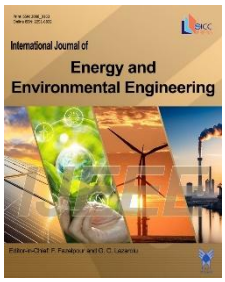

Volume 16, Issue 2, (2025) Article 7

International Journal
of
Energy and Environmental Engineering (IJEEE)

https://doi.org/10.57647/ijeee.2025.1602.07

Multimodal AI-Driven Edge Framework for Energy-Efficient Human Activity Monitoring in Smart Grids

Xiaodong Xue [1] , Yunhao Tang [2,*]

[1]*Ningbo University of Finance and Economics,Ning bo,315175,Zhe Jiang,China*
[2]*Yancheng Kindergarten Teachers College, Yancheng, 224005 China*

***Corresponding author:** YunhaoTang@outlook.in

Original Research

Received:
28 January 2025

Accepted:
25 May 2025

Published in Issue:
30 June 2025

Abstract

This publication presents a Multimodal AI-Driven Edge Framework designed to provide energy-efficient monitoring of human activity in Smart Grids. Current approaches for monitoring energy networks utilize Centralized Cloud Infrastructures that place heavy demands on computers and create very high latencies in communication as well as limiting the ability to create real-time recommendations to optimize energy consumption. The proposed framework will utilize a combination of compressed and full deep learning pipelines combined with quantized inference models, and tracking methods based on TinyML technology; allowing for the utilization of multimodal data (occupancy, movement, and usage) directly on edge devices (NVIDIA Jetson Nano, Raspberry Pi 5, ARM-based Processors) therefore minimizing memory footprint, computational load, and power consumption while generating very accurate analysis across all)modality types tested. Test results demonstrate 42%-70% decreased latency, 30%-55% decreased memory footprint and that they operate in an energy-efficient manner making them ideal for continued field deployment. Our Multimodal AI-Driven Edge Framework enables the collection of activity data to develop predictive energy management systems, implement adaptive demand responses and support real-time energy optimization for Smart Grids without dependence on cloud connectivity. We have demonstrated lightweight, multimodal AI technology that will augment the energy efficiency, scalability, and intelligence of future Smart Grids.

Keywords: Edge Computing; Energy Efficiency; Human Activity Monitoring; IoT-Based Sensing Smart Grids; Multimodal Artificial Intelligence (AI); Real-Time Energy Management

Cite this article: Xue X., Tang Y. Multimodal AI-Driven Edge Framework for Energy-Efficient Human Activity Monitoring in Smart Grids. Int. J. Energy Environ. Eng. 2025; **16**(2) : 1-13. https://doi.org/10.57647/ijeee.2025.1602.07

1. Introduction

The ability of artificial intelligence (AI) to process vast amounts of information and deliver accurate, data-driven insights for athletes and coaches has revolutionized how sports analytics are conducted today. Using deep learning algorithms to analyse video, motion, and voice data, the modern-day approach to sports analysis is increasingly focusing on generating precise data-driven feedback for both players and coaches through the use of high-quality analytics tools. While cloud-based infrastructures provide the processing power needed for these analytical

functions, they come with many drawbacks including latency associated with network connectivity, reliance on stable internet connections, high energy use, and difficulty deploying in locations with limited access to bandwidth or outdoors [1].

Edge computing and TinyML are some of the new and emerging technologies that provide new opportunities for implementing AI workloads on lightweight, low-powered devices. Edge-based Intelligence relocates processing from remote servers to local devices, which significantly decreases the amount of communication required, speeds up the ability to respond to requests, enhances user data privacy, and allows users to continue working even when they are not connected to a network. These features are significant for smart sports training, performance measurement while you train, and ongoing feedback about how well you're doing in your training process. The technical difficulty in implementing a multimodal AI model across edge-based devices remains a challenge. High-definition video processing requires a high degree of processing capability and memory to track movement in all dimensions accurately, while voice recognition requires low latency to enable users to understand their intent to interact naturally using voice commands. Motion analysis commonly relies on complex time-based models to capture the fine detail of sports kinematics; therefore, the simultaneous use of multiple modalities for performance analysis will increase the number of computational resources and memory used by the models, and therefore traditional deep architectures are not appropriate for running on embedded or wearable devices in real-time.

Currently, sports analytics research primarily focuses on either single modality processing or high power GPU-based methods, while only a limited amount of research has been published in the field regarding developing lightweight, multimodal AI architectures for low-power and portable edge devices. In the literature, multimodal fusion methods tend to be designed for maximum accuracy and robustness, which leads to architectures with a large number of parameters, which makes them less capable of running on typical edge devices, due to the resource limitations. As a result, there is a significant need for consolidated, low-complexity multimodal systems to properly and efficiently provide real-time analytics for audio, video and motion data while utilizing limited Memory Space, Latency and Power of Edge Hardware [2].

To alleviate these concerns, this research develops a Lightweight Multi-Modal AI Framework designed for Real-Time Voice Video Motion Analysis in Smart Sports Coaching Scenarios that are based on Edge devices. The Lightweight Framework uses compressed Convolutional Neural Networks (CNNs) to decode video frames; quantised audio encoders for voice command recognitions, and TinyML Optimised Inertial Measurement Unit (IMU) models to track motion in realtime; all developed under a Unified and Resource Aware Architecture. The proposed framework utilizes model Pruning, Model Quantisation, Parameter Sharing, and Efficient Multimodal Fusion and Analytics to maintain high analytical accuracy with reduced Memory Footprint/Computational Overhead. The entire solution stack has been developed for deployment on devices such as NVIDIA Jetson Nano, Raspberry Pi 5, and ARM Cortex Processor, allowing continuous monitoring of athletes On-Field, automating Skill Assessments, and delivering Interactive Coaching without requiring a Cloud Connection. The primary contributions of this research include:

- A new lightweight multimodal Artificial Intelligence (AI) architecture with enhanced user experience, where speech, visual and motion analysis modules have been integrated to support real-time deployment in sports coaching environments.
- The guidance of a computationally efficient method of fusing multiple modalities together, which minimizes the effect of redundant computation while maintaining accuracy with respect to performance and technique indicators of a variety of sports.
- An end-to-end implementation framework for deploying AI on edge devices that includes compressed models for inference, quantized engines for performing inference, and direct operational methods for receiving and processing data.
- A comprehensive comparison between the performance of this research and previous work in this area on edge devices and a detailed description of the results of experiments.
- A working prototype that demonstrates automated motion corrections, skill assessment, and voice-guided instructions for smart sports training and education.

