big data and cognitive computing

MDPI

Article

MATIC: A Lightweight Decision-Level Attention Model as a Strong Baseline for Multimodal Time-Series Classification

Rabin Nepal * and Bonny Banerjee

Institute for Intelligent Systems, and Department of Electrical & Computer Engineering, University of Memphis, Memphis, TN 38152, USA; bonnybanerjee@yahoo.com
* Correspondence: rnepal@memphis.edu

Abstract

Multimodal time-series classification has been successfully applied to diverse domains, yet recent research has increasingly favored complex feature-level fusion architectures whose computational cost often limits practical deployment. In this work, we investigate whether such complexity is necessary through a comprehensive cross-domain empirical study of decision-level fusion. We introduce MATIC (Multimodal Attention-based Time-series Classifier), a lightweight framework in which channel-specific convolutional neural networks independently classify each signal stream and a self-attention module adaptively fuses the resulting class-probability distributions. This modular formulation decouples feature extraction from multimodal fusion, enabling efficient training, reduced model complexity, and straightforward scalability across heterogeneous modalities. We evaluate the proposed approach on six public benchmark datasets spanning emotion recognition, human activity recognition, and stock movement prediction. Across all benchmarks, MATIC achieves competitive performance while requiring substantially fewer trainable parameters (0.11–1.95 M) than many specialized multimodal architectures. Comparisons with conventional decision fusion strategies and architectural ablations show that adaptive attention outperforms static and MLP-based fusion in most settings, and robustness analyses characterize its sensitivity to noisy and missing inputs. These findings suggest that lightweight decision-level attention fusion constitutes a strong baseline for multimodal time-series classification and provides a simpler alternative to increasingly complex multimodal fusion architectures.

Keywords: multimodal learning; time-series classification; decision-level fusion; self-attention; cross-domain generalization

Check for updates

Academic Editor: Nik Bessis

Received: 10 September 2026
Revised: 24 October 2026
Accepted: 4 November 2026
Published: 17 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Time-series data consist of sequences of observations indexed over time and are central to a wide range of tasks across domains such as finance [1], meteorology [2], transportation [3,4], healthcare [5,6], and others. In these applications, accurate decision-making often depends on the ability to recognize temporal patterns that evolve continuously over time. In practice, many real-world systems generate time series of unbounded or effectively infinite length, making it infeasible to process the entire sequence at once. As a result, time-series classification is commonly performed using a sliding window of fixed length, where local temporal segments are analyzed sequentially to enable timely and scalable predictions [7].

Despite the importance of multimodal time-series classification in real-world applications, many methods are designed under a unimodal assumption, where a single data

source or sensor stream is used for prediction. This limits their ability to capture complementary and correlated information that is often distributed across multiple heterogeneous signals [8]. In practical systems, such as healthcare monitoring, human activity recognition, or financial analysis, individual modalities may be noisy, incomplete, or insufficient on their own, making unimodal models fragile and less reliable.

Recent deep learning approaches have extended time-series classification to multimodal settings; however, many of these models are highly specialized for specific datasets or domains. They often rely on complex fusion mechanisms, large architectures, or handcrafted features, resulting in increased computational cost and reduced scalability [9]. Such designs hinder deployment in streaming and resource-constrained environments, where low latency, memory efficiency, and general-purpose applicability are critical. Moreover, the need to redesign model architectures for each new task limits their practicality as foundational solutions for multimodal time-series classification. Decision-level fusion offers a simpler, modular alternative by combining predictions from individual channels or modalities. However, conventional static and MLP-based fusion methods do not explicitly model sample-dependent relationships among channel predictions, motivating a systematic cross-domain evaluation of self-attention-based decision fusion.

Our Contributions

This work does not introduce a fundamentally new multimodal learning paradigm or model. Instead, it provides a rigorous empirical reassessment of decision-level fusion for multimodal time-series classification, which establishes a principled and reproducible baseline. The key contributions are as follows:

1. **A Lightweight and Modular Decision-Level Attention Fusion Framework.** We present a simple yet effective multimodal architecture—Multimodal Attention-based TIme-series Classifier (MATIC)—in which each channel is independently modeled using compact convolutional neural networks (CNNs), and their predicted class-probability distributions are fused through a self-attention mechanism. The design emphasizes modularity, parameter efficiency, and ease of scalability across heterogeneous modalities.
2. **A Systematic Cross-Domain Empirical Evaluation.** We conduct a comprehensive experimental study across multiple benchmark datasets spanning three distinct application domains: physiological-signal-based emotion recognition, wearable-sensor-based activity recognition, and financial time-series prediction. This cross-domain evaluation enables a robust assessment of generalization capability beyond a single task setting.
3. **Competitive Performance with Reduced Complexity.** Across all evaluated tasks, the proposed lightweight decision-level attention fusion approach achieves performance comparable to state-of-the-art multimodal models while using significantly fewer trainable parameters. These results demonstrate that complex feature-level fusion architectures are not always necessary to obtain competitive performance.
4. **Ablation and Comparative Analysis of Fusion Strategies.** We provide detailed ablation studies and comparisons against conventional static aggregation rules and MLP-based fusion methods, highlighting the practical advantages of attention-based decision fusion in terms of accuracy–complexity trade-offs.
5. **Practical Insights for Multimodal Time-Series Modeling.** Through consistent experimentation and analysis, we offer evidence that compact decision-level attention fusion constitutes a strong, computationally efficient, and reproducible baseline for heterogeneous multimodal time-series classification tasks.

https://doi.org/10.3390/bdcc10110393

The rest of this paper is organized as follows. The next section presents a survey of related work. The proposed model, MATIC, is described in Section 3, followed by our experimental setup in Section 4. MATIC is evaluated on benchmark datasets in Section 5.

2. Related Works

Time-series classification has been a long-standing area of research within the fields of data mining and machine learning. Over the years, a wide range of methods have been developed, evolving from traditional distance-based approaches to sophisticated deep learning architectures. This section provides an overview of existing approaches, with particular emphasis on unimodal and multimodal methods, recent advances in deep learning-based models, and emerging work in cross-domain multimodal learning.

Multimodal time-series classification (MTSC) integrates heterogeneous data sources, such as physiological signals, medical imaging, or textual records, to capture complementary and context-rich information that improves predictive performance over unimodal models [8]. Despite its promise, MTSC poses significant challenges, including heterogeneity in data formats, mismatched temporal resolutions, modality-specific noise, and the difficulty of learning aligned joint representations across streams.

To address these challenges, prior work has explored a range of fusion strategies. Early fusion concatenates raw inputs, intermediate fusion combines learned feature representations, and decision-level fusion aggregates modality-specific predictions. Recent advances increasingly leverage attention mechanisms, contrastive alignment, and hybrid CNN-RNN (convolutional neural network-recurrent neural network) architectures to model cross-modal dependencies and temporal structure [10]. InstructTime [11], for example, reframes time-series classification as a generative problem by aligning temporal signals with textual representations through discretization and semantic bridging.

Self-attention has emerged as a particularly effective mechanism for adaptive multimodal fusion. Poria et al. [12] demonstrated that self-attention can dynamically weight modalities based on reliability, especially under incomplete data scenarios via modality dropout. Their work motivates the use of attention as a principled alternative to static fusion, enabling data-driven modulation of modality contributions without explicit supervision.

Subsequent studies have extended self-attention fusion across diverse domains. Zhu et al. [13] introduced a low-rank tensor fusion framework with self-attention to reduce computational complexity while preserving inter-modal interactions, achieving strong performance on multimodal emotion recognition benchmarks. Liu et al. [14] proposed SFusion, an N-to-One attention-based fusion block that accommodates variable modality availability without imputation, demonstrating robustness across human activity recognition and medical imaging tasks. Domain-specific applications further illustrate the versatility of attention-based fusion, including remote-sensing scene classification [15], multimodal 3D reconstruction [16], and sequential recommendation systems [17].

While self-attention has been widely adopted for multimodal fusion, existing approaches often focus on either early or intermediate fusion, where modality embeddings are jointly transformed at multiple layers. Such designs can be effective but frequently introduce architectural complexity, increased parameter counts, and tight coupling between modalities, which reduces modularity and scalability. Moreover, many attention-based fusion models assume strong cross-modal alignment or require carefully designed interaction blocks, limiting their applicability as general-purpose solutions for diverse multimodal time-series tasks.

In contrast, our framework adopts a late-stage self-attention fusion strategy, where modality-specific temporal representations are first learned independently and then fused through a shared multi-head self-attention mechanism at the decision level. This design

preserves modality autonomy, enables plug-and-play extensibility, and reduces architectural overhead, while still allowing the model to learn dynamic, data-driven weighting of modalities. Unlike prior approaches that tailor fusion mechanisms to specific modality types or domains, our method provides a general-purpose and lightweight fusion paradigm applicable across heterogeneous multimodal time-series settings.

Multimodal time-series classification has leveraged datasets like DEAP [18] and UTD-MHAD [19] to evaluate models that integrate diverse sources such as electroencephalogram (EEG), peripheral physiological signals (PPS), videos, and inertial sensors. These datasets enable the development of models that learn rich, dynamic interactions across modalities, facilitating tasks like emotion and human activity recognition.

Ma et al. [20] proposed MMResLSTM, a multimodal residual LSTM network for emotion classification using EEG and peripheral signals. It employed weight-sharing LSTMs to capture inter-modal correlations and residual paths to stabilize training. Evaluated on the DEAP dataset, it achieved 92.87% accuracy for arousal and 92.30% for valence. Zhang et al. [21] introduced HFCNN, which fused CNN-extracted hierarchical features with handcrafted statistics. Achieving 84.71% (valence) and 89.00% (arousal), their method emphasized hybrid fusion and multiscale feature extraction.

Liu et al. [22] presented MSMDFN, which models inter-modal dependencies through a multi-stage fusion strategy. It uses separate extractors for EEG, electrooculogram (EOG), and electromyogram (EMG) signals, followed by staged dynamic fusion guided by correlation coefficients. On the DEAP dataset, MSMDFN reached 93.52% for valence and 94.21% for arousal, outperforming earlier methods. Its design effectively preserves intra-modal information while modeling fine-grained inter-modal interactions. There are also highly specialized models for the DEAP dataset using only the EEG modality. DSE-Mixer [23] utilizes dynamic spatial and channel mixing layers to capture global spatial information, achieving 95.19% accuracy for arousal and 95.22% for valence. Similarly, TPRO-NET [24] leverages differential entropy features and enhanced transformer encoders to achieve even higher average accuracies of 97.63% for valence and 97.47% for arousal.

For human action recognition (HAR), several methods have shown strong performance on the UTD-MHAD dataset. HAMLET [25] used a hierarchical self-attention mechanism to model spatio-temporal features, achieving 95.12% accuracy. MuMu [26] incorporated cooperative multitask learning, guiding feature extraction via an auxiliary grouping task, and reached 97.60%. MAVEN [27] employed memory banks to capture long-term spatiotemporal dependencies across modalities, attaining 97.81% accuracy which was the highest among the evaluated models.

Stock movement prediction has increasingly explored multimodal approaches that combine numerical data with textual or social media sources. For instance, Xu et al. [28] integrated historical prices with tweets to improve prediction accuracy, demonstrating that combining complementary sources can enhance performance. Ma et al. [29] proposed the Multi-source Aggregated Classification (MAC) model, which incorporates news from both target companies and related firms via graph convolutional networks (GCNs) and combines them with 31 technical indicators using BiLSTM with attention-based fusion. Their approach showed that pre-training news embeddings with actual market movements and incorporating inter-company relationships can substantially improve prediction.

