Master's Thesis

Multi-Modal Deep Learning for Badminton Winning Prediction Point-by-Point

Yutong WANG

Supervisor Professor Shogo Okada

Division of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

September, 2025

Abstract

Predicting outcomes in competitive sports is a complex and high-impact task that benefits from recent advances in deep learning and multi-modal data analysis. However, fine-grained outcome prediction in sports like badminton remains underexplored, especially at the point-by-point level, where rapid player movements and subtle contextual cues make prediction challenging. This study aims to address this gap by developing a model that can accurately predict point outcomes using lightweight, scalable inputs derived from broadcast videos.

To this end, we propose a novel point-by-point winning prediction model for badminton, leveraging two complementary modalities: player posture and environmental audio. The Badminton Winning Prediction dataset, consisting of 12 full-length singles matches, was manually annotated and preprocessed to extract synchronized features from both video and audio streams.

Our proposed architecture consists of two key components. First, a Cross-Modal Fusion Module integrates Mel-Frequency Cepstral Coefficient (MFCC) and delta features extracted from environmental audio, with 2D body keypoints extracted using MMPose, through a bi-directional cross-modal attention mechanism, to form stage-level representations. Second, a Cross-Stage Fusion Module combines rally and non-rally segments via attention and gated mechanisms to capture temporal interactions across gameplay phases. The fused features are then aggregated over a sliding window of consecutive points and passed into a classification head (Linear, Long Short-Term Memory (LSTM), or Transformer) for point-wise outcome prediction.

Extensive experiments show that our multi-modal fusion strategy significantly outperforms single-modal baselines and simple fusion methods. Notably, the inclusion of non-rally segments and audio information contributes to improved accuracy and F1-score, highlighting the value of contextual cues beyond active play. Our best model achieves an accuracy of 88.75% in binary classification, demonstrating the practical feasibility of fine-grained prediction using lightweight and scalable inputs.

These results demonstrate that our multi-modal fusion approach effectively captures predictive information in badminton matches. In particular, even a simple linear classifier achieves excellent performance, highlighting the strength of the fused features. The significant improvements gained by incorporating audio cues and non-rally segments confirm the importance of contextual information beyond active play.

The proposed framework not only advances badminton analytics but also offers a generalizable approach for other racket sports such as tennis and table tennis, especially in scenarios with limited camera views or sensor inputs.

Keywords: Badminton, Multi-modal Fusion, Winning Prediction, Attention Mechanism, Sports Analytics

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Okada, for his continuous support, insightful guidance, and invaluable feedback throughout my research and academic journey at Japan Advanced Institute of Science and Technology (JAIST). His patience and encouragement have been a constant source of motivation.

I would also like to thank all the faculty members and staff in JAIST for their helpful discussions and generous assistance. My sincere appreciation extends to the members of the Okada Laboratory, whose camaraderie and collaboration have made my research experience both enriching and enjoyable.

I am especially grateful to our laboratory Assistant Professor, Candy-sensei and Li-sensei, for guiding me through the research process and offering kind and practical help whenever I encountered difficulties. Their mentorship and encouragement have played an important role in my growth as a researcher.

I am grateful to the Japan Advanced Institute of Science and Technology for providing an excellent academic environment and facilities that enabled me to carry out this work.

Special thanks go to my family for their unwavering support and understanding, especially during the difficult times. Their love and belief in me have been the cornerstone of my perseverance.

Finally, I would like to thank all my friends and colleagues who have accompanied me on this journey. Your encouragement and inspiration have meant a great deal to me.