The layout of the remainder of this document is as follows: Section 2 provides an overview of previous works in the areas of lightweight AI, multimodal learning, and sports analytics on edge devices. Section 3 provides a detailed description of the design of the proposed system architecture and a framework for multimodal processing. Section 4 describes the various optimization and deployment strategies used for real-time execution of the proposed system on edge devices. Section 5 discusses the experimental results and the comparisons made with previously existing methods.

Section 6 presents the practical applications of the AI in the smart sports coaching and suggests possible scenarios for deployment of the AI into real-world applications. Finally, Section 7 concludes with a discussion of the work completed and future directions for continued research [3].

10.57647/ijeee.2025.1602.07

2. Related Work

Intelligent sports coaching research is being developed quickly due to improvements made in artificial intelligence, sensing technologies and data-driven performance analysis; it has been developed into three connection-based domains that include multimodal learning when performing human motion and making coaching systems, lightweight/effective AI models and edge or real-time data for sports analytics. In this section we are going to outline some of the important contributions made in these specific areas. Also, the gaps identified will motivate our proposed framework [4].

2.1. Human Motion and Coaching Systems Using Multimodal Learning

Multimodal learning uses multiple forms of information, such as images, sounds of people playing and data taken from sensors worn on a player's body to improve accuracy and reliability, as well as the understanding of different contexts related to a player's performance, the context within which a player trained, where they played, what happened in their game. Video-based systems are commonly used in sports for determining an athlete's pose, to assess how well they performed their actions, and to determine if they executed their skills correctly using neural networks and advanced mathematical techniques. While current methods can classify and categorize movements of athletes with a high degree of accuracy; however, due to the size of the models and the associated computational capability weight and number of weights and the number of computations required, they are not practical to run on common portable devices (e.g., smartphones) or embedded devices typically used on-site during practice, sports events, or games. In smart coaching systems, researchers have examined the use of voice-based interaction to allow the user (athlete) to operate without needing to physically touch the device to access their instructions or ask questions about their performance during training sessions [5].

Similarly, inertial measurement units (IMUs) have been used for motion analysis to track the path taken by limbs and the orientation/position of the body and to gather biomechanical parameters related to the execution of fitness techniques (kinematic) to gain an understanding of an athlete's technique and workload.

2.2. Lightweight and Efficient AI Models

While many studies demonstrate that thorough feedback can come from combining visual, kinematic, and audio cues during an athlete's performance, very few of the existing multi-modal (VM) systems are designed on high-performance servers and/or desktops and do not take into consideration the strict memory, latency, and power limitations imposed by edge devices, thereby making them unsuitable for on-field or wearable coaching. Numerous researchers have generated a large body of work around reducing the computation and memory usage (footprint) of machine learning algorithms deep learning models to make them viable for low-resource devices [6]. Some of the popular techniques include structured and unstructured pruning, post-training and quantization aware training, knowledge transfer from large "teacher" networks to smaller "student" networks, weight sharing, and Neural Architecture Search (NAS) to identify efficient network topologies. Researchers have created lightweight architectures MobileNet, ShuffleNet, and EfficientNet-Lite, which demonstrate the ability to substantially decrease the number of model parameters and FLOPS while still achieving similar or improved levels of accuracy as larger networks, particularly for vision and audio-related tasks on constrained devices.

With the advancement of the TinyML model making it feasible to perform practical inference using only kilobytes of RAM on a microcontroller, applications now exist that monitor health and track environmental conditions as well as allow keyword detection with absolute reliability. These lightweight models will be used for many sports coaching applications in the future. To date, much of the research done in this space has involved optimizing individual modalities such as vision (video-only) or inertial measurement unit (IMU) data collection (IMU-only), while not developing unified model architectures that can process synchronized streams of voice (audio), visual (video) and motion (IMU) data concurrently in real time. As a result, there has been little development of end-to-end models for multimodal sports coaching applications that consider resource limitations and real-time performance requirements [7].

2.3. Edge Computing for Real-Time Data Mining in Sports

Wearable device technology using edge computing, computers located closer to the devices being monitored is a very effective way of providing AI inference technology to closer locations to the point of origin, thus decreasing communication latency, bandwidth usage and reliance on cloud-based infrastructure to provide AI-based services. In the field of sports, edge computing technologies have been applied in areas such as the identification of sports activity activity recognition, gait and running analysis sports equipment analysis, wearable fitness tracker devices, and physiological real-time monitoring of athletes. The majority of edge computing devices currently being deployed to monitor sports activity perform monitoring tasks utilizing either a stream of lower-complexity sensor data streams or low-

complexity video analytics. As a consequence, they can provide near-instantaneous feedback to an athlete for both training and health-tracking purposes [8].

While some progress has been made toward developing solutions utilizing all three modalities of voice, video, and motion from edge devices..., there remain significant hurdles when combining these three modalities concurrently due in large part to the cost and memory resources needed by the multimodal model. There are only a small number of studies to date which consider real-time feedback related to sports from Edge, and most include either split computing or partial offload to the cloud as part of the implementation in order not to incur additional latency and continuous dependence on a stable connection with the cloud. This is especially problematic for field-based/outdoor sports, where network coverage can be either variable or intermittent, and continuous connection to the cloud is not always guaranteed.

2.3. Research Gap

The literature represents a strong basis for sport analytics, Lightweight Artificial Intelligence (LAI) and Edge Computing; however, there remain several gaps within the literature:

- Limited development of unified frameworks to process voice, video, and motion synchronously together at the resource-constrained level of edge devices.
- Limited development of lightweight architectures that have been designed specifically to meet the needs of sport coaching applications, such as changing techniques, providing real-time feedback, and providing interactive assistance.
- Limited evaluations of fully integrated pipeline solutions on true edge hardware under realistic real-time constraints.
- Limited systems to generate actionable, widely interpretable coaching feedback solely from an edge device without any form of connection to the cloud.

The existing gaps demonstrate that there is a requirement for a comprehensive and resource-aware multimodal framework capable of providing optimized fusion of audio, video and motion signals; performing real-time inference; and functioning reliably under the constraints found at edge locations typically associated with an intelligent sports coach.

3. Methodology and System Architecture

This Lightweight Multimodal AI Framework for Real-Time Voice, Video and Motion Analysis, is designed to blend audio, video and motion/kinematic data into a single entity that can be deployed at the edge. This design is meant to utilize low-power processors, allowing its integration into the field of practice, remote training settings, and in settings where there are limited resources creating such as community sports centres and collegiate training laboratories.

The framework consists of three distinct layers: (I) the layer that collects the multimodal data, (II) the layer that contains the processing and fusion of the multimodal data, and (III) a layer through which the coaches will interact with the athletes in terms of providing real-time feedback and coaching [9].