Several studies have shown that deep learning models using only numerical technical indicators can achieve strong stock prediction performance. Wu et al. [30] proposed a graph-based CNN-LSTM (SACLSTM) using historical prices and leading indicators, while Hoseinzade and Haratizadeh [31] employed CNNs on indicators like RSI and MACD and Agrawal et al. [32] used LSTMs with adaptive STIs and Correlation-Tensors for feature selection. Zhang et al. [33] integrated variational mode decomposition (VMD) and triangulated

maximally filtered graph (TMFG) with LSTM to process 43 indicators efficiently. Collectively, these works demonstrate that treating each technical indicator as a separate channel or modality, and aggregating them via attention or graph-based fusion, can effectively capture complementary information across numerical features.

In contrast, our work focuses exclusively on numerical market variables, including technical indicators and fundamental ratios as separate channels or modalities, without relying on textual or social data. Each indicator is treated as an independent input to the self-attention fusion model, allowing the network to dynamically weight their contributions for accurate prediction. This approach simplifies the input space while still leveraging multimodal fusion principles to capture complementary and temporally correlated patterns across financial indicators.

3. Materials and Methods

3.1. Problem Statement

Time-series classification across heterogeneous modalities presents significant challenges due to differences in signal characteristics and temporal dependencies. Formally, let $M_1, M_2, \ldots, M_m$ denote m modalities, where each modality M_i (for $i = 1, 2, \ldots, m$) may contain one or more channels. Each channel in modality M_i is represented as $C_{i1}, C_{i2}, \ldots, C_{in_i}$, where n_i is the number of channels in modality M_i. For each channel C_{ij}, a classifier λ_{ij} estimates posterior probabilities for a set of classes $\{\omega_1, \omega_2, \ldots, \omega_K\}$ as $P_{ij} = \left[P(\omega_k \mid C_{ij})\right]_{k=1}^{K} \in \mathbb{R}^K$, where $k = 1, 2, \ldots, K$.

The goal of multimodal time-series classification is to learn a function g that assigns an input pattern X to one of the K classes by effectively integrating information across all channels and modalities [34]. Achieving this requires capturing both intra-channel temporal patterns and inter-channel correlations while handling the heterogeneity of the input data.

To address this challenge, we propose a recognition framework that combines a CNN-based classifier for each channel with a self-attention based fusion model, which aggregates the posterior probabilities from all channels to produce a final prediction as shown in Figure 1. This approach allows the model to exploit both local temporal features and global cross-channel dependencies, improving classification performance in multimodal time-series tasks. Further architectural details are provided in Appendix D.

3.2. Dilated Residual CNN (Channel Model)

For each channel C_{ij} in modality M_i, we employ a 1-D Dilated Residual Convolutional Neural Network (CNN) [35,36] as the base classifier λ_{ij}. This model is designed to capture both local and long-range temporal dependencies in time-series data, while preserving information flow through residual connections.

Input Layer

The input to the model is a single-channel time series $C_{ij} \in \mathbb{R}^{T \times 1}$, where T denotes the number of timesteps.

Initial Convolution and Pooling

The input is first processed through a 1D convolutional layer with dilation, followed by a non-linear activation and temporal downsampling (pooling). This layer extracts local temporal features while enlarging the receptive field.

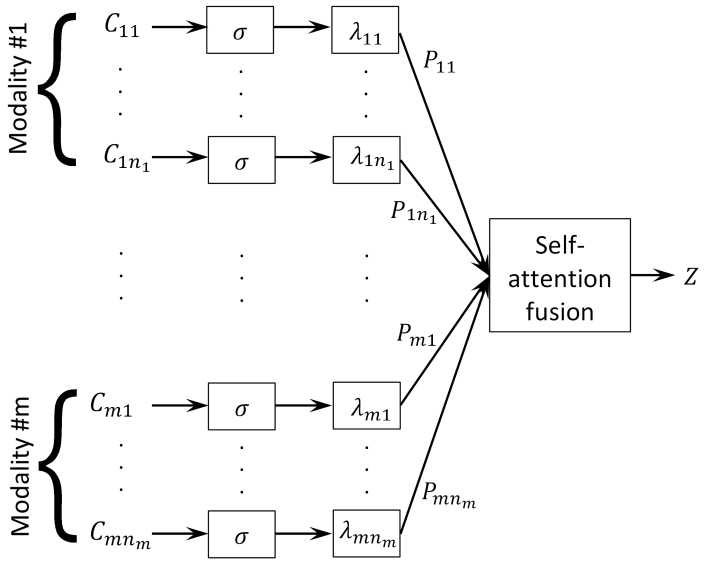


Figure 1. Block diagram for our multimodal time-series classification model using decision-level fusion, where C_{ij} is the input signal from the jth channel of the ith modality ($i = 1, 2, \ldots, m$), ith modality contains n_i channels, σ is a normalization function, λ_{ij} is a CNN classifier, P_{ij} is a probability distribution over the class labels for C_{ij}, and Z is the final probability distribution over the class labels for all channels and modalities.

Residual Block

To facilitate gradient flow and enable deeper networks, a residual block is applied. The input to the block is processed through a sequence of dilated convolutions and non-linear activations. The output of this block is summed with a skip connection (residual) from the block input, preserving original information and supporting deep feature learning. Temporal downsampling is applied after the residual addition.

Deep Feature Extraction

The downsampled representation is further processed through additional dilated convolutional layers with non-linear activations to capture higher-level temporal patterns. Global feature aggregation, such as Global Average Pooling, is then applied along the temporal dimension to produce a fixed-length feature vector.

Fully Connected Layers and Output

The aggregated features are passed through fully connected layers with non-linear activations and regularization (e.g., dropout) to learn abstract representations. The final layer is a dense layer followed by a softmax layer, producing the posterior class probabilities:

$$P_{ij} = [P(\omega_1 \mid C_{ij}), \ldots, P(\omega_K \mid C_{ij})].$$

Loss Function

The network is trained using categorical cross-entropy:

$$\mathcal{L}(\theta) = -\sum_{k=1}^{K} y_k \log([P_{ij}]_k),$$

where y_k is the ground-truth label indicator for class ω_k. This enables each channel classifier λ_{ij} to learn reliable posterior probabilities for subsequent fusion.

3.3. Self-Attention Fusion Model

After obtaining posterior probability vectors from each channel classifier λ_{ij}, we fuse these outputs across channels and modalities using a self-attention mechanism. The fusion model learns adaptive weights over channel predictions, allowing the network to emphasize the most informative signals for the final decision.

Input Representation

For each channel C_{ij}, let

$$P_{ij} = [P(\omega_1 \mid C_{ij}), \ldots, P(\omega_K \mid C_{ij})] \in \mathbb{R}^K$$

denote the posterior probability vector predicted by classifier λ_{ij}.

We stack all channel outputs to form

$$\mathbf{P} = [P_{11}, \ldots, P_{1n_1}, \ldots, P_{m1}, \ldots, P_{mn_m}] \in \mathbb{R}^{C \times K},$$

where $C = \sum_{i=1}^{m} n_i$ is the total number of channels across all modalities.

Linear Projection

The stacked probabilities are projected into a latent feature space:

$$Q = \mathbf{P}W_Q, \quad K = \mathbf{P}W_K, \quad V = \mathbf{P}W_V,$$

where $W_Q, W_K, W_V \in \mathbb{R}^{K \times d}$ are trainable matrices and d is the attention dimension.

Scaled Dot-Product Self-Attention

Attention weights are computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V,$$

yielding context-aware channel representations.

Multi-Head Attention

To model diverse inter-channel relationships, we employ multi-head attention:

$$\text{MultiHead}(\mathbf{P}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O,$$

where each head is computed independently using separate projections.

Global Feature Aggregation

The attention outputs are pooled across channels:

$$z = \text{GlobalAveragePooling}(\text{MultiHead}(\mathbf{P})).$$

Output Layer

The pooled representation is passed through fully connected layers, producing the final fused posterior distribution:

$$Z = f(\mathbf{P}) \in \mathbb{R}^K.$$

Training Objective

The fusion model is optimized using categorical cross-entropy:

$$\mathcal{L}_{fusion} = -\sum_{k=1}^{K} y_k \log(Z_k),$$

where y_k is the ground-truth label indicator for class ω_k. The corresponding end-to-end objective, in which the channel classifiers and the fusion module are trained jointly, is derived in Appendix A.

4. Experimental Setup

4.1. Classification

For each dataset, the classification task is defined as follows: for DEAP and DREAMER, each input segment is classified into high or low levels of three emotional dimensions: valence, arousal, and dominance; for UTD-MHAD, each segment is assigned to one of 27 human actions; for mHealth, each segment is classified into one of 12 physical activities; for the stock market datasets (S&P500, DJI, NASDAQ, and selected Chinese companies), each input window is classified based on the next-day price movement as up or down. Inputs consist of multimodal channels for each dataset, segmented into fixed-length windows, and models are evaluated on held-out test data or using cross-validation where applicable.

4.2. Training Procedure

4.2.1. Datasets

We evaluate our model across three real-world application domains: emotion recognition from physiological signals, action recognition using wearable sensor data, and stock price movement prediction, as summarized in Table 1.

Table 1. Summary of all datasets used in this study.

Dataset	N	Modalities	Sampling Rate	Total Instances	Labels
DEAP	32	EEG, EOG, EMG, GSR, PLET, RESP, TEMP	128 Hz	32 participants × 40 videos = 1280 samples	9-point scale for Arousal, Valence, Dominance, Liking
DREAMER	23	EEG, ECG	256 Hz	23 participants × 18 videos = 414 samples	Valence, Arousal, Dominance (5-point scale)
UTD-MHAD	8	Inertial, RGB Video, Depth, Skeletal	Inertial: 50 Hz; Skeletal/RGB/Depth: 30 Hz	27 actions × 8 participants × 4 sessions = 864 samples	27 action classes (e.g., wave, push, walk, lunge)
mHealth	10	Acc, Gy, Mag, ECG	50 Hz	10 participants × 12 activities = 120 samples	12 activities (e.g., walking, jogging, sitting)
Stock Indices	–	28 technical indicators	Daily	S&P500, DJI, NASDAQ (2000–2024)	Up/Down labels
Stock Market	–	31 technical indicators	Daily	6 Chinese companies (2018–2021)	Up/Down labels

N: number of participants.

DEAP

The DEAP (Dataset for Emotion Analysis using Physiological Signals) [18] is a widely used multimodal dataset for emotion recognition. It contains recordings from 32 participants, each watching 40 one-minute music videos designed to elicit various emotional responses. Participants rated arousal, valence, dominance, and liking on a 9-point scale. The dataset includes EEG (32 channels), EOG, EMG, galvanic skin response (GSR), plethysmograph (PLET), respiration (RESP), and skin temperature (TEMP) signals, preprocessed as in [20,23,24,37] by downsampling to 128 Hz, band-pass filtering, and normalization using baseline recordings. Ratings were binarized into low (1–4) and high (5–9) emotional states, the first 3 s of each trial were discarded as baseline, and data were segmented into non-overlapping 1 s windows, resulting in 60 windows per trial. For each participant, subject-dependent stratified 5-fold cross-validation ($k = 5$) was performed over the 1 s windows. Subject-wise classification metrics were computed for each participant and aver-

aged across all subjects, following the evaluation protocol used in previous state-of-the-art studies [20,23,24,37].