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence

BWF Badminton World Federation

BWP Badminton Win Prediction

CLSTM Chained Long Short-Term Memory

CNN Convolutional Neural Network

DCT Dscrete Cosine Transform

HPE Human Pose Estimation

IoU Intersection over Union

JAIST Japan Advanced Institute of Science and Technology

LSTM Long Short-Term Memory

MFCC Mel-Frequency Cepstral Coefficient

NBA National Basketball Association

SHAP SHapley Additive exPlanations

Chapter 1

Introduction

1.1 Background and Significance

The development of Artificial Intelligence (AI) has spanned over seven decades since Alan Turing proposed the concept in 1950. Interestingly, the first known system for sports analysis is believed to have emerged around 1951, marking the inception of AI applications in sports. Since then, the rapid advancement of AI technologies has had a profound impact on the sports industry. AI has now been tightly integrated with a wide range of sports, offering invaluable support to coaches, athletes, referees, and spectators. These technologies have made sports safer, fairer, smarter, and more engaging for audiences [11].

In everyday exercise and fitness activities, AI plays a crucial role. The widespread adoption of wearable devices has made it possible to collect physiological signals from the human body, enabling real-time monitoring of exercise conditions and the development of healthier training plans [13]. Beyond recreational fitness, AI applications in professional sports events are even more significant. For example, in National Basketball Association (NBA) games, multi-view 3D reconstruction systems provide immersive replay experiences for spectators and serve as a visual aid for referees in blind spots [8]. In football, AI-driven visual analysis offers real-time match insights to coaches, helping devise targeted strategies during games and review performances afterward [28]. In training scenarios, motion capture and recognition technologies assist athletes in identifying incorrect movements, allowing for more effective skill refinement [18].

However, single-modal information has inherent limitations in its descriptive power. To obtain more comprehensive insights, multi-modal AI technologies have increasingly been adopted in sports. For instance, combining visual

data with physiological sensors allows for accurate fatigue detection, reducing the risk of injuries caused by over-training [15]. The integration of audio, visual, and positional signals provides a more complete understanding of athletes' physical and psychological states, improving coaching effectiveness and analytical precision.

The proliferation of advanced sensors and sophisticated data processing techniques in sports has also led to a surge in time-series data generation [16]. This data, representing sequences of athlete behaviors, game events, or environmental conditions, opens new avenues for time-series modeling. Such models can uncover underlying patterns and trends in performance, strategy, and training plans. More importantly, time-series models enable the forecasting of future events or outcomes, transforming traditional approaches to match prediction, performance optimization, and injury prevention.

Among these applications, match outcome prediction is one of the most impactful and challenging problems. In sports such as tennis, time-series models have already been applied with success [23]. A match's outcome is primarily influenced by a player's performance (e.g., posture or movement) but is also affected by environmental factors, such as crowd noise or atmosphere. Therefore, it is essential to integrate both athlete behavior and contextual environmental signals into a unified multi-modal framework for accurate prediction.

Winning prediction is a crucial task in sports analytics. However, most existing approaches focus on analyzing the outcomes of previous games, often relying on post-match statistics or video analysis [12]. This retrospective nature limits their applicability during ongoing matches, making point-by-point prediction especially challenging. As the fastest racket sport, badminton poses a distinct challenge, with rapid exchanges and short rallies that require precise, real-time data analysis.

1.2 Research Problem

Despite the increasing application of AI in sports analytics, several unresolved challenges remain, particularly in the context of point-by-point prediction for fast-paced games like badminton. First, the majority of existing methods concentrate on full-match outcome prediction. These models typically rely on post-match statistics or highlight data, which limits their applicability in live match scenarios. Real-time or intra-match prediction, especially at the level of individual points, remains underexplored.

Second, while many studies utilize visual information such as player position, ball trajectory, or action recognition, they often overlook the potential of

multi-modal data integration. Crowd noise, environmental sounds, and even players' vocal reactions can reflect momentum shifts, psychological states, or fatigue levels, all of which may influence game outcomes. Ignoring such auditory information reduces the depth and accuracy of predictive models.

Third, badminton itself poses unique challenges. Unlike tennis or football, badminton rallies are extremely short, with intense back-and-forth exchanges and frequent shifts in rhythm. This fast-paced nature makes it difficult to extract and process meaningful temporal features from a single modality. Furthermore, due to the dominance of single-camera footage in publicly available datasets, it becomes crucial to develop methods that can operate effectively under limited visual perspectives while still leveraging multi-modal information.