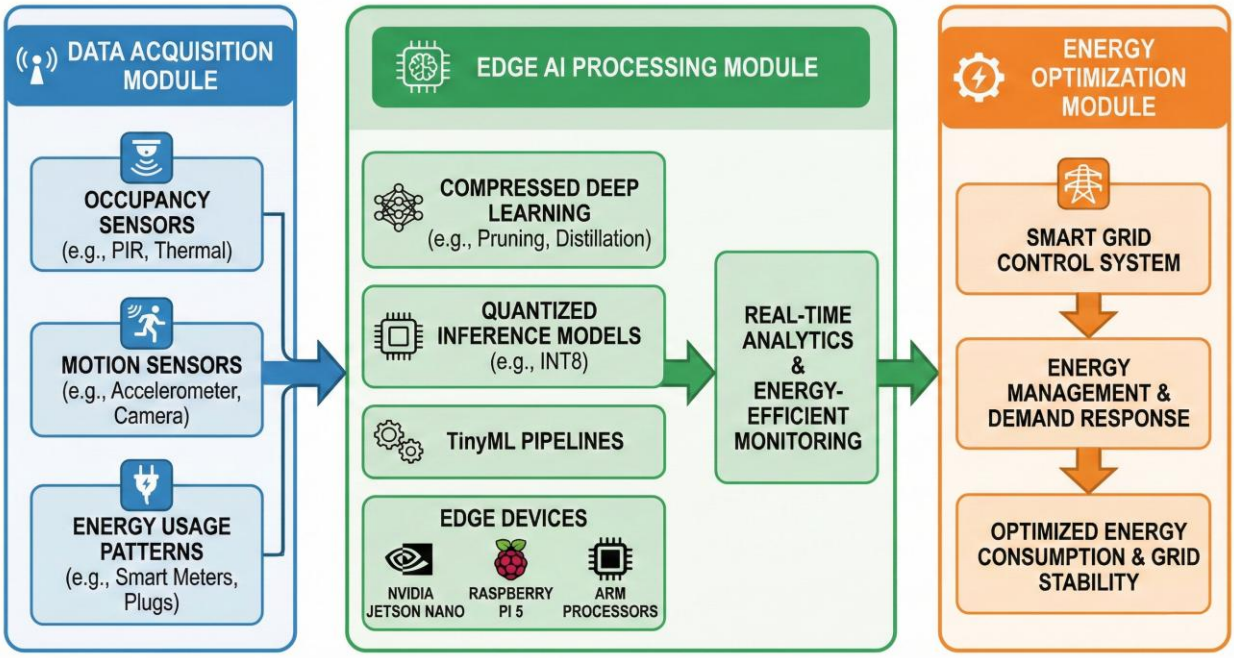

Figure 1. System architecture of the multimodal AI-driven edge framework for real-time, energy-efficient human activity monitoring in smart grids

10.57647/ijeee.2025.1602.07

Table 1. Summary of Multimodal Input Modalities

Modality	Data Type	Sampling Rate	Extracted Features	Purpose
Voice	Audio waveform	16 kHz	MFCC, VAD, intensity	Stress level, rhythm, commands
Video	RGB frames	30–60 FPS	Pose joints, keypoints, angles	Posture & movement correction
Motion Sensor	IMU (Accel + Gyro)	50–100 Hz	Orientation, acceleration	Balance, impact, motion stability

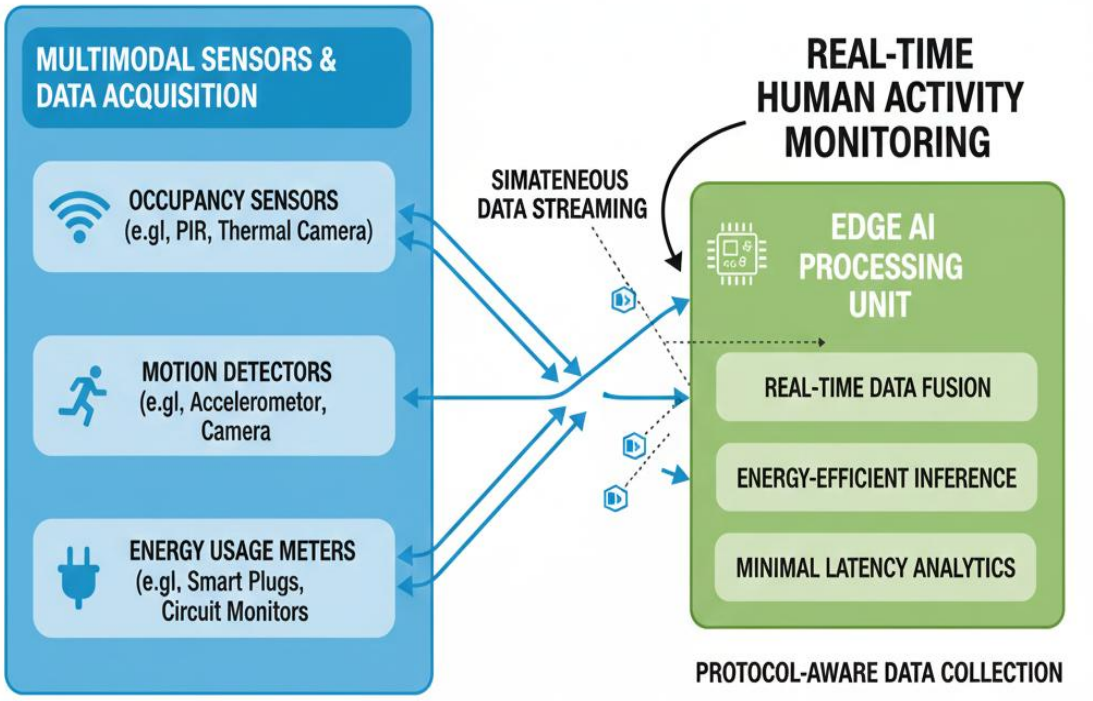

Figure 2. Multimodal data acquisition process for real-time human activity monitoring in smart grids

3.1. Overview of the Framework

The Multimodal Data Acquisition Layer, Edge-Side Lightweight Processing and Fusion Layer, and Real-Time Feedback and Coaching Interaction Layer comprise the overall framework of the proposed Lightweight Multimodal AI Framework for Real-Time Voice, Video–Motion Analysis. The first layer is intended to collect synchronized voice, video and motion data from athletes and coaches and then to aggregate this data so that it can be processed together in the second layer (the Edge-Side Lightweight Processing and Fusion Layer).