DREAMER

The DREAMER dataset [38] contains recordings from 23 participants who watched 18 emotional video clips. Participants provided self-assessments of valence, arousal, dominance, and liking on a 5-point scale. EEG signals (14 channels) were recorded at 128 Hz, and electrocardiogram (ECG) signals (2 channels) at 256 Hz. The ratings were binarized into low (1–2) and high (3–5) emotional states. Signals were band-pass filtered, normalized, and segmented into non-overlapping 1 s windows, following preprocessing and evaluation protocols used in prior studies [23,24]. Subject-dependent stratified 5-fold cross-validation ($k = 5$) was performed separately for each participant over the 1 s windows. Subject-wise classification metrics were computed and averaged across participants.

UTD-MHAD

The UTD-MHAD dataset [19] is a benchmark for multimodal human action recognition, containing recordings from 8 subjects performing 27 distinct actions. Each action is captured with RGB video, depth images, inertial sensors (3-axis accelerometer and gyroscope), and 3D skeletal joint positions. The dataset facilitates evaluation across visual and motion-based modalities, supporting research in multimodal activity recognition. Inertial sensor signals are 6-channel time series sampled at 50 Hz and normalized using z-score standardization. Skeletal data consist of 20 joint 3D coordinates captured at 30 fps, normalized by translating joints relative to the hip center for translation invariance and scaling by mean bone length for size invariance [39]. RGB and depth videos are downsampled to reduce computational cost and normalized to the [0,1] range. Following the standard protocol [19], subjects {1, 3, 5, 7} are used for training and {2, 4, 6, 8} for testing, with 20% of training data reserved for validation.

mHealth

The mHealth [40–42] dataset contains recordings from 10 subjects performing 12 physical activities using three body-worn inertial measurement units (IMUs): chest-mounted sensors (3-axis accelerometer, 2-lead ECG), right ankle sensors (3-axis accelerometer, gyroscope, magnetometer), and left wrist sensors (3-axis accelerometer, gyroscope, magnetometer), yielding 23 channels total sampled at 50Hz. We preprocess the data as follows as done in [43–45]: Activities labeled as 0 (null class) are removed, and the continuous time series is segmented into 2.72 s windows with 50% overlap (136 timesteps per window, 68-timestep stride). Each window is assigned the majority-vote label from its constituent samples. We employ leave-one-subject-out (LOSO) cross-validation, training on 9 subjects and testing on the held-out subject.

Stock Indices

Daily trading data for the Dow Jones Industrial Average (DJI), S&P 500, and NASDAQ spanning 2000–2024 were used. For each index, 28 technical indicators are extracted, including price-based, volume-based, momentum, trend, and volatility features; the complete list of indicators and their groupings is provided in Table A1. Data preprocessing, labeling, and windowing follow the protocol in [29], with binary labels indicating next-day price movement (0 = down, 1 = up) as in [28,29]. The data are temporally split into training (up to 2014), validation (2015–2019), and testing (2020–2024) sets and segmented into 6-day sliding windows with Min–Max normalization.

Chinese Stock Companies

Data from six companies listed on the Shenzhen Stock Exchange were used: 000063 ZTE Corporation (Manufacturing of communications), 000651 Gree Electric Appliances Inc. (Manufacturing of electrical machinery), 601800 China Communications Construction Ltd. (Civil engineering construction), 000876 New Hope Liuhe Ltd. (Farm and sideline food processing), 600104 SAIC Motor Ltd. (Automobile industry), and 601933 Yonghui Superstores Ltd. (Wholesale and retail trade). Each company had 31 features: 28 technical indicators, similar to stock indices, and 3 fundamental ratios (Price-to-Earnings, Price-to-Book, and Price-to-Sales) as in [29]. Binary labels indicated next-day price movement (0 = down, 1 = up). The data were temporally split as follows: training from 2 January 2018 to 21 October 2020, validation from 22 October 2020 to 18 February 2021, and testing from 19 February 2021 to 18 June 2021. Data were segmented into 6-day sliding windows, and each feature was treated as an independent channel for multimodal fusion.

MATIC can be trained using two complementary strategies: sequential training and end-to-end training, as outlined in Algorithms 1 and 2.

Algorithm 1 Sequential Posterior Probability Fusion

Require: Dataset $\mathcal{D} = \{(X^{(n)}, y^{(n)})\}_{n=1}^{N}$, modalities $\{M_1, M_2, \ldots, M_m\}$, classes $\{\omega_1, \omega_2, \ldots, \omega_K\}$
Ensure: Final fused posterior distribution Z
Channel-wise classifier learning
for each modality M_i **do**
  for each channel $C_{ij} \in M_i$ **do**
    Learn channel classifier: $\lambda_{ij} \leftarrow \arg\min_{\lambda} \mathbb{E}_{(X,y)\sim\mathcal{D}}[\mathcal{L}(y, P_{ij})]$
  end for
end for
Posterior probability estimation
for each channel C_{ij} **do**
  Compute posterior vector: $P_{ij}(X) \leftarrow [P(\omega_1 \mid C_{ij}), \ldots, P(\omega_K \mid C_{ij})]$
end for
Posterior tensor construction
Form posterior set: $\mathbf{P}(X) \leftarrow \{P_{ij}(X) \mid \forall i, j\}$
Fusion model learning
Learn fusion function: $f \leftarrow \arg\min_{f} \mathbb{E}_{(X,y)\sim\mathcal{D}}[\mathcal{L}(y, f(\mathbf{P}(X)))]$
Inference
Compute final fused posterior distribution: $Z \leftarrow f(\mathbf{P}(X))$
return Z

Sequential Fusion Training

Each channel-specific CNN is trained independently to produce posterior probabilities for each class. Once trained, the CNN weights are frozen, and the predicted probabilities from all channels are stacked to form the input to the self-attention fusion model. The fusion model learns to combine the channel-level posterior probabilities and outputs the final prediction probabilities:

$$\hat{P}(\omega_k|X) = f_{\text{fusion}}\Big(\{P(\omega_k|C_{ij})\}\Big)$$

where X is the input pattern across all channels. This strategy allows the fusion model to leverage already learned channel-specific features. The fusion model is trained using posterior probabilities generated from the same training examples used to train the channel classifiers, while a validation split held out from the training folds is used for early stopping; the test fold is not used during training.

Algorithm 2 End-to-End Attention-Based Fusion

Require: Dataset $\mathcal{D} = \{(X^{(n)}, y^{(n)})\}_{n=1}^{N}$, modalities $\{M_1, M_2, \ldots, M_m\}$, classes $\{\omega_1, \omega_2, \ldots, \omega_K\}$
Ensure: Final fused posterior distribution Z
Joint model definition
for each modality M_i **do**
  for each channel $C_{ij} \in M_i$ **do**
    Define channel classifier: $\lambda_{ij} : C_{ij} \to \mathbb{R}^K$
  end for
end for
Define attention-based fusion function f using Attn
Let Θ denote all trainable parameters of $\{\lambda_{ij}\}_{\forall i,j}$ and f
Posterior probability estimation
for each channel C_{ij} **do**
  Compute posterior vector: $P_{ij}(X) \leftarrow [P(\omega_1 \mid C_{ij}), \ldots, P(\omega_K \mid C_{ij})] = \lambda_{ij}(C_{ij})$
end for
Posterior tensor construction
Form posterior set: $P(X) \leftarrow \{P_{ij}(X) \mid \forall i, j\}$
Attention-based fusion
Compute fused posterior distribution: $Z \leftarrow f(P(X))$
Joint optimization
Learn all parameters jointly: $\Theta \leftarrow \arg\min_{\Theta} \mathbb{E}_{(X,y)\sim\mathcal{D}}[\mathcal{L}(y, Z)]$
Inference
Compute final fused posterior distribution: $Z \leftarrow f(P(X))$
return Z

End-to-End Training

In this strategy, all channel-specific CNNs and the self-attention fusion model are trained jointly. Inputs from all channels are fed simultaneously, and the network learns both local channel-specific features and global cross-channel dependencies. Formally, the model is optimized to directly map the full input pattern Z to final posterior probabilities:

$$\hat{P}(\omega_k|Z) = f_{\text{e2e}}(\{C_{ij}\})$$

where f_{e2e} represents the unified end-to-end model. This approach allows gradients from the fusion layer to influence channel-specific feature extraction during training, potentially improving joint feature learning.

4.3. Evaluation Metrics

After training, each model is evaluated on the test inputs by computing posterior probabilities for all classes, from which predicted labels are obtained for performance assessment. For the emotion and action recognition tasks, classification performance is evaluated using accuracy, precision, recall, and F1-score, following commonly used evaluation practices in these domains [18–20,23–25,34,37,45,46]. For the stock-movement prediction tasks, accuracy and Matthews Correlation Coefficient (MCC) are reported, following [29].

- **Accuracy (*Acc*)**: Represents the ratio of correctly predicted instances to the total number of instances.

$$Acc = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

- **Precision (*P*)**: Measures the proportion of true positive predictions among all instances predicted as positive.

$$P = \frac{TP}{TP + FP}$$

where TP is the number of true positives, and FP is the number of false positives.

- **Recall (*R*)**: Represents the proportion of true positives correctly identified from all actual positives.

$$R = \frac{TP}{TP + FN}$$

where FN is the number of false negatives.

- **F1-Score (*F1*)**: Provides a harmonic mean of Precision and Recall, offering a single measure that balances both metrics.

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

- **Matthews Correlation Coefficient (MCC)**: Measures the quality of binary (or multi-class via generalization) classifications, taking into account all four confusion matrix categories.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TN is the number of true negatives. MCC returns a value between -1 and 1, where 1 indicates perfect prediction, 0 indicates random prediction, and -1 indicates total disagreement between prediction and observation.

5. Results

5.1. Evaluation for Accuracy

Tables 2–5 summarize the predictive performance of the proposed framework across emotion recognition, action recognition, and financial time-series classification benchmarks. Results for MM-ResLSTM [20], DSE-Mixer [23], NAS [37], TPRO-NET [24], MLP with modality selection [34], HAMLET [25], DBN [45], and MAC [29] are taken from the original publications, whose reported experimental setups were followed as closely as possible. Baseline LSTM, Baseline MLP, and all fusion strategies were reproduced using identical data splits, preprocessing, and channel classifiers. Differences in input settings are also considered: NAS, DSE-Mixer, and TPRO-NET use EEG only, MM-ResLSTM includes peripheral signals, and DSE-Mixer and TPRO-NET use hand-crafted EEG features, whereas MATIC uses raw signals [20,23,24,37]. For MAC, we compare with its numerical-indicator variant [29]. Overall, the results demonstrate that the proposed model achieves competitive performance across diverse multimodal tasks while maintaining a compact and general-purpose architecture. Detailed classification metrics for the emotion and activity recognition datasets are provided in Table A2.

On the emotion recognition benchmarks, the proposed framework shows highly competitive performance on both DEAP and DREAMER. Strong results are obtained under EEG-only settings, indicating that the channel-specific temporal encoders are able to extract informative representations from neural signals alone. When peripheral physiological signals are incorporated, performance further improves in most tasks, highlighting the benefit of integrating complementary modalities through the proposed decision-level fusion mechanism. Notably, performance remains strong across valence, arousal, and dominance prediction, with only a modest decrease for DEAP arousal under sequential training when PPS are added.

Table 2. Performance comparison on emotion recognition benchmarks (DEAP and DREAMER). Results are reported as classification accuracy (%) in mean, std format. PPS (Peripheral Physiological Signal).