Therefore, the central research problem of this thesis lies in how to effectively fuse visual and auditory modalities in a time-series framework to enable point-by-point winning prediction in badminton matches using limited-view data.

1.3 Research Objectives

This thesis aims to propose a novel multi-modal point-by-point winning prediction model tailored to badminton, a sport with rapid dynamics and limited visual coverage. The first objective is to extract player posture information from single-camera video footage using pose estimation techniques. This allows for capturing subtle changes in athlete movement and posture over time, which are indicative of their performance and physical condition.

The second objective is to incorporate contextual auditory features by extracting sound-based information such as crowd reactions, racket sounds, and court ambiance. These signals provide additional insights into match intensity and situational dynamics. By aligning the audio and visual data streams across temporal sequences, the model seeks to exploit both modalities for richer predictive power.

A further objective is to develop and evaluate a multi-modal fusion framework capable of learning joint representations from both modalities. This framework will employ deep learning techniques to handle sequential input and output a binary classification result for each point, predicting whether the player of interest will win or lose that point.

Ultimately, this research aims not only to improve point-level prediction accuracy in badminton but also to provide a generalizable multi-modal modeling approach that can be adapted to other racket sports such as tennis or table tennis, particularly in scenarios with constrained video sources or

real-time requirements.

1.4 Originality

While prior studies have made significant progress in sports analytics, most existing models either rely on single-modal information or primarily focus on specific gameplay phases such as rally segments. For instance, Yu et al. [23] introduced Chained Long Short-Term Memory (CLSTM), a chained LSTM-based model for predicting stage-wise winning percentages in tennis, capturing the sequential nature of match progression. Similarly, Nitin et al. [19] proposed an end-to-end system for analyzing player movements based solely on video data during rally periods in badminton.

In contrast, this research presents a novel point-by-point prediction framework tailored specifically for badminton, which inherently involves rapid and complex shifts between rally and non-rally stages. The proposed approach integrates multi-modal data—including visual features (player posture and motion) and auditory cues (audience response, environmental noise)—to achieve a more holistic representation of match dynamics. Furthermore, the model introduces a new cross-modal global-local fusion mechanism that adaptively focuses on modality relevance depending on the game stage, thereby enhancing prediction accuracy across diverse contexts. This stage-aware, multi-modal fusion strategy represents a significant departure from existing methods and contributes a new direction for fine-grained performance analysis in racket sports.

1.5 Organizations of Thesis

This thesis is organized into six chapters as follows:

Chapter 1 introduces the background and significance of the research, defines the core problem, sets out the research objectives, and highlights the originality of this work.

Chapter 2 presents a literature review covering the foundations of badminton, existing research on win prediction in racket sports, and relevant methods in pose estimation and multi-modal learning, which collectively form the basis for this study.

Chapter 3 describes the process of constructing the Badminton Winning Prediction (BWP) dataset, including data collection, manual annotation, feature extraction, and preprocessing steps.

Chapter 4 describes the proposed prediction model in depth, including

its overall architecture, the cross-modal and cross-stage fusion modules, and three types of classification heads (Transformer block, LSTM, and linear classifier). Baseline methods are also introduced for comparative analysis.

Chapter 5 describes the experimental setup, main results, and ablation studies. It includes comparisons of fusion methods, classification strategies, modality contributions, and input stage designs. A discussion of the results highlights the implications and effectiveness of the proposed approach.

Finally, Chapter 6 summarizes the main contributions of this work, discusses its limitations, and proposes directions for future research in the domain of multi-modal winning prediction in racket sports.

Chapter 2

Literature Review

2.1 Foundation of Badminton

2.1.1 Badminton Rules

Badminton is a fast-paced racket sport played in either singles (one player per side) or doubles (two players per side), as defined by the Badminton World Federation (BWF) [4]. In this thesis, we focus exclusively on singles matches. The game is played on a rectangular court divided by a net, and players aim to score points by hitting the shuttlecock over the net into the opponent's side such that it cannot be returned [4].