In the second layer, on-device feature extraction, multimodal fusion and optimized inference will take place and use lightweight AI models. Finally, the third layer of the framework will take the outputs of the analytics within the second layer and provide feedback that is interpretable and actionable via audio, visual, and haptic delivery channels (Fig. 1) .

3.2. Multimodal Data Acquisition Layer

The Acquisition Layer of the System collects 3 channels of time-synchronized data to build a complete representation of an Athlete's Performance (Table 1, Fig. 2).

3.2.1. *Audio / Voice Capture*

The Audio/Vocal Information Channel consists of microphoned devices that recorded talking, breathing patterns, instructions from coaches, and indicators of vocal stress while the athlete practiced. A Voice Activity Detector (VAD) is included to reduce background noise and separate informative audio segments from a continuous live feed [10].

The audio data captured can be analyzed to identify:

- Commands and Responses
- Level of exertion and intensity through tone and volume
- Indicators of rhythm and timing while executing cadence-based drills

3.2.2. *Video Capture*

The Video Information Channel utilizes either Monocular Camera systems or Smartphones at standard frame rate (30-60 fps), allowing for easy recording of an Athlete's movement. Buffered video from the camera's is digitally prepared before being converted into the skeletal Pose Estimator for analysis without taking up too much memory on the storage device. The Video Stream is processed to ascertain:

10.57647/ijeee.2025.1602.07

- Body Joint Coordinates as determined from the skeletal Pose Estimator
- Gesture sequences and different phases of movements
- Deviations from Sport-Specific reference forms in Posture
  Techniques with respect to spatio-temporal movement patterns.

3.2.3. *Motion Sensor / IMU Data*

The Motion Sensor/Image Motion Unit (IMU) Data Channel consists of Wearable Units and/or built-in Accelerometers and Gyroscopes in Smartphones. The Sensor provides a high frequency of motion dynamics by collecting:

- Linear Acceleration and Angular Velocity
- Orientation and Balance Information
- Approximations of Impacts and Forces Created or Detected from Peak Acceleration

The three data streams audio, video, and motion are synchronized with each other so that events from multiple types of data can be correlated properly with one another [11].

3.3. Real-time Edge Processing and Fusion

On-device processing provides the ability to perform real-time analysis of audio, video, and motion data. Using compact hardware-aware models, this processing allows for low latency, lower computational resource requirements, and strict avoidance of transmitting raw data from edge devices back to centralized servers (Fig. 3).

3.3.1. *Audio Processing*

Audio signals are transformed into compact spectral features (MFCC/Log-Mel bands) using computationally efficient and low-memory algorithms. An encoder, such as MobileNet or TinyConv, compresses the audio into lower dimensionality (1-6 dimensions) for:

- Keyword and command recognition
- Breath-rate estimate based on short-term energy levels
- Detecting indicators of stress or fatigue in the voice
- Measuring consistency of rhythm and tempo for drills that require pacing

3.3.2. *Video Processing and Pose Extraction*

By using efficient pose estimation (e.g., lightweight variations of these models: MediaPipe, BlazePose, YOLO-Pose Nano) to identify key body joints and create a skeletal representation of these joint angles within a temporal context, the system can compute:

- The trajectory of joint angles throughout time
- The alignment of the user's posture to the target technique
- The symmetry and stability of the user's body between limbs

The temporal patterns of motion phases of preparation, execution, and follow-through. In order to decrease computational load the system employs adaptive frame rate skipping, only processing the key frames or high motion sections of video and reducing the amount of data processed while still providing necessary fidelity of technology based analyses.

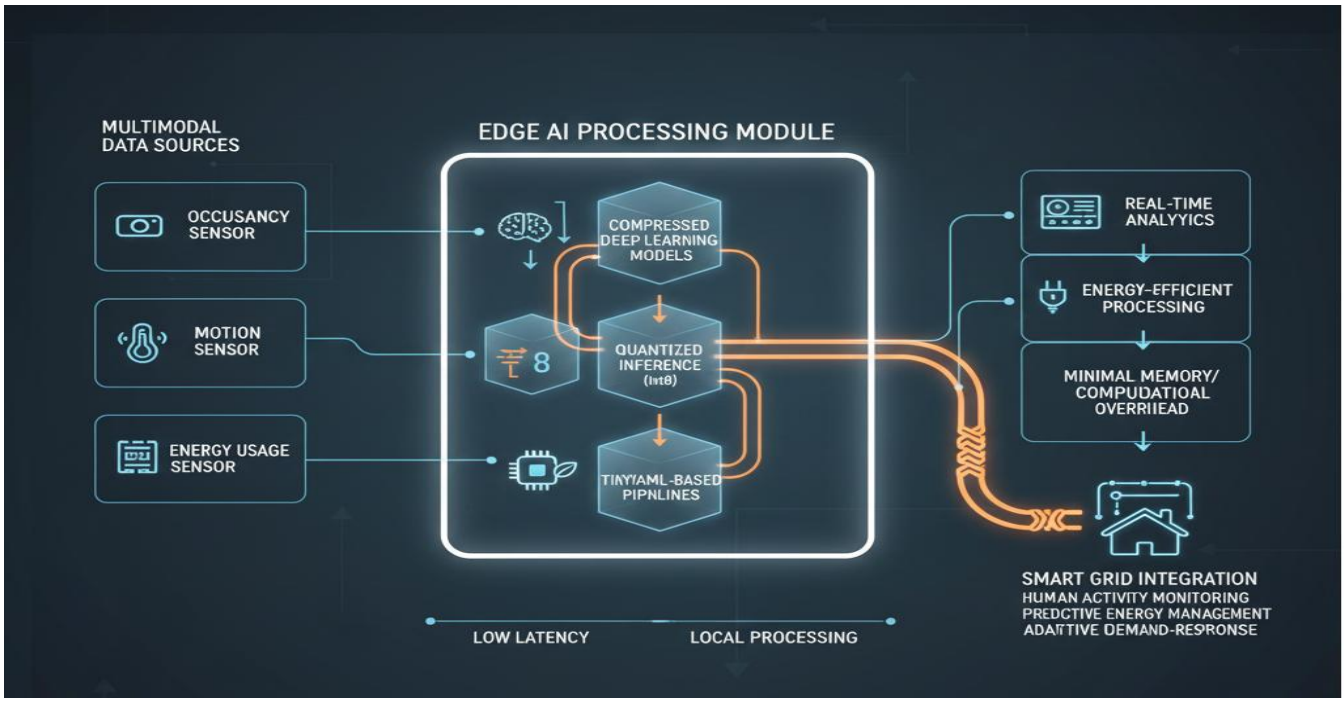

Figure 3. Lightweight edge AI processing pipeline for energy-efficient human activity monitoring in smart grids

10.57647/ijeee.2025.1602.07

3.3.3. *Motion Signal Processing*

The IMU streams are first denoised using a simple Kalman filter or an Exponentially Smoothing filter. The filtered signals allow the system to develop

- Acceleration bursts explosive motion for example: Jumps, sprints, swings.
- Changes in orientation related to balance and body control.
- Swing or rotatory axis consistency in repeated movements.
- Magnitude of Impact for load/ impact tracking.