Dataset	Methods	Modalities	Valence	Arousal	Dominance
DEAP	MM-ResLSTM [20]	EEG, PPS	92.30, 1.55	92.87, 2.11	–
	DSE-Mixer [23]	EEG	95.18	95.22	–
	NAS [37]	EEG	97.94, 1.04	97.74, 1.02	97.82, 1.20
	TPRO-NET [24]	EEG	97.87, 1.89	98.08, 1.83	98.33, 1.55
	MATIC (seq)	EEG	97.70, 1.72	**98.24, 0.79**	98.53, 1.80
	MATIC (seq)	EEG, PPS	98.41, 1.70	97.11, 2.36	98.66, 2.24
	MATIC (e2e)	EEG	97.83, 3.37	97.33, 3.07	97.99, 3.24
	MATIC (e2e)	EEG, PPS	**98.86, 1.29**	97.93, 2.36	**98.82, 2.49**
DREAMER	NAS [37]	EEG	96.62, 3.52	96.29, 3.82	96.61, 4.04
	TPRO-NET [24]	EEG	98.86, 0.57	98.97, 0.49	98.93, 0.69
	MATIC (seq)	EEG	98.94, 0.21	99.17, 0.18	99.60, 0.36
	MATIC (seq)	EEG, PPS	99.71, 0.31	99.70, 0.32	99.72, 0.41
	MATIC (e2e)	EEG	99.19, 0.42	99.24, 0.21	99.63, 0.47
	MATIC (e2e)	EEG, PPS	**99.78, 0.14**	**99.78, 0.14**	**99.74, 0.41**

Table 3. Performance comparison on human activity recognition datasets. Results are reported as classification accuracy (%). Best results for each dataset are shown in bold.

Dataset	Methods	Modalities	Acc (%)
UTD-MHAD	MLP with modality selection [34]	Inertial, Skeletal, RGB, Depth	92.40
	Hamlet [25]	Inertial, Skeletal, RGB, Depth	**95.12**
	MATIC (seq)	Inertial, Skeletal, RGB, Depth	93.19, 2.04
	MATIC (e2e)	Inertial, Skeletal, RGB, Depth	94.19, 1.77
mHealth	DBN [45]	Acc, Gy, Mag, ECG	94.67
	MATIC (seq)	Acc, Gy, Mag, ECG	96.03, 4.06
	MATIC (e2e)	Acc, Gy, Mag, ECG	**96.10, 3.29**

Table 4. Performance comparison of our sequential and end-to-end variants against Baseline LSTM and Baseline MLP on major stock indices in terms of accuracy (Acc) and Matthews correlation coefficient (MCC). The majority-class baseline always predicts the most frequent class of the test set, $\mathrm{Acc}_{\mathrm{maj}} = n_{\mathrm{maj}}/n_{\mathrm{test}}$, and has MCC = 0.

Index	Majority	Baseline LSTM		Baseline MLP		MATIC (seq)		MATIC (e2e)	
(2000–2024)	$n_{\mathrm{maj}}/n_{\mathrm{test}}$ (%)	Acc (%)	MCC	Acc (%)	MCC	Acc (%)	MCC	Acc (%)	MCC
S&P500	672/1252 (53.67)	51.32, 1.08	−0.0154	50.16, 1.01	0.0258	53.91, 1.69	0.0263	**54.63, 1.03**	**0.0315**
DJI	675/1252 (53.91)	54.81, 0.65	−0.0027	49.64, 1.43	−0.0156	53.67, 1.16	0.0418	**55.11, 2.16**	**0.0466**
NASDAQ	693/1252 (**55.35**)	52.88, 0.92	0.0269	50.96, 0.95	0.0014	52.56, 2.21	**0.0621**	54.49, 1.02	0.0557

For action/activity recognition, the model also performs favorably on UTD-MHAD and mHealth. Competitive results on UTD-MHAD indicate that the proposed fusion strategy can effectively combine heterogeneous sources such as inertial measurements, skeletal trajectories, RGB video, and depth information, although it does not achieve the highest accuracy among the compared specialized methods on this dataset. Similarly, strong performance on mHealth demonstrates that the framework remains effective in wearable sensing scenarios involving noisy multi-sensor streams. These findings support the suitability of the model for practical human-centered sensing applications where multiple modalities are commonly available.

Table 5. Performance comparison of MAC [29] and our sequential and end-to-end variants across Chinese stock companies in terms of accuracy (Acc) and Matthews correlation coefficient (MCC). The majority-class baseline always predicts the most frequent class of the test set, $\mathrm{Acc}_{\mathrm{maj}} = n_{\mathrm{maj}}/n_{\mathrm{test}}$, and has MCC = 0.

Stock Code	**Majority**	**MAC [29]**		**MATIC (seq)**		**MATIC (e2e)**	
(2018–2021)	$\boldsymbol{n_{\mathrm{maj}}/n_{\mathrm{test}}}$ **(%)**	**Acc (%)**	**MCC**	**Acc (%)**	**MCC**	**Acc (%)**	**MCC**
000063	63/121 (52.07)	57.79	**0.1531**	57.33, 0.00	0.1081	**58.39, 0.36**	0.1406
000651	61/121 (50.41)	56.52	0.1274	**58.67, 2.19**	**0.1952**	58.33, 3.14	0.1156
601800	69/121 (57.02)	58.03	0.1274	59.87, 0.96	0.1207	**60.38, 2.05**	**0.1289**
000876	74/121 (61.16)	58.03	0.1274	64.01, 1.06	0.1431	**64.79, 3.01**	**0.1506**
600104	65/121 (53.72)	57.41	0.1408	**61.63, 2.10**	**0.1451**	60.94, 4.61	0.1306
601933	70/121 (57.85)	57.41	0.0864	58.77, 0.03	0.1291	**59.87, 1.17**	**0.1311**

The proposed framework further generalizes to financial time-series classification, where multimodal inputs are represented by diverse technical indicators and market variables. Across both Chinese stock companies and major U.S. market indices, the proposed method performs competitively against prior baselines and learning-based approaches. For the Chinese stocks, MATIC generally improves upon the majority-class baseline and performs comparably to or better than MAC, whereas performance on the U.S. indices remains closer to the majority-class baseline. This is particularly notable because financial prediction tasks are substantially different from physiological and activity recognition problems, yet the same overall architecture can be applied without domain-specific redesign.

Ablation Study

We analyze the contribution of individual channels and the effectiveness of different fusion strategies across all tasks (Table 6).

Across the evaluated tasks, attention-based fusion generally achieves the strongest or among the strongest performance compared with single-channel baselines and conventional fusion methods, including averaging, weighted averaging, product fusion, and majority voting. While simple aggregation methods provide improvements in several cases, learned fusion approaches, particularly self-attention, generally provide larger gains, supporting the benefit of adaptive, input-dependent weighting.

Compared to MLP-based fusion, self-attention generally improves performance or remains comparable, suggesting that explicitly modeling inter-channel relationships can provide an advantage over treating concatenated predictions independently. Notably, these gains are achieved without increasing the complexity of the channel-wise classifiers, as all methods operate on the same set of posterior probabilities. Because the same channel classifiers are used across the fusion strategies, this comparison isolates the contribution of the fusion mechanism itself.

Overall, the results support the effectiveness of self-attention-based decision fusion compared with single-channel, static, and MLP-based fusion strategies while keeping the channel classifiers unchanged.

Table 6. Comparison of fusion strategies across multiple datasets with corresponding best and worst single-channel baselines. Results are reported as mean and standard deviation (mean, std) of classification accuracy (%).

Dataset	Task	Best Ch.	Worst Ch.	Avg	Wtd Avg	Product	Maj. Vote	MLP (seq)	MLP (e2e)	Self-Attn (seq)	Self-Attn (e2e)
DEAP	Valence	83.67 (Ch24)	53.31 (Ch39)	82.70, 10.39	92.83, 4.73	87.32, 9.65	73.32, 9.87	93.20, 4.69	96.40, 3.35	98.41, 1.70	**98.86, 1.29**
	Arousal	84.25 (Ch2)	54.58 (Ch39)	78.47, 9.63	90.23, 5.94	83.50, 9.07	72.43, 9.52	92.58, 4.61	95.68, 3.93	97.11, 2.36	**97.93, 2.36**
	Dominance	84.64 (Ch21)	53.74 (Ch36)	80.51, 10.82	90.58, 7.18	84.70, 9.74	74.02, 12.27	93.13, 5.07	96.16, 4.15	98.66, 2.24	**98.82, 2.49**
DREAMER	Valence	95.36 (Ch6)	77.12 (Ch12)	97.46, 3.37	99.16, 0.98	98.61, 2.52	96.10, 3.81	99.54, 0.37	99.63, 0.38	99.71, 0.31	**99.78, 0.14**
	Arousal	96.11 (Ch3)	78.10 (Ch7)	95.97, 4.35	98.45, 1.73	97.43, 2.96	95.28, 4.79	99.50, 0.39	99.62, 0.36	99.70, 0.32	**99.78, 0.14**
	Dominance	96.40 (Ch3)	77.40 (Ch13)	96.12, 3.77	98.88, 1.25	97.89, 2.54	95.57, 4.16	99.50, 0.61	99.58, 0.64	99.72, 0.41	**99.74, 0.41**
mHealth	–	88.46 ($Ankle(Acc_z)$)	30.59 ($Wrist(Gy_x)$)	95.67, 3.23	95.62, 3.23	94.09, 4.57	91.96, 3.96	94.14, 3.15	**96.17, 3.57**	96.03, 4.06	96.10, 3.29
UTD-MHAD	–	88.18 (RGB)	57.98 (Gy_z)	81.53, 1.66	87.39, 4.39	87.99, 4.08	84.48, 1.75	90.27, 3.66	92.87, 3.41	93.19, 2.04	**94.19, 1.77**
US Stock	S&P500	52.56 (EMA)	46.01 (Slowk)	53.66, 1.60	53.58, 0.00	53.84, 0.00	52.81, 0.00	52.08, 0.40	53.81, 0.00	53.91, 1.69	**54.63, 1.03**
	DJI	54.23 (OP)	45.77 (DEA)	54.52, 1.17	54.08, 0.00	54.51, 0.00	52.58, 0.00	53.03, 1.14	53.39, 2.16	53.67, 1.16	**55.11, 2.16**
	NASDAQ	51.35 (MACD)	44.57 (RSI12)	54.09, 1.94	53.84, 0.00	54.25, 0.00	53.27, 0.00	53.19, 3.16	53.80, 2.54	52.56, 2.21	**54.49, 1.02**
CN Stock	000063	53.33 (DA)	42.67 (LowerBand)	49.11, 1.49	49.29, 1.43	49.11, 1.49	48.75, 1.34	56.14, 2.87	**58.93, 1.43**	57.33, 0.00	58.39, 0.36
	000651	55.67 (SlowK)	40.01 (MA10)	49.40, 1.83	51.97, 3.76	49.40, 1.74	51.79, 1.39	51.08, 1.40	55.85, 4.59	**58.67, 2.19**	58.33, 3.14
	601800	57.01 (OP)	41.33 (MFI)	56.41, 0.00	56.41, 0.00	56.41, 0.00	56.41, 0.00	55.16, 3.07	56.41, 4.57	59.87, 0.96	**60.38, 2.05**
	000876	62.11 (CP)	38.67 (MiddleBand)	38.80, 0.42	39.32, 1.08	38.80, 0.42	39.32, 1.32	57.02, 1.06	57.30, 4.19	64.01, 1.06	**64.79, 3.01**
	600104	57.67 (SlowK)	44.00 (MA20)	54.01, 0.02	54.70, 0.98	54.70, 0.61	53.19, 0.22	53.49, 3.56	54.70, 3.05	**61.63, 2.10**	60.94, 4.61
	601933	58.67 (CP)	41.97 (MA10)	44.19, 0.46	49.58, 0.04	49.75, 0.95	51.70, 0.14	56.07, 1.19	56.18, 2.01	58.77, 0.03	**59.87, 1.17**

5.2. Evaluation for Efficiency

We evaluate computational efficiency in terms of model size and inference complexity using parameter count and floating-point operations (FLOPs). Figure 2 provides a visual comparison across tasks, while Table 7 reports the exact values for each dataset configuration.