Matches follow the 3-games $\times$ 21-points rally scoring system, officially adopted by the BWF in 2006: each game is played to 21 points, with every rally awarding a point regardless of serve. A player must win by at least a 2-point margin, and if the score reaches 29-29, the first to reach 30 wins that game [2].

Each rally begins with a serve, and ends when the shuttlecock hits the ground or a fault is committed. Serving alternates based on rally winners: if the server wins, they continue serving; otherwise, the receiver becomes the server [2]. Matches are officiated under BWF Laws of Badminton by a chair umpire and service judge, who enforce rules and declare faults [32].

Additional operational rules include: the winner of each game serves first in the next game; players change ends after the first game, and in the third game when one side reaches 11 points; and the serve must be delivered underarm, shuttle height below 1.15m, feet stationary until contact, and no undue delay in service motion as per Laws 7–9 of the official statutes [32].

2.1.2 Characteristics of Match Stages

A badminton match can be conceptually divided into two alternating stages:

Rally Stage

The rally stage refers to the active play period when the shuttlecock is in motion, starting from the serve until the point ends [19]. During this stage, both players are moving quickly, executing strokes, footwork, and positioning. This is where the most dynamic and informative actions occur in terms of player performance, strategy, and physical execution.

Non-Rally Stage

The non-rally stage refers to the period between two rallies [19]. It includes the time immediately after a point ends and before the next serve begins. During this stage, players may:

- Retrieve the shuttlecock
- Walk back to position
- Wipe sweat, adjust equipment
- Receive verbal guidance or instructions from the umpire

Although this stage contains less visible activity, it includes useful contextual information:

- Auditory signals such as umpire announcements of scores
- Verbal cues indicating fouls or service faults
- Emotional or fatigue-related body language

In our model, both stages are treated as equally important components. Each stage (rally + preceding non-rally) is grouped and labeled with a win/loss outcome depending on which player scored in the rally. This grouping enables the model to learn predictive cues not only from rally dynamics but also from transitional moments and acoustic context.

2.1.3 Scoring and Victory Conditions

To summarize the core mechanics:

- Each game is played to 21 points, win by 2.
- If 29–29, the player who scores the 30th point wins the game.
- A match is best of 3 games.
- Points are awarded at the end of each rally (rally point scoring system).
- The serve changes sides based on rally outcomes.

The combination of technical play during rallies and strategic behavior during non-rally stages creates a rich set of modalities from which winning potential can be inferred.

2.2 Review of Machine Learning Applications in Sports and Badminton

Zhao et al. [31] conducted a comprehensive survey on the applications of machine learning in sports. They reviewed a wide range of machine learning algorithms, datasets, and virtual environments currently used in the sports domain. Their study covered more than ten different sports and the corresponding machine learning methods applied to each. Furthermore, Zhao et al. [31] expressed optimistic prospects for future research in multimodal approaches, practical applications, and synthetic data generation in sports analytics.

Tan et al. [5] conducted a comprehensive survey on the analysis of the sport of badminton. With the development of technology, manual analysis of badminton players has been gradually replaced by advanced methods such as badminton smashing analysis, badminton service recognition, and badminton swing and shuttle trajectory analysis.

Building on these foundational studies, recent research has increasingly explored the integration of multiple data modalities—such as video, audio, and sensor signals—to enhance the accuracy and robustness of sports performance analysis. Multimodal deep learning has shown potential in capturing complex interactions between players and their environments, enabling more fine-grained predictions of match outcomes and player behavior.

Motivated by these trends, the present study proposes a multimodal deep learning framework for point-by-point prediction of badminton match outcomes. By combining pose and audio features, this work contributes to

the expanding field of intelligent sports analytics through fine-grained, data-driven modeling.

2.3 Winning Prediction in Racket Sports

2.3.1 Tennis Winning Prediction

Kovalchik evaluated the predictive performance of 11 published models using data from 2,395 singles matches, offering guidance for the future development of tennis win prediction models [3]. His study compared multiple machine learning methods, including logistic regression and ELO-based approaches, revealing that probabilistic models often outperform naive baselines in professional tennis.