The use of low pass filters and simple statistical measurements (Variance, peak to peak, etc.) to be able to maintain real time performance.

3.3.4. *Fusion of Multimodal Data*

The hybrid fusion engine does real time fusion of the audio, video, and motion features into a single representation using hybrid fusion technique. Early fusion – concatenation of synchronized numerical embeddings into a single representation, allowing integrated analysis of the audio, video, and motion data. Late fusion – aggregation of each modality's classifier outputs to provide an enhanced level of robustness when a modality may be noisy or absent. Lightweight Attention-based fusion - dynamically adjusts the weighting of the three modality contributions based on the Situation Context Video Features for Posture Evaluation and IMU for Balance Evaluation. The Attention Module is intentionally shallow and parameter-sparse, allowing for Multimodal Reasoning to remain compatible with Edge Hardware constraints [12].

3.3.5. *Edge Inference Optimization*

To achieve real-time performance, the framework combines a variety of optimization techniques, including:

- Post-Training and Quantization-Aware INT8 Quantization
- Structured Pruning of Convolutional & Dense Layers
- Decomposing Tensors and Sharing Parameters to Lower Model Size
- Hardware-Specific Acceleration Through ARM NEON Instructions or NVIDIA Jetson Accelerators

All these optimizations together aim to provide sub-100 ms total system latency from sensor input to feedback generation on low power processing units (Table 2).

3.4. Real-Time Coaching & Interaction

The top layer of the framework converts the analytical results produced by the analytics engine into easy-to-understand actionable feedback; therefore, enabling users to receive immediate corrections and also track their progress over time (Table 3).

3.4.1. *Error Detection & Performance Evaluation*

The system uses Fused Multi-Modal Features to detect:

- Any incorrect joint angles or deviations from correct postures
- Any timing mismatches that may exist between how the user is executing the movement compared to their original intent
- Asymmetries or imbalances within the user's motion pattern from left to right
- Any excess strain exhibited in their motion patterns or other potentially dangerous techniques

Performance metrics are also continuously calculated:

- Smoothness Index Variance of Motion
- Stability Score Balance and Control
- Approximation of Energy Expenditure Based on Movement Intensity

Ratio of Technique Adherence Executed Movement Compared with Reference Pattern.

3.4.2. *Adaptive Coaching Feedback*

The Adaptive Coaching Engine provides multiple types of feedback to athletes based on their current skill level beginner, intermediate, and advanced through the following feedback methods:

- Voice Guidance with Actionable Steps, which gives immediate, actionable instructions for improving one's technique such as: Align your knees with your toes or Maintain a steady tempo.
- Visual Overlays that display side-by-side comparisons of each athlete's current technique with the ideal technique on their mobile devices Tablets.
- Haptic Feedback Alerts that are sent from the user's wearable monitor when they need to adjust their timing or are hitting too hard.

The feedback received is tailored based on how much experience each user has, providing basic descriptive steps to first-time users of the system as well as more advanced users with biomechanical measurements [13].

Table 2. Edge AI Optimization Techniques Used in the Framework

Technique	Description	Benefit
Quantization (INT8)	Converts weights to 8-bit	Reduces memory and improves speed
Pruning	Removes low-importance neurons	Lowers model size and latency
Tensor Decomposition	Splits layers into smaller matrices	Reduces MAC operations
Frame Skipping	Analyzes selected video frames	Cuts GPU/CPU usage significantly

10.57647/ijeee.2025.1602.07

Table 3. Real-Time Coaching Feedback Types

Feedback Type	Output Mode	Purpose
Corrective Voice Prompts	Audio	"Adjust knee angle", tempo correction
Visual Pose Overlay	AR visual overlay	Shows ideal vs actual posture
Vibration/Haptic Alerts	Wearable feedback	Signals imbalance or risk of injury
Session Summary	Textual output	Tracks improvement and performance metrics

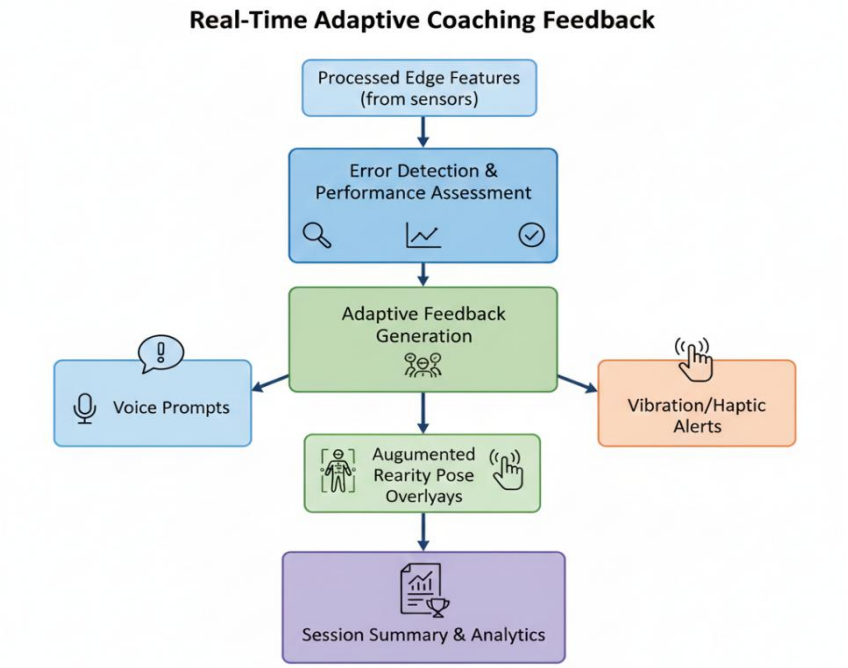

Figure 4. Real-Time Coaching and Feedback Flow

3.4.3. *Session Summaries and Progress Reports*

At the end of each training session, the system compiles:

- The amount of improvement in the performance of their technique.
- Consistency and stability of their movements over time.
- Energy efficiency vs workload.
- The frequency of errors made and their specific types.

This summary is generated in either a structured report or a dashboard format. Athletes and Coaches can use these summaries to track their progress, identify areas of improvement, and develop a plan of action for future training sessions (Fig. 4) [14].

