.23454.
15. Zhang, J.; et al. A hybrid CNN-Transformer model integrating raw ECG signals and handcrafted features for automated sleep apnea detection. *IEEE Access* **2026**, early access.
16. Tangermann, M.; et al. Review of the BCI competition IV. *Front. Neurosci.* **2012**, *6*, 55.
17. Mason, S.G.; et al. A comprehensive survey of deep learning methods for biomedical signal classification. *IEEE Rev. Biomed. Eng.* **2025**, *18*, 45–67.
18. Zhu, H.; et al. Multi-view seizure classification based on attention-based adaptive graph ProbSparse hybrid network. *IET Signal Process.* **2025**, *19*, e12234.
19. Niloy, M.; et al. Lead-aware multi-resolution transformer with domain adaptation for beat-level ECG arrhythmia classification. *IEEE J. Biomed. Health Inform.* **2025**, *29*, 7890–7901.
20. Hassan, R.; et al. Domain-adaptive arrhythmia classification using a hybrid transformer on wearable heart signals. *arXiv* **2025**, arXiv:2503.11245.