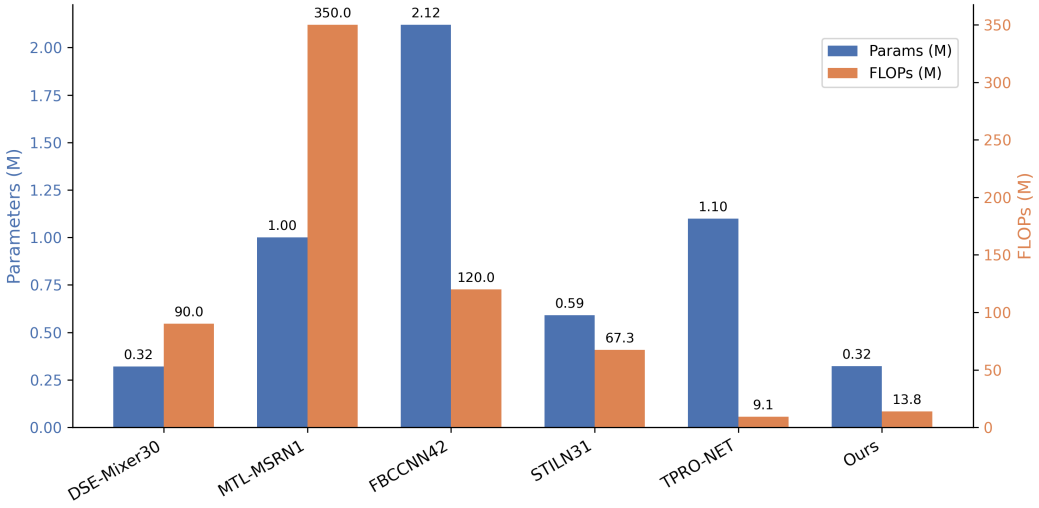


Figure 2. Comparison of model complexity for different models for DEAP dataset. Blue bars show the number of parameters (in millions) and orange bars show the FLOPs (in millions).

Table 7. Model configuration and computational cost for different datasets.

Dataset	# Channels	# Parameters	MFLOPs
Emotion Recognition			
DREAMER (EEG, ECG)	16	0.123 M	5.51
DREAMER (EEG)	14	0.108 M	4.82
DEAP (EEG, PPS)	40	0.322 M	13.77
DEAP (EEG)	32	0.262 M	11.02
Activity Recognition			
UTD-MHAD	28	1.948 M	5482.86
mHealth	23	0.179 M	15.42
Stock Prediction			
US Stock Indices	28	0.215 M	0.51
Chinese Stocks	31	0.236 M	0.57

The proposed framework remains parameter-efficient across all application domains, with fewer than 2M parameters in every setting. This highlights an important advantage of the proposed design: the same general architecture can be applied to heterogeneous multimodal time-series tasks without requiring large task-specific networks.

The emotion recognition configurations are the most compact. Both DREAMER and DEAP require only a small number of parameters and low computational cost, even when multiple physiological modalities are used. This indicates that decision-level fusion can effectively integrate complementary signals without substantially increasing model complexity.

For activity recognition, computational cost increases primarily when high-dimensional visual streams are included. The UTD-MHAD configuration has the highest FLOPs because RGB and depth inputs require spatiotemporal feature extraction, whereas the wearable-sensor-based mHealth setting remains comparatively efficient. These results suggest that

the proposed framework scales naturally with input complexity while remaining practical for real-world multimodal sensing tasks.

The stock prediction models are the most computationally efficient, requiring only a fraction of a million parameters and sub-MFLOP complexity. This demonstrates the low computational cost of the framework for numerical time-series inputs.

Taken together, the results show that the proposed model achieves competitive performance across diverse tasks while keeping the overall architecture compact, with computational cost mainly determined by the complexity of the input data.

Despite its relatively small size and simple design, the proposed framework performs competitively across a wide range of datasets. This performance is supported by three main design choices.

1. Channel- and modality-independent feature extraction: Each input channel is processed independently using a compact convolutional network, enabling effective extraction of local temporal patterns while reducing interference from unrelated channels. This preserves informative signals and limits redundancy across heterogeneous modalities.
2. Attention-based decision-level fusion: Multi-head self-attention adaptively weighs the predictions of different channels or modalities according to their relevance to each input sample. Unlike static fusion rules, it models relationships among the channel predictions while operating only on their low-dimensional posterior probabilities.
3. End-to-end learning from raw inputs: The framework learns directly from raw sequential data, reducing the need for handcrafted features and task-specific preprocessing. This also allows the same overall architecture to be used across physiological sensing, human activity recognition, and financial forecasting.

Collectively, these design choices allow the framework to remain compact while integrating information from heterogeneous inputs. Its compact size and modest computational requirements make it a promising candidate for resource-constrained and latency-sensitive applications. The same general design can therefore be applied across the evaluated domains without requiring a different architecture for each task.

5.3. Robustness and Sensitivity Analysis

To further assess the reliability of the proposed lightweight decision-level attention fusion framework, we conduct a systematic robustness and sensitivity analysis. Specifically, we evaluate (i) robustness to input perturbations and channel corruption, and (ii) sensitivity to architectural and training hyperparameters. All experiments are performed consistently across the three application domains considered in this study.

5.3.1. Robustness to Input Noise

Additive Gaussian Noise

To evaluate robustness to signal perturbations, zero-mean Gaussian noise with increasing standard deviation ($\sigma \in \{0.01, 0.05, 0.1, 0.2\}$, applied after signal normalization) is injected into each channel at inference time. Figure 3 illustrates classification accuracy as a function of noise level.

Performance decreases as the noise level increases, although the degree of degradation varies across datasets. Most datasets show a gradual decline, whereas UTD-MHAD is considerably more sensitive, with accuracy dropping to approximately 65% even at $\sigma = 0.01$. This indicates that robustness to additive noise depends strongly on the characteristics of the input data and is not uniform across tasks.

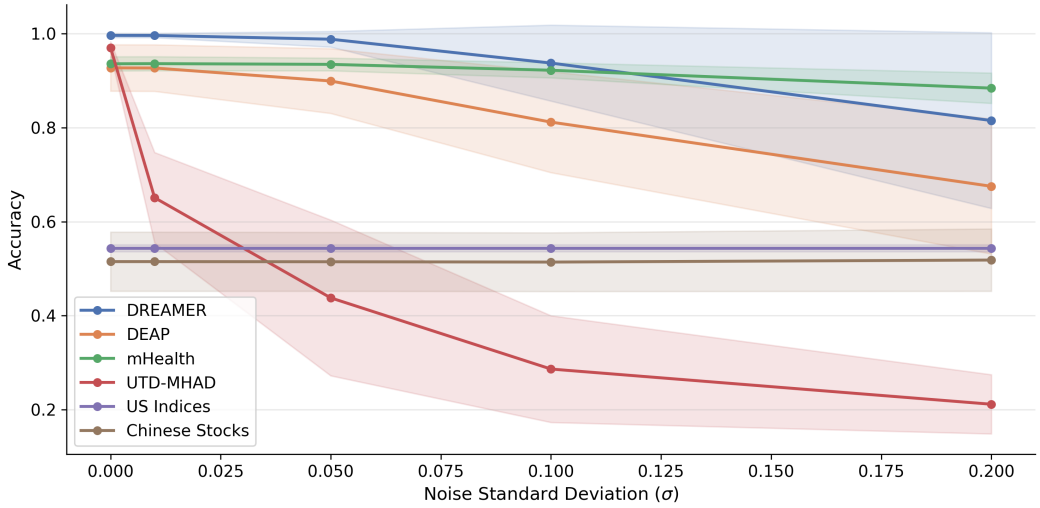


Figure 3. Robustness analysis under additive Gaussian noise. Shaded bands indicate the standard deviation across repeated runs.

Channel Dropout and Temporal Corruption

To simulate realistic sensor degradation or temporary failure, we randomly remove 10%, 20%, 30%, 40%, and 50% of the input channels at inference time without retraining the model (Figure 4). Removed channels are excluded from the attention input. In addition, random temporal masking is applied to selected signal windows (Figure 5).

As the proportion of missing channels increases, performance generally decreases, with the extent of degradation varying across datasets. Importantly, the model continues to operate under partial channel loss without retraining, showing that the decision-level fusion framework can accommodate missing inputs to some extent as illustrated in Figure 4. Similarly, Figure 5 shows that performance is retained under partial temporal corruption, although the degree of degradation varies across tasks.

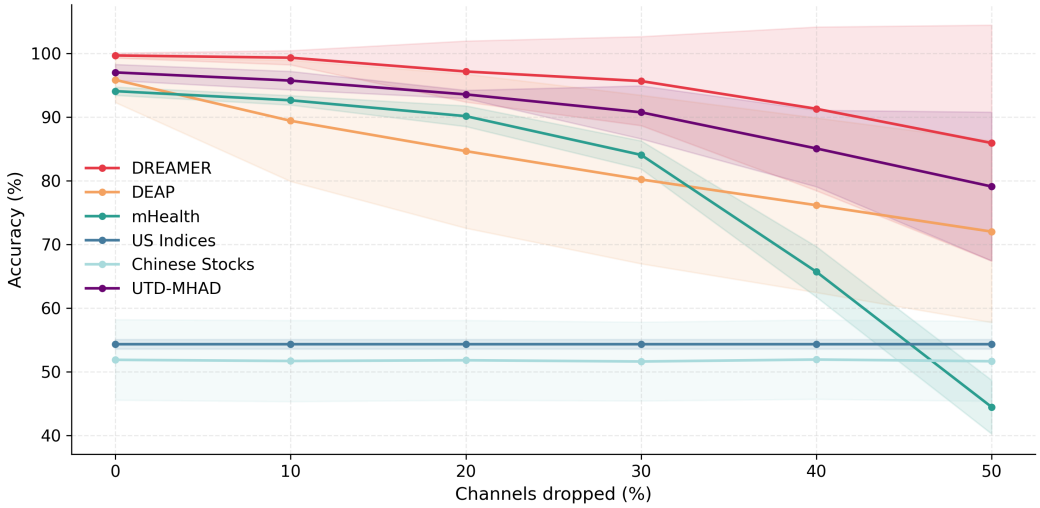


Figure 4. Impact of random channel dropout on classification performance. Shaded bands indicate the standard deviation across repeated runs.

Stability Across Random Initializations

To assess training stability, all experiments are repeated over five independent runs with different random seeds. Table 6 reports the mean and standard deviation of classification accuracy.

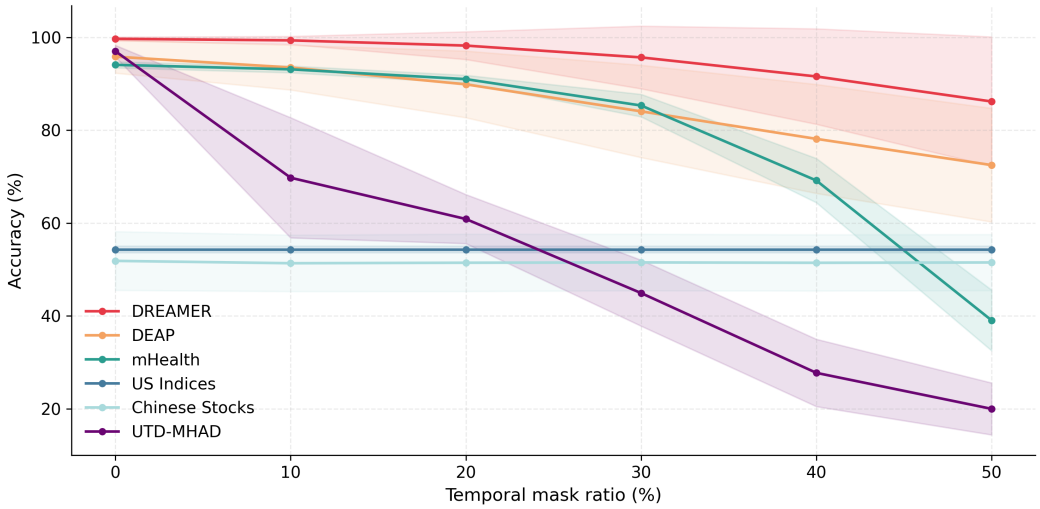


Figure 5. Effect of temporal masking on model performance. Shaded bands indicate the standard deviation across repeated runs.