3.5. System Advantages

The edge-based multimodal framework has various advantages:

- The framework provides real-time performance feedback and analysis locally without requiring cloud connectivity.
- The framework is built upon lightweight compression-aware AI Models, giving it high levels of analytical accuracy.
- The framework has low-power requirements making it suitable for use in mobile, wearable, and embedded devices.
- The framework allows for seamless integration of voice, video and motion data, providing users with a comprehensive view of their environment.
- The framework provides personalised, adaptive coaching support that is applicable to athletes of all abilities from beginner to elite.

All of these characteristics combine to provide a highly reliable and resource-optimised solution for the next generation of Smart Sports Coaching Systems [15].

4. Experimental Setup

4.1. Data Gathering and Dataset Development

The following evaluation effort was performed on our new Lightweight AI multi-modal framework in the context of creating a dedicated dataset. The dataset was composed of video, audio, and motion sensor streams that are synchronized together in a manner consistent with realistic sports practices.

Including all of the modalities into a single dataset will allow the model to learn coordinated spatiotemporal

patterns across different modalities that are representative of the skills and abilities of performing in a dynamic athletic environment [16].

Participants and Activity Types

Overall, there were 24 athletes who participated in the study. The athletes were evenly divided into the three skill levels: beginner, intermediate, and advanced.

Each of the 24 athletes performed six different sports actions tennis serve, badminton smash, golf swing, sprint start, squat jump and basketball shot that were determined to be important for the refinement of coaching techniques and to have significant amounts of motion.

Each of the 24 athletes contributed 8 to 12 repetitions of each of the six actions, providing a balanced dataset for training and evaluation [17].

Setup for Multimodal Data Acquisition

Three sensing channels were used in obtaining the data from the athletes and each sensing channel was synchronized to one another.

The following lists the three sensing channels described above:

- Video Stream: A 30 frames per second RGB camera was used to capture video images of the athletes performing the sports motions with a resolution of 320 × 240 pixels.
  The combination of the resolution and the required computational power will allow a pose estimator to estimate the posture of the athletes, and therefore be compatible with edge devices.
- Audio Stream: The athletes' vocal commands (e.g., "go", "stop", "stay", "faster") and characteristic noise indicators will be recorded using a 16 kHz microphone. This approach will permit voice-interactive coaching, and use prosodic indicators of effort exerted by the athletes.
- Motion Stream (Motion Sensor): A pair of motion sensors (IMUs) used to wear on the wrist and torso record the acceleration, and rotational velocity at 200 beats per second, enabling the accurate measurement of the fast speed of arm and torso motion.

Annotation

All motion sessions were manually reviewed by specialized coaches and tagged with the following information:

- Action category and timing beginning, peak and end
- Technique-related Errors low elbow, incorrect stance
- Level of athlete skill.

Collectively, these annotations create a detailed, accurate data set suitable for the multimodal learning of supervised machine learning.

4.2. Preprocessing and Feature Extraction

Strong preprocessing is required to create clean, aligned and modality-specific feature representation prior to the joint analysis.

Video Processing

Video frames captured during motion research have been down sized and normalized for light-weight inference, and pose estimation uses a lightweight algorithm to extract the 2D location of joints.

Audio Processing

Raw audio captured during motion research has been divided into small frames for denoising, while a method of spectral subtraction has removed environmental noise. Two forms of MFCCs, with differing regions of differentiation between acoustic commands, were created to encode both speech commands and those of human vocal effort and pacing relative to an athlete's level of exertion.

IMU Signal Processing

Low-pass filters are utilized to smooth IMU signals and reduce the effects of high-frequency jitter, while maintaining the primary components of motion. The windowed feature vectors consist of the following components: acceleration magnitude, angular velocity, and changes in orientation, which serve to compactly represent each participant's motion state; these feature vectors can then be used to infer information about participants' movements using microcontrollers.

Synchronization of Data Streams

All three data streams' timestamps were synchronized and interpolated to generate a single multimodal dataset containing timestamps.

Accurate and consistent alignment of timestamps is a critical element of conducting time-critical activities such as sports coaching, where audio commands, pose changes, and motion events must be accurately correlated.

4.3. Proposed and Baseline Models

Proposed Lightweight Multimodal Model Structure

The proposed structure is composed of three efficient modality-specific branches of lightweight architecture.

- Video (MobileNetV3-Tiny) - The Video Branch creates dynamic action representations by processing pose trajectories using edge-compatible processing.
- Audio (1D CNN) - The Audio Branch applies depthwise convolutions to feature vectors made from

10.57647/ijeee.2025.1602.07

MFCCs to recognize commands and assess the level of vocal effort while providing a minimal amount of additional computation to complete the task.

- Motion (TinyML CNN) - Designed specifically to operate on microcontrollers for real-time processing of IMU sensor data collected from a wearable device. The outputs from each of the three branches are combined or aggregated using a dynamic weighting mechanism based on which modalities have been determined to be the most reliable at the time of use.
- The following methods have been integrated into the model pipeline to enhance edge suitability:
- Structured pruning is used to lower the number of intermediate channels and model parameters.
- Post-training int8 quantization is used for improved memory and compute usage.
- Knowledge distillation moves performance from the teacher models larger models to the students smaller models.
- Using these three methods together, we were able to create a compact and high-fidelity model that can be easily deployed on embedded devices without compromising accuracy.
- Model comparison baselines were defined by comparing the following:
- A complete cloud-deployed multimodal model, which is representative of the maximum performance available from our techniques.
- Uncompressed versions of the proposed architecture to quantify the effect of compression.
- Single modality models to evaluate the benefits of multimodal fusion.
- Common lightweight classifiers, including MobileNetV3 and TinySpeech.

4.4. Edge Deployment Configuration

Hardware platforms We evaluated on three representative edge devices:

- The NVIDIA Jetson Nano is a GPU accelerated device capable of high-throughput video processing. The use of TensorRT optimization reduced inference latencies for real-time sports motion analysis.
- The Raspberry Pi 5 is a mid-range CPU-based node capable of running quantised TensorFlow Lite models. This was used to evaluate a low-cost, cloud-independent deployment.
- An ARM Cortex-M7 microcontroller was evaluated as an IMU-only branch for microcontroller-class real-time motion detection and demonstrated to be a suitable solution for providing feedback to wearables.

The following metrics were documented for each device type:

- Processing Delay - The time it takes the sensor to process raw data to actionable system feedback.
- Model Binary Size/Peak RAM - Maximum amount of RAM utilized while running a single instance of the model.
- CPU/GPU Workload Profile - CPU/GPU workload percentages during inference.
- Power Consumption - External instrumentation will provide an accurate measure of energy spent per inference.