Gao et al. compiled, cleaned, and utilized the largest known tennis match database to date, employing the Random Forest method to predict match outcomes [10]. They extracted over 20 features including player rankings, recent performance, and surface type. Their results showed that ensemble methods provide better generalization and robustness in noisy sports environments.

Yu et al. proposed combining a chained model with an LSTM network and introduced a novel dynamic model (CLSTM), as shown in Figure 2.1, which leverages the sequential interdependence of matches to predict the winning probability at each stage of a tennis match [23]. Their framework incorporated time-dependent player state features and demonstrated superior performance in stage-wise predictions compared to conventional static models. However, a notable shortcoming is that by focusing solely on sequential dependencies within rallies, the model overlooks potentially informative relationships that occur outside of rally segments, such as between-point strategies or momentum shifts.

2.3.2 Badminton Winning Prediction

Sharma et al. proposed a badminton match result prediction model using the Naive Bayes method trained on historical match records [12]. Although simple, their model demonstrated that probabilistic learning can capture basic outcome patterns based on past matchups, player names, and win-loss ratios.

Jo et al. developed a sequential win probability prediction model based on the Extended Sequential Probability Ratio Test (EXSPRT) [30]. They introduced a scoring difficulty index calculated from technical, situational, and

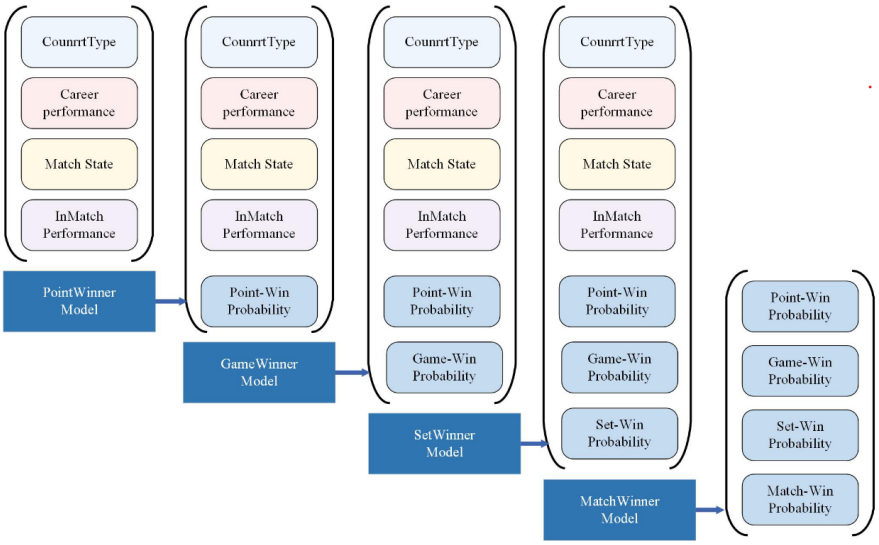

Figure 2.1: The Work Flow of CLSTM [23].

temporal indicators such as shot type, rally length, and current score. This model dynamically adjusted win probabilities after each rally and showed improved interpretability over traditional classification methods. The framework, shown in Figure 2.2, emphasizes the importance of rally context in point-level prediction.

Yuan et al. focused on elite player An Se-young and trained machine learning models to classify rally outcomes (win/loss) in women's singles [22]. They extracted handcrafted features such as stroke type frequency, player movement speed, and opponent error statistics. Among several classifiers, SVM and XGBoost achieved high accuracy, highlighting the viability of feature engineering in badminton analysis.

Sheng et al. designed a recognition system for 23 badminton technical actions using vision-based learning and applied the recognized actions in match outcome prediction [27]. Their pipeline used convolutional networks for action detection and integrated temporal dynamics to link action sequences with game results. This study demonstrated the potential of fine-grained action recognition in supporting tactical analysis and prediction.