The proposed model exhibits low performance variance across runs, showing relatively stable performance across different random initializations. This consistency is observed across the evaluated datasets despite differences in input type and task.

5.3.2. Sensitivity to Architectural Parameters

CNN Capacity

We vary the filter sizes and number of convolutional blocks of the individual channel classifiers across a range of compact and larger CNN configurations. Table 8 reports the performance of these individual channel classifiers before fusion, allowing us to examine how CNN capacity alone affects classification performance.

Table 8. Ablation study of CNN architectures with varying filter sizes and number of convolutional blocks across all datasets. Results are reported as mean and standard deviation (mean, std) of classification accuracy (%) for the individual channel CNNs without fusion, averaged across channels.

Filters	Blocks	Params (K)	DREAMER	DEAP	mHealth	UTD-MHAD	US Indices	Chinese Stocks
[16, 32]	1	7.7	91.17, 9.21	71.56, 8.86	86.94, 0.41	83.18, 0.37	54.71, 0.80	51.10, 4.49
[16, 32]	2	13.9	91.35, 9.04	72.45, 9.32	88.37, 1.02	84.41, 0.52	54.67, 0.81	51.58, 4.33
[32, 64]	1	25.2	92.05, 8.39	72.79, 8.50	88.37, 1.84	88.07, 1.12	54.67, 0.81	51.93, 5.16
[32, 64]	2	49.9	92.19, 8.23	73.16, 8.95	87.55, 1.43	87.78, 1.14	54.67, 0.81	52.34, 5.19
[16, 32, 64]	1	15.9	91.55, 8.94	73.18, 9.22	89.59, 1.02	86.55, 0.97	54.67, 0.81	49.93, 4.73
[16, 32, 64]	2	22.1	91.60, 9.03	73.04, 9.43	90.82, 0.20	89.59, 1.16	54.67, 0.81	51.52, 5.16
[32, 64, 128]	1	54.0	92.33, 8.21	74.08, 8.81	90.00, 0.61	89.92, 1.86	54.76, 0.76	52.62, 5.10
[32, 64, 128]	2	78.7	91.99, 8.39	73.77, 9.19	91.02, 0.41	90.11, 1.41	54.67, 0.81	51.52, 3.67

Results indicate that increasing CNN capacity does not lead to consistent improvements across datasets. In several cases, relatively small CNNs perform similarly to larger configurations, suggesting that compact channel classifiers are sufficient to capture much of the useful information from individual input signals. Because the fusion stage subsequently

combines predictions across channels, strong overall performance does not require each channel classifier to be highly complex. Additional CNN capacity therefore provides limited benefit in many of the evaluated settings.

Attention Configuration

We further analyze the impact of the attention mechanism by varying the number of attention heads ($h \in \{1, 2, 4, 8\}$), the attention embedding dimension ($d_{\text{model}} \in \{16, 32, 64\}$), and the presence or absence of a feed-forward sublayer after attention. Table 9 summarizes the resulting variation in classification accuracy while keeping the CNN backbone fixed.

Across datasets, performance varies noticeably with the attention configuration, and no single setting is universally optimal. Configurations with one or two attention heads generally perform better than those with eight heads, while increasing the attention dimension does not consistently improve accuracy. These results suggest that larger attention modules are not necessarily beneficial and that relatively compact configurations are often sufficient for effective decision-level fusion.

Table 9. Ablation study of attention module configurations with varying model dimensions (d_{model}), number of heads, and feed-forward network (FFN) usage across all datasets. Results are reported as mean and standard deviation (mean, std) of classification accuracy (%) across five runs with different random seeds. Parameter counts are shown as ranges because they vary slightly with the number of attention heads and output dimensions.

d_{model}	Heads	FFN	Params (K)	DREAMER	DEAP	mHealth	UTD-MHAD	US Indices	Chinese Stocks
16	2	No	1.2–2.0	95.65, 2.81	97.70, 2.57	94.88, 2.41	91.52, 2.88	53.84, 0.92	52.41, 1.76
16	1	No	1.2–2.0	95.86, 2.52	98.18, 1.12	95.37, 1.83	92.08, 2.47	54.12, 0.64	51.24, 0.00
16	8	No	1.2–2.0	89.94, 4.10	91.28, 4.04	90.41, 3.75	86.73, 4.55	51.96, 1.07	48.86, 2.19
16	1	Yes	2.4–4.4	96.64, 4.34	97.76, 4.18	95.92, 1.44	92.61, 2.04	54.67, 0.81	55.37, 0.83
16	8	Yes	2.4–4.4	85.80, 4.36	91.21, 3.07	89.63, 4.28	85.44, 4.93	52.48, 0.95	49.72, 2.61
16	4	Yes	2.4–4.4	88.90, 6.33	90.05, 6.98	91.12, 3.96	87.39, 4.62	52.81, 1.11	50.38, 3.07
16	2	Yes	2.4–4.4	94.29, 4.05	95.68, 4.43	94.47, 2.26	90.76, 3.21	53.43, 0.78	52.17, 1.94
32	8	No	4.4–6.0	93.33, 0.42	93.62, 0.65	92.84, 2.07	89.44, 3.58	53.09, 0.63	54.61, 2.84
32	1	No	4.4–6.0	92.48, 3.02	96.88, 1.82	94.73, 1.66	91.37, 2.89	53.74, 0.71	60.33, 8.26
32	2	No	4.4–6.0	91.47, 1.04	91.44, 1.15	93.23, 1.48	88.14, 4.41	52.31, 0.52	54.06, 0.54
32	4	No	4.4–6.0	91.81, 1.56	92.97, 1.30	93.42, 2.18	88.92, 3.02	52.96, 0.88	53.75, 2.43
32	2	Yes	6.6–9.0	91.83, 4.53	94.57, 2.06	94.15, 1.87	90.53, 2.74	53.37, 0.95	55.08, 3.11
32	8	Yes	6.6–9.0	89.77, 4.08	93.02, 2.54	92.66, 2.41	88.01, 3.65	52.74, 1.02	54.29, 2.67
32	4	Yes	6.6–9.0	91.62, 3.30	94.33, 3.61	93.91, 1.95	90.11, 2.58	53.56, 0.89	55.72, 3.24
32	1	Yes	6.6–9.0	93.08, 4.63	95.92, 2.68	95.08, 1.63	91.84, 2.21	54.03, 0.84	56.14, 2.98
64	4	No	17.0–20.2	88.21, 5.20	93.06, 3.10	92.77, 2.54	89.73, 3.43	52.68, 0.91	54.87, 2.72
64	8	No	17.0–20.2	88.85, 4.27	94.14, 3.52	93.41, 2.08	90.32, 2.94	53.11, 0.87	55.29, 2.95
64	1	No	17.0–20.2	92.97, 2.50	94.01, 3.89	94.89, 1.46	92.53, 2.12	53.94, 0.69	56.41, 1.82
64	2	No	17.0–20.2	97.69, 1.73	99.22, 0.52	96.03, 1.11	93.19, 2.04	54.39, 1.03	58.77, 0.03
64	4	Yes	21.1–24.3	93.30, 3.60	94.70, 3.92	94.52, 1.88	91.25, 2.51	53.78, 0.96	56.92, 2.61
64	8	Yes	21.1–24.3	90.93, 3.58	92.86, 4.42	93.06, 2.34	89.67, 3.36	53.22, 1.04	55.66, 2.84
64	2	Yes	21.1–24.3	94.59, 5.94	95.73, 6.36	95.43, 1.29	92.47, 2.33	54.21, 0.88	57.33, 0.00
64	1	Yes	21.1–24.3	95.22, 1.75	97.50, 1.01	95.81, 1.54	92.87, 3.41	54.08, 0.77	57.85, 0.00

6. Discussion

Effective multimodal time-series modeling depends critically on how information from multiple channels and modalities is integrated. Some channels may be more informative than others for a given instance, while others may be noisy or redundant. Across the evaluated tasks, self-attention-based decision-level fusion generally provides stronger performance than single-channel models and conventional static or MLP-based fusion strategies while maintaining a compact architecture. However, the observed gains vary across datasets, model configurations, and input conditions. In this section, we discuss the advantages of adaptive decision-level

fusion, the trade-off between predictive performance and computational complexity, and the conditions under which this approach may be less suitable.

6.1. Why the Proposed Fusion Outperforms Other Fusion Methods?

Let $\{p_1, p_2, \ldots, p_N\}$ denote the predicted probability vectors from N channel-specific classifiers, where $p_i \in \mathbb{R}^C$ and C is the number of classes. These vectors are combined through a self-attention mechanism to produce a fused prediction $\hat{p}$. Unlike simple averaging or weighted summation, where the contribution of each channel is fixed or globally learned, the attention mechanism computes an input-dependent weighting matrix A, where each entry A_{ij} encodes how strongly the prediction from channel j should influence the fused decision of channel i.

This design allows the model to dynamically modulate contributions of competing or uncertain channels and emphasize reliable or complementary information on a per-sample basis. In addition, by explicitly modeling pairwise interactions through A, the network can capture relationships between channel predictions that concatenation-based MLPs or fixed product-of-experts fusion must learn implicitly, resulting in a richer function class and more expressive fusion [14,15]. The self-attention block achieves this efficiently, scaling as $O(N^2C)$ rather than the exponential growth incurred by tensor-based fusion, making it both computationally tractable and highly expressive.

The empirical results support this distinction: across the evaluated datasets, self-attention generally performs better than static aggregation rules and MLP-based fusion when all methods use the same channel classifiers. These results suggest that input-dependent modeling of relationships among channel predictions provides an advantage over fixed or shallow learned aggregation. However, because fusion occurs at the decision level, the framework does not explicitly model interactions between intermediate features or aligned time steps across modalities, which may limit its effectiveness when such fine-grained cross-modal interactions are important.

6.2. Why the Proposed Fusion Outperforms Single-Channel Models?

Each single-modality classifier p_i is limited by the information available in its own input. Let y denote the target label, and consider the joint prediction matrix $P = [p_1; \ldots; p_N]$. When different channels provide complementary predictive information, combining their posterior distributions can provide a more informative representation of the target than relying on any individual channel alone.

The self-attention fusion module exploits this complementarity by learning the attention weights A_{ij}, which act as adaptive gates that modulate each channel's contribution based on sample-specific reliability and context. This allows the model to emphasize informative channels, reduce the influence of less useful predictions, and capture dependencies among channel-level posterior distributions. The results in Table 6 support this interpretation, as the fused models generally outperform the corresponding best individual-channel classifiers across the evaluated tasks. Importantly, these gains are obtained without increasing the complexity of the individual channel classifiers. Instead, additional predictive information is extracted by combining their outputs at the decision level, suggesting that effective multimodal integration can provide substantial gains even when each unimodal classifier remains compact [13,14].