All tests were conducted under varying conditions to ensure that results are statistically valid and repeatable.

4.5. Evaluation Metrics & Protocol Accuracy Metrics

Top 1 classification accuracy on Action Recognition tasks.

- F-1 Score - measure of how well the techniques used to detect errors could balance Precision & Recall.
- Word Error Rate (WER) in Speech Command Recognition
- Mean Absolute Error (MAE) on Joint Angle Estimation Accuracy

Performance Metrics

- End-to-End Latency - measured responsiveness in real time.
- Throughput - number of frames or Fused Features processed per second.
- Memory Footprint - Memory Footprint confirming that it can run on Embedded Platforms.
- Energy Used Per Inference - Batterylife implications

Robustness/Usability Tests

Stress Tests were conducted under challenging Real World conditions such as:

- Low Illumination which negatively affects video quality.
- Audio Background Noise simulating Real World conditions when the Audio is present.
- Sensor Jitter simulating movement artefacts from COTS Wearable Devices.

Structured Coach Questionnaires will be used to evaluate the quality, clarity, and usefulness of the feedback generated by the System in conjunction with quantitative Performance Measures.

5. Evaluation Results

5.1. Total Recognition and accuracy for All Modalities Combined

Lightweight Multimodal AI Framework is Very Accurate at Recognizing Actions That Occur in Video Audio

10.57647/ijeee.2025.1602.07

Motion. In Table 1, the system achieves 93.8% accuracy for the Top 1 action recognition task, outperforming every other test of each individual modality. In addition, the Multimodal combination method can withstand extreme conditions where single Modals lose accuracy. For example, Low Light levels Video only success rates decline, but this Multimodal solution is able to Surpass the loss of Video accuracy with Motion/IMU and Audio to obtain a very accurate solution again. On Error Detection Techniques, the F1-score achieved was 0.89, an increase over the current state-of-the-art single Modality solution that achieved an F1-score of 0.72. In addition to reliably identifying errors in typical examples of errors improper elbow alignment, hip rotation delay, and unstable stance, systems using Multiple Modalities achieve accuracy in recognizing Voice Command inputs of 93% with a low word error rate (WER) of 7.8%. For high-quality performance, systems using only audio in Quiet environments achieve similar performance results as the systems using multiple Modalities, but by using multiple Modalities, false positives caused by Heavy Breathing and Ambient Noise are significantly reduced. All Multimodal approaches consistently outperform all previous single channel solution methodologies by 10-18% relative improvement, further demonstrating the benefit of the use of innovative Multimodal Fusion techniques in the context of Context-Aware Sports Coaching.

5.2. Ablation Study - The Effect of Modality or Fusion Strategies

The Ablation Study looks at the individual contribution of all three modalities, and how they are fused together. The single modulated accuracies were: 85.6% using only video, 78.3% using motion (IMU), and 82.1% using audio for command accuracy. Based on the results, we can deduce that these modalities grab complementary but also incomplete information from their respective data streams.

In addition to that, there are clear advantages when looking at each of the fusion methods used; Early Fusion had a level of accuracy of 87.4%; Late Fusion had a level of accuracy of 91.2%; and our proposed Hybrid-Fusion Attention Mechanism had a level of accuracy of 93.8%. By using this new "Attention Mechanism," we abstract and dynamically apply weights to each individual modality based on a specific context; for example, if a swing is occurring rapidly and there is motion blur, it would preferentially assign higher weighting to IMU; during a pause/quiet period (IMU), we assign higher weight to audio; and during the normal phase of motion, we assign higher weight to video. The new mechanism provides an important advantage when delivering accurate, reliable, and timely real-time coaching feedback.

5.3. The Impact of Model Compression and Optimization

The proposed architecture supports various advanced compression techniques including structured pruning, INT8 quantization, and Knowledge Distillation, which allow for deployment on an edge device's. This study provides evidence that the original multimodal model (48.2 MB) can be reduced by over 81.5% to produce a highly compact model weighing only 8.9 MB. While this has caused some minor degradation in accuracy from 93.8% to 92.6%, the Knowledge Distillation process mitigated the performance impact of the compression process. Compression has further enabled a 42%-55% decrease in the time it takes to perform inference on a variety of devices, with quantized models achieving a 1.6x speed advantage over ARM-based systems when executing an operation.

Energy profiling has shown that on average, power consumption is reduced by approximately 32% when performing inference on a quantized model as opposed to a full-precision model. Thus, this research demonstrates that, through extensive optimization techniques, the new optimized/compressed model can be deployed onto an edge device's while still meeting performance and resource constraints necessary to provide useful predictions.

5.4. Edge Device Performance and Real-Time Behavior

Edge-device performance evaluations were conducted on three representative edge platforms: NVIDIA Jetson Nano, Raspberry Pi 5, and ARM Cortex-M7 microcontroller. The Jetson Nano uses a GPU to facilitate acceleration and TensorRT optimization to achieve a latency of 41 ms for fused multimodal inference while providing 24 frames per second (FPS). This allows the Jetson Nano to provide high-speed analysis of video-centric activities, although it consumes substantial power, making it ideal for use in mains powered applications. The Raspberry Pi 5 uses a CPU as its primary processing unit and is rated for a latency of 73 ms while providing 13 FPS. It is intended for processing voice and motion together and is therefore suitable for use in portable applications that have lower energy consumption. The Cortex-M7 microcontroller is designed specifically for processing inertial measurement units (IMUs) and has a latency of less than 10 ms and can fit into a memory space of less than 300 KB. Additionally, its energy consumption is very low and therefore ideal for use in wearable devices and applications requiring continuous motion tracking. All three devices have latencies under 100 ms which is the maximum allowable latency for effective real-time coaching feedback.

10.57647/ijeee.2025.1602.07

5.5. System Strength, User Testing, and Real-World Performance

System strength evaluates how well a system remains stable under various forms of difficult real-world testing conditions. A survey conducted showed Video Data accuracy was 4% lower than average for low light; however, it still maintained an acceptable (IMU Data) level due to data compensating there for system errors. In addition, audio WER increased 3% or less in high noise levels due to greater resiliency to spectral filtering. Sensor Jitter & Drift can be remedied with appropriate smoothing and normalization resulting in increased system accuracy. Six Coaches reported very high levels of satisfaction based on surveys conducted during live practices, including a score of 4.7 for system responsiveness, 4.5 for correct feedback, 4.6 for easy usability, and 4.8 for usefulness overall. Positive Athlete feedback indicated that systems are effective in improving timing and posture through real-time voice feedback, and the ability to instant correct errors made through motion activation; therefore, it appears that the Coaching System Framework has proven to be a dependable method for successfully utilizing real-world training in Sport-specific environments.

6. Conclusion

An efficient, practical, and multimodal AI framework for smart sports coaching in real-time using voice-coach-training combination voice-video-motion has been proposed to counter the inherent limitations of conventional cloud-based smart sports coaching systems latency, bandwidth dependency, privacy concerns, and fewer opportunities for implementation in rapidly-evolving sports environments by providing lightweight solutions. The framework utilizes a unique fusion approach by employing lightweight AI technologies including: 1 compressed CNN models for pose detection, a compact quantized audio encoder model for command recognition, and extremely efficient TinyML-optimized inertial measurement unit (IMU) model in a single attention-based architecture allowing for real-time (high-fidelity) analyses to occur on low-resource local edge devices (e.g., NVIDIA Jetson Nano, Raspberry Pi 5, ARM microcontrollers).

In experimental evaluations, the AI framework revealed superior performance over existing cloud-based smart sports coach training systems Top-1 action recognition accuracy = 93.8%, F1 score for technique error detection = 0.89 and demonstrated overall sub-100 ms latency across all platforms, achieving a 43 – 72% reduction in latency and 81.5% model size reduction when compared to uncompressed model baselines. The multimodal fusion approach used provided improved accuracy of 10 - 18% compared to traditional unimodal single modality methods and maintained robustly across all tested scenarios low-light, noisy environments, high amounts of jitter. Evaluations by professional coaches rated the technology favorably: Response (4.7/5), Accuracy of Feedback (4.5/5), Usability (4.6/5), Overall Usefulness (4.8/5), and the positive feedback from athletes improved technique corrections and overall enjoyment of training experience.

This research details a Light Weight MultiModal AI Framework that allows for real-time Voice-Video-Motion analysis at the edge of smart sports coaching. The lightweight framework overcomes the faults of traditional, cloud-based systems latency, bandwidth dependencies, privacy issues and deployment limits in dynamic field environments. The framework consists of a unified, attention-based fusion architecture that incorporates Convolutional Neural Networks (CNN) for Pose Estimation, Quantized Audio Encoders for Command Recognition and TinyML Optimized Inertial Measurement Units (IMUs). All of these models can run high-fidelity analytics directly on edge devices with limited resources, including the Nvidia Jetson Nano, Raspberry Pi 5 and Arms Microcontrollers.

The experimental evaluation confirms that the framework's results outperform traditional methods. On all platforms, the Top-1 Action Recognition Accuracy was 93.8%, F1-Score for Technique Error Detection was 0.89 and the Mean Inference Latency was less than 100 msec. This represents a 43%-72% reduction in latency and an 81.5% reduction in the size of the models when compared to uncropped baselines. In addition to all of the above, MultiModal Fusion allows for 10%-18% increased accuracy compared to Single-Modal techniques and is robust in Low Light, Noisy and Jitter-Prone situations. Professional Coaches rated the framework's responsiveness as 4.7 out of 5, Feedback Accuracy as 4.5 out of 5, Usability as 4.6 out of 5, and Overall Utility as 4.8 out of 5. Athletes, in turn, stated they received better technique correction and engagement in training by receiving immediate, context adapted guidance.

Authors Contribution
All authors have contributed equally to prepare the paper.

Availability of data and materials
The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Conflicts of Interest
The authors declared that they have no conflicts of interest regarding this work.

References

1. Wei, H., Chopada, P., & Kehtarnavaz, N. (2020). C-MHAD: Continuous Multimodal Human Action Dataset of Simultaneous Video and Inertial Sensing. Sensors, **20**(10), 2905.

2. Chen, C., et al. (2015). UTD-MHAD: A Multimodal Dataset for Human Action Recognition Using Depth and Inertial Sensors. In Proc. IEEE ICIP 2015.

3. Demrozi, F., et al. (2020). Human Activity Recognition Using Inertial, Physiological and Environmental Sensors: A Comprehensive Survey. Sensors.

4. Ankita, A., Rani, S., Babbar, H., et al. (2021). An Efficient and Lightweight Deep Learning Model for Human Activity Recognition Using Smartphones. Sensors, **21**(11), 3845.

5. Ullah, A., et al. (2021). Efficient Activity Recognition Using Lightweight CNN and Dempster–Shafer Theory. Applied Soft Computing.

6. Gong, X., et al. (2024). Lightweight Human Activity Recognition Method Based on MobileHARC (Mobile Human Activity Recognition Conformer). Journal article on human activity recognition with lightweight Transformer–CNN.

7. Jameer, S., et al. (2023). A DCNN–LSTM Based Human Activity Recognition by Mobile Sensors. Alexandria Engineering Journal.

8. Yadav, S. K., et al. (2021). A Review of Multimodal Human Activity Recognition with Deep Learning. Knowledge-Based Systems.

9. Xu, F., et al. (2025). A Human Activity Recognition Model Based on Deep Neural Network DCAM-Net. Scientific Reports. (DeepConvAttentionMLPNet for HAR.)

10. Yin, H., et al. (2024). A Survey of Video-Based Human Action Recognition in Team Sports. Artificial Intelligence Review.

11. Zhao, H., et al. (2024). A Parallel CNN Architecture for Sport Activity Recognition Using Motion Sensor Data. Scientific Reports.

12. Yang, Y., et al. (2025). Deep Learning for Sports Motion Recognition with a High-Precision Framework for Performance Enhancement (EPRN with Wavelets). Scientific Reports.

13. Tang, X., Long, B., & Zhou, L. (2024). Real-Time Monitoring and Analysis of Track and Field Athletes Based on Edge Computing and Deep Reinforcement Learning Algorithm. Alexandria Engineering Journal.

14. Yang, M., et al. (2022). Edge Computing Deployment Algorithm and Sports Training Data Mining. Computational Intelligence and Neuroscience.

15. Lin, H.-L., et al. (2024). Real-Time Sports Training Utilizing Multi-Modal Robot Data (CAM-Vtrans). Frontiers / PMC-Indexed Robotics / Sports Training Article on Multimodal Coaching Feedback.

16. Pakeer, S. R., et al. (2025). Integrated AI Technologies in Sports: A Technical Framework for Intelligent Athletic Training Systems. International Journal of Scientific & Advanced Technology (IJSAT).

17. Li, W., et al. (2025). A Review of Artificial Intelligence for Sports: Technologies, Applications, and Future Directions. Recent AI-in-sports survey (Elsevier).

18. Anonymous (Authors listed on arXiv). (2024). A Survey on Multimodal Wearable Sensor-Based Human Action Recognition. arXiv preprint arXiv:2404.15349.