6.3. Accuracy–Complexity–Robustness Trade-Off

To provide a comprehensive assessment, we jointly consider model size, predictive performance, and robustness under input perturbations. Across heterogeneous time-series datasets with different input characteristics, MATIC generally achieves competitive per-

formance with a relatively small number of trainable parameters, providing a favorable trade-off between accuracy and model complexity.

However, this behavior is not uniform across all datasets or input conditions. Performance varies with the attention configuration, and UTD-MHAD shows substantial sensitivity to additive Gaussian noise. MATIC also does not achieve the highest accuracy on UTD-MHAD, indicating that more complex multimodal architectures may provide an advantage when richer feature-level or temporally aligned interactions are important. Because MATIC performs fusion at the level of class-posterior distributions, such fine-grained cross-modal interactions are not explicitly modeled.

Overall, MATIC is most suitable when a balance between predictive performance, architectural simplicity, and parameter efficiency is desired. The results across the evaluated domains support its use as a strong lightweight baseline rather than a universal replacement for more complex feature-level fusion methods.

7. Conclusions

In this work, we systematically evaluated a lightweight and modular self-attention-based decision-level fusion framework for multimodal time-series classification. By combining channel-specific CNN classifiers with attention-based decision-level fusion, the model effectively captures both temporal patterns and cross-channel interactions while maintaining low computational complexity.

Experiments across diverse real-world datasets, including emotion recognition, action recognition, and stock price movement prediction, show that the proposed approach achieves competitive performance compared to existing methods while using fewer parameters than most compared models. Although it does not uniformly outperform every baseline across all datasets and efficiency measures, it provides a favorable balance between predictive performance and computational complexity. Ablation studies further confirm that attention-based fusion generally outperforms single-channel baselines and conventional fusion strategies.

Overall, this framework provides a simple and scalable baseline for multimodal time series analysis, and its compact size makes it a promising candidate for resource-constrained and real-time settings.

Author Contributions: Conceptualization, R.N. and B.B.; methodology, R.N. and B.B.; software, R.N.; validation, R.N. and B.B.; formal analysis, R.N. and B.B.; investigation, R.N. and B.B.; resources, B.B.; data curation, R.N.; writing—original draft preparation, R.N. and B.B.; writing—review and editing, R.N. and B.B.; visualization, R.N.; supervision, B.B.; project administration, B.B.; funding acquisition, B.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets analyzed in this study are publicly available and cited in the manuscript. DEAP is available at https://www.eecs.qmul.ac.uk/mmv/datasets/deap/ (accessed on 26 August 2024); DREAMER is available at https://zenodo.org/record/546113 (accessed on 27 August 2024); UTD-MHAD is available at https://personal.utdallas.edu/~kehtar/UTD-MHAD.html (accessed on 29 August 2024); mHealth is available from the UCI Machine Learning Repository at https://archive.ics.uci.edu/dataset/319/mhealth+dataset (accessed on 29 August 2024); the StockNet benchmark datasets are described in the associated cited publication. No new data were created in this study.

Acknowledgments: During the preparation of this study, the authors used ChatGPT (GPT-5) for limited rephrasing, grammar correction, improving sentence structure and flow, and LaTeX format-

ting.The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MATIC	Multimodal Attention-based Time-series Classifier
CNN	Convolutional Neural Network
MLP	Multilayer Perceptron
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
RNN	Recurrent Neural Network
GCN	Graph Convolutional Network
MTSC	Multimodal Time-Series Classification
HAR	Human Activity/Action Recognition
NAS	Neural Architecture Search
seq	Sequential (training)
e2e	End-to-end (training)
FLOPs	Floating Point Operations
MCC	Matthews Correlation Coefficient
DEAP	Database for Emotion Analysis using Physiological Signals
DREAMER	Database for Emotion Recognition through EEG and ECG signals
UTD-MHAD	UT Dallas Multimodal Human Action Dataset
mHealth	Mobile Health Dataset
EEG	Electroencephalogram
ECG	Electrocardiogram
EOG	Electrooculogram
EMG	Electromyogram
GSR	Galvanic Skin Response
PLET	Plethysmograph
RESP	Respiration
TEMP	Skin Temperature
PPS	Peripheral Physiological Signals
IMU	Inertial Measurement Unit
ACC	Accelerometer
Gy	Gyroscope
Mag	Magnetometer
RGB	Red, Green, Blue (color video)
S&P500	Standard and Poor's 500 Index
DJI	Dow Jones Industrial Average
NASDAQ	National Association of Securities Dealers Automated Quotations (Composite Index)
MAC	Multi-source Aggregated Classification
RSI	Relative Strength Index
MACD	Moving Average Convergence Divergence
VMD	Variational Mode Decomposition
TMFG	Triangulated Maximally Filtered Graph

Appendix A. End-to-End Training Objective

The loss function for the end-to-end attention-based fusion model can be derived as follows:

$$\mathcal{L}_{\text{E2E}} = -\mathbb{E}_{(X,y)\sim\mathcal{D}}[\log p(y \mid X)] \tag{A1}$$

$$= -\mathbb{E}_{(X,y)\sim\mathcal{D}}\left[\log \prod_{k=1}^{K} \hat{P}(\omega_k \mid X)^{y_k}\right] \tag{A2}$$

$$= -\mathbb{E}_{(X,y)\sim\mathcal{D}}\left[\sum_{k=1}^{K} y_k \log \hat{P}(\omega_k \mid X)\right] \tag{A3}$$

$$\hat{P}(\omega_k \mid X) = \text{Softmax}_k(\mathbf{W}\mathbf{z} + \mathbf{b}) \tag{A4}$$

$$\mathbf{z} = \text{Attn}(\mathbf{P}(X)) \tag{A5}$$

$$P_{ij}(X) = [P(\omega_1 \mid C_{ij}), \ldots, P(\omega_K \mid C_{ij})] = \lambda_{ij}(C_{ij}) \tag{A6}$$

$$\mathbf{P}(X) = [P_{11}(X), \ldots, P_{mn_m}(X)] \in \mathbb{R}^{C\times K}, \qquad C = \sum_{i=1}^{m} n_i \tag{A7}$$

$$\Theta^* = \arg\min_{\Theta} \mathcal{L}_{\text{E2E}}, \tag{A8}$$

where Θ denotes the joint set of trainable parameters of all channel classifiers $\{\lambda_{ij}\}_{\forall i,j}$, the attention-based fusion function, and the output layer.

Appendix B. Indicator Details for Stock Prediction Experiments

Table A1. Technical indicators and financial features grouped by category for stock price movement prediction.

Feature	Description	Category
OP	Opening Price	Price
CP	Closing Price	Price
HP	Highest Price	Price
LP	Lowest Price	Price
DA	Daily Amplitude	Price
Vol	Volume of Trade	Volume
TurR	Turnover Rate	Volume
DR	Daily Return	Returns
MA10	Moving Average (10-day)	Trend
MA20	Moving Average (20-day)	Trend
MA30	Moving Average (30-day)	Trend
WMA	Weighted Moving Average	Trend
EMA	Exponential Moving Average	Trend
DIFF	MACD Difference	Trend
DEA	MACD Signal Line (DEA)	Trend
MACD	MACD Histogram	Trend
RSI6	Relative Strength Index (6-day)	Momentum
RSI12	Relative Strength Index (12-day)	Momentum
RSI24	Relative Strength Index (24-day)	Momentum
MFI	Money Flow Index	Momentum
Slowk	Stochastic Indicator KDJ (Slow K)	Momentum
Slowd	Stochastic Indicator KDJ (Slow D)	Momentum
WR	Williams %R Indicator	Momentum
TR	True Range	Volatility
ATR	Average True Range	Volatility
UpperBand	Bollinger Band (Upper)	Volatility
MiddleBand	Bollinger Band (Middle)	Volatility
LowerBand	Bollinger Band (Lower)	Volatility
P/E	Price-to-Earnings Ratio	Valuation
P/B	Price-to-Book Ratio	Valuation
P/S	Price-to-Sales Ratio	Valuation

Appendix C. Detailed Classification Metrics

Table A2. Detailed classification metrics of MATIC on the emotion recognition and activity recognition datasets for sequential (seq) and end-to-end (e2e) training. Results are reported as mean and standard deviation (mean, std) in % across five runs with random seeds.

Dataset	Target	Method	Accuracy	F1	Precision	Recall
Emotion recognition						
DEAP	Valence	MATIC (seq)	98.41, 1.70	98.16, 2.73	98.34, 2.58	98.41, 1.70
		MATIC (e2e)	98.86, 1.29	98.97, 1.20	98.99, 1.17	98.86, 1.29
	Arousal	MATIC (seq)	97.11, 2.36	97.34, 2.94	97.69, 2.53	97.11, 2.36
		MATIC (e2e)	97.93, 2.36	97.28, 2.62	97.44, 3.27	97.93, 2.36
	Dominance	MATIC (seq)	98.66, 2.24	98.99, 2.25	98.24, 2.00	98.66, 2.24
		MATIC (e2e)	98.82, 2.49	98.72, 2.91	98.84, 2.64	98.82, 2.49
DREAMER	Valence	MATIC (seq)	99.71, 0.31	99.54, 0.37	99.54, 0.37	99.71, 0.31
		MATIC (e2e)	99.78, 0.14	99.78, 0.14	99.78, 0.14	99.78, 0.14
	Arousal	MATIC (seq)	99.70, 0.32	99.50, 0.39	99.51, 0.38	99.70, 0.32
		MATIC (e2e)	99.78, 0.14	99.78, 0.14	99.78, 0.14	99.78, 0.14
	Dominance	MATIC (seq)	99.72, 0.41	99.49, 0.64	99.49, 0.65	99.72, 0.41
		MATIC (e2e)	99.74, 0.41	99.72, 0.40	99.73, 0.38	99.74, 0.41
Activity recognition						
mHealth	12 activities	MATIC (seq)	96.03, 4.06	96.94, 3.85	96.31, 4.05	96.03, 4.06
		MATIC (e2e)	96.10, 3.29	96.14, 3.45	96.28, 3.94	96.10, 3.29
UTD-MHAD	27 actions	MATIC (seq)	93.19, 2.04	93.75, 3.30	93.58, 3.40	93.19, 2.04
		MATIC (e2e)	94.19, 1.77	94.57, 2.53	94.31, 2.57	94.19, 1.77

Appendix D. Architecture Details

Each input channel is processed by its own channel classifier: a 1D CNN for time series and a 3D CNN for the UTD-MHAD depth and RGB videos. The softmax outputs of all C channel classifiers are then fused by a self-attention module to produce the final prediction. All models were trained using the Adam optimizer with cross-entropy loss, a batch size of 32, and a maximum of 300 epochs, with early stopping using a patience of 15 epochs. The learning rate was set to 10^{-3}, except for the channel classifiers during end-to-end training, where a learning rate of 10^{-4} with weight decay 10^{-4} was used.

1D CNN (per channel).

Input $(1 \times T) \rightarrow$ Conv1D$(16, k = 3, d = 2)$, ReLU $\rightarrow$ MaxPool$(2) \rightarrow$ Residual block [Conv1D$(32, k = 3, d = 2)$, ReLU $\rightarrow$ Conv1D$(32, k = 3, d = 2)$, ReLU; shortcut Conv1D$(32, k = 1)$] $\rightarrow$ Global average pooling $\rightarrow$ FC(64), ReLU, Dropout$(0.2) \rightarrow$ FC(K). Here, k is the kernel size, d the dilation rate, T the window length, and K the number of classes.

3D CNN (UTD-MHAD depth and RGB).

Input $(c \times T \times H \times W) \rightarrow$ Conv3D$(32, 3 \times 3 \times 3$, stride $1 \times 2 \times 2)$, BN, ReLU $\rightarrow$ Conv3D$(64, 3 \times 3 \times 3$, stride $2)$, BN, ReLU $\rightarrow$ Conv3D$(64, 3 \times 3 \times 3)$, BN, ReLU $\rightarrow$ Conv3D$(128, 3 \times 3 \times 3$, stride $2)$, BN, ReLU $\rightarrow$ Conv3D$(128, 3 \times 3 \times 3)$, BN, ReLU $\rightarrow$ Global average pooling $\rightarrow$ FC(64), ReLU, Dropout$(0.3) \rightarrow$ FC(K). Here, c is the number of image channels (1 for depth, 3 for RGB), T the number of frames, and $H \times W$ denotes the spatial frame dimensions after preprocessing.

Self-attention fusion.

Input $(C \times K)$ channel probabilities $\rightarrow$ Linear$(d_{model}) \rightarrow$ Multi-head self-attention (h heads, scaled dot-product) across channels $\rightarrow$ Mean pooling over channels $\rightarrow$ FC(d_{ff}), ReLU,

Dropout(0.3) → FC(K). Here, d_{model} is the embedding dimension, h the number of attention heads, and d_{ff} the hidden size of the feed-forward layer.

References

1. Sezer, O.B.; Gudelek, M.U.; Ozbayoglu, A.M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Appl. Soft Comput.* **2020**, *90*, 106181. [CrossRef]
2. Barrera-Animas, A.Y.; Oyedele, L.O.; Bilal, M.; Akinosho, T.D.; Delgado, J.M.D.; Akanbi, L.A. Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting. *Mach. Learn. Appl.* **2022**, *7*, 100204. [CrossRef]
3. Ma, C.; Dai, G.; Zhou, J. Short-term traffic flow prediction for urban road sections based on time series analysis and LSTM_BILSTM method. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 5615–5624. [CrossRef]
4. Zhao, X.; Li, Z.; Zhao, C.; Wang, C. Real-time multistep time-series prediction of driver's head pose during IVIS secondary tasks for Human–Machine codriving and distraction warning systems. *IEEE Sens. J.* **2022**, *22*, 24364–24379. [CrossRef]
5. Wang, W.K.; Chen, I.; Hershkovich, L.; Yang, J.; Shetty, A.; Singh, G.; Jiang, Y.; Kotla, A.; Shang, J.Z.; Yerrabelli, R.; et al. A systematic review of time series classification techniques used in biomedical applications. *Sensors* **2022**, *22*, 8016. [CrossRef] [PubMed]
6. An, Q.; Rahman, S.; Zhou, J.; Kang, J.J. A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges. *Sensors* **2023**, *23*, 4178. [CrossRef] [PubMed]
7. Heidari Kapourchali, M.; Banerjee, B. Unsupervised feature learning from time-series data using linear models. *IEEE Internet Things J.* **2018**, *5*, 3918–3926. [CrossRef]
8. Sleeman, W.C., IV; Kapoor, R.; Ghosh, P. Multimodal classification: Current landscape, taxonomy and future directions. *ACM Comput. Surv.* **2022**, *55*, 150. [CrossRef]
9. Mohammadi Foumani, N.; Miller, L.; Tan, C.W.; Webb, G.I.; Forestier, G.; Salehi, M. Deep learning for time series classification and extrinsic regression: A current survey. *ACM Comput. Surv.* **2024**, *56*, 217. [CrossRef]
10. Jiang, Y.; Ning, K.; Pan, Z.; Shen, X.; Ni, J.; Yu, W.; Schneider, A.; Chen, H.; Nevmyvaka, Y.; Song, D. Multi-modal Time Series Analysis: A Tutorial and Survey. *arXiv* **2025**, arXiv:2503.13709. [CrossRef]
11. Cheng, M.; Chen, Y.; Liu, Q.; Liu, Z.; Luo, Y.; Chen, E. InstrucTime: Advancing Time Series Classification with Multimodal Language Modeling. In Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, Hannover, Germany, 10–14 March 2025; pp. 792–800.
12. Chumachenko, K.; Iosifidis, A.; Gabbouj, M. Self-attention fusion for audiovisual emotion recognition with incomplete data. In *Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR)*; IEEE: Piscataway, NJ, USA, 2022; pp. 2822–2828.
13. Zhu, H.; Wang, Z.; Shi, Y.; Hua, Y.; Xu, G.; Deng, L. Multimodal Fusion Method Based on Self-Attention Mechanism. *Wirel. Commun. Mob. Comput.* **2020**, *2020*, 8843186. [CrossRef]
14. Liu, Z.; Wei, J.; Li, R.; Zhou, J. SFusion: Self-attention based n-to-one multimodal fusion block. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2023; pp. 159–169.
15. Cao, R.; Fang, L.; Lu, T.; He, N. Self-attention-based deep feature fusion for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 43–47. [CrossRef]
16. Lin, X.; Chao, S.; Yan, D.; Guo, L.; Liu, Y.; Li, L. Multi-sensor data fusion method based on self-attention mechanism. *Appl. Sci.* **2023**, *13*, 11992. [CrossRef]
17. Liu, C.; Li, X.; Cai, G.; Dong, Z.; Zhu, H.; Shang, L. Noninvasive self-attention for side information fusion in sequential recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 4249–4256.
18. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* **2011**, *3*, 18–31. [CrossRef]
19. Chen, C.; Jafari, R.; Kehtarnavaz, N. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP)*; IEEE: Piscataway, NJ, USA, 2015; pp. 168–172.
20. Ma, J.; Tang, H.; Zheng, W.L.; Lu, B.L. Emotion recognition using multimodal residual LSTM network. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 176–183.
21. Zhang, Y.; Cheng, C.; Zhang, Y. Multimodal emotion recognition using a hierarchical fusion convolutional neural network. *IEEE Access* **2021**, *9*, 7943–7951. [CrossRef]
22. Chen, S.; Tang, J.; Zhu, L.; Kong, W. A multi-stage dynamical fusion network for multimodal emotion recognition. *Cogn. Neurodynamics* **2023**, *17*, 671–680. [CrossRef] [PubMed]
23. Lin, K.; Zhang, L.; Cai, J.; Sun, J.; Cui, W.; Liu, G. DSE-Mixer: A pure multilayer perceptron network for emotion recognition from EEG feature maps. *J. Neurosci. Methods* **2024**, *401*, 110008. [CrossRef] [PubMed]

24. Zhang, X.; Cheng, X.; Liu, H. TPRO-NET: An EEG-based emotion recognition method reflecting subtle changes in emotion. *Sci. Rep.* **2024**, *14*, 13491. [CrossRef] [PubMed]
25. Islam, M.M.; Iqbal, T. HAMLET: A Hierarchical Multimodal Attention-based Human Activity Recognition Algorithm. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Virtual, 24 October–24 January 2020; pp. 10285–10292. [CrossRef]
26. Islam, M.M.; Iqbal, T. MuMu: Cooperative multitask learning-based guided multimodal fusion. *Proc. Conf. AAAI Artif. Intell.* **2022**, *36*, 1043–1051. [CrossRef]
27. Islam, M.M.; Yasar, M.S.; Iqbal, T. MAVEN: A Memory Augmented Recurrent Approach for Multimodal Fusion. *IEEE Trans. Multimed.* **2023**, *25*, 3694–3708. [CrossRef]
28. Xu, Y.; Cohen, S.B. Stock movement prediction from tweets and historical prices. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 1970–1979.
29. Ma, Y.; Mao, R.; Lin, Q.; Wu, P.; Cambria, E. Multi-source aggregated classification for stock price movement prediction. *Inf. Fusion* **2023**, *91*, 515–528. [CrossRef]
30. Wu, J.M.T.; Li, Z.; Herencsar, N.; Vo, B.; Lin, J.C.W. A graph-based CNN-LSTM stock price prediction algorithm with leading indicators. *Multimed. Syst.* **2021**, *27*, 1025–1042. [CrossRef]
31. Hoseinzade, E.; Haratizadeh, S.; Khoeini, A. U-cnnpred: A universal cnn-based predictor for stock markets. *arXiv* **2019**, arXiv:1911.12540. [CrossRef]
32. Agrawal, M.; Khan, A.U.; Shukla, P.K. Stock price prediction using technical indicators: A predictive model using optimal deep learning. *Int. J. Recent Technol. Eng.* **2019**, *8*, 2297–2305. [CrossRef]
33. Zhang, Z.; Liu, Q.; Hu, Y.; Liu, H. Multi-feature stock price prediction by LSTM networks based on VMD and TMFG. *J. Big Data* **2025**, *12*, 74. [CrossRef]
34. Baruah, M.; Banerjee, B. Modality selection for classification on time-series data. *MileTS* **2020**, *20*, 6.
35. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122. [CrossRef]
36. Kiranyaz, S.; Avci, O.; Abdeljaber, O.; Ince, T.; Gabbouj, M.; Inman, D.J. 1D convolutional neural networks and applications: A survey. *Mech. Syst. Signal Process.* **2021**, *151*, 107398. [CrossRef]
37. Li, C.; Zhang, Z.; Song, R.; Cheng, J.; Liu, Y.; Chen, X. EEG-based emotion recognition via neural architecture search. *IEEE Trans. Affect. Comput.* **2021**, *14*, 957–968. [CrossRef]
38. Katsigiannis, S.; Ramzan, N. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 98–107. [CrossRef] [PubMed]
39. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A large scale dataset for 3D human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019. [CrossRef]
40. Banos, O.; Garcia, R.; Saez, A. UCI Machine Learning Repository. *MHEALTH* **2014**. [CrossRef]
41. Banos, O.; Garcia, R.; Holgado-Terriza, J.A.; Damas, M.; Pomares, H.; Rojas, I.; Saez, A.; Villalonga, C. mHealthDroid: A novel framework for agile development of mobile health applications. In *Proceedings of the 6th International Work-Conference on Ambient Assisted Living and Active Ageing (IWAAL 2014), Belfast, Northern Ireland, 2–5 December 2014*; Springer: Cham, Switzerland, 2014; pp. 91–98. [CrossRef]
42. Banos, O.; Villalonga, C.; Garcia, R.; Saez, A.; Damas, M.; Holgado-Terriza, J.A.; Lee, S.; Pomares, H.; Rojas, I. Design, implementation and validation of a novel open framework for agile development of mobile health applications. *BioMed. Eng. Online* **2015**, *14*, S6. [CrossRef] [PubMed]
43. Papagiannaki, A.; Zacharaki, E.I.; Kalouris, G.; Kalogiannis, S.; Deltouzos, K.; Ellul, J.; Megalooikonomou, V. Recognizing Physical Activity of Older People from Wearable Sensors and Inconsistent Data. *Sensors* **2019**, *19*, 880. [CrossRef] [PubMed]
44. Pippa, E.; Mporas, I.; Megalooikonomou, V. Feature Selection Evaluation for Light Human Motion Identification in Frailty Monitoring System. In Proceedings of the ICT4AgeingWell, Rome, Italy, 21–22 April 2016; pp. 88–95.
45. Almujally, N.A.; Khan, D.; Al Mudawi, N.; Alonazi, M.; Alhasson, H.F.; Jalal, A.; Liu, H. Wearable sensors-based assistive technologies for patient health monitoring. *Front. Bioeng. Biotechnol.* **2025**, *13*, 1437877. [CrossRef] [PubMed]
46. Bota, P.J.; Wang, C.; Fred, A.L.N.; Plácido Da Silva, H. A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals. *IEEE Access* **2019**, *7*, 140990–141020. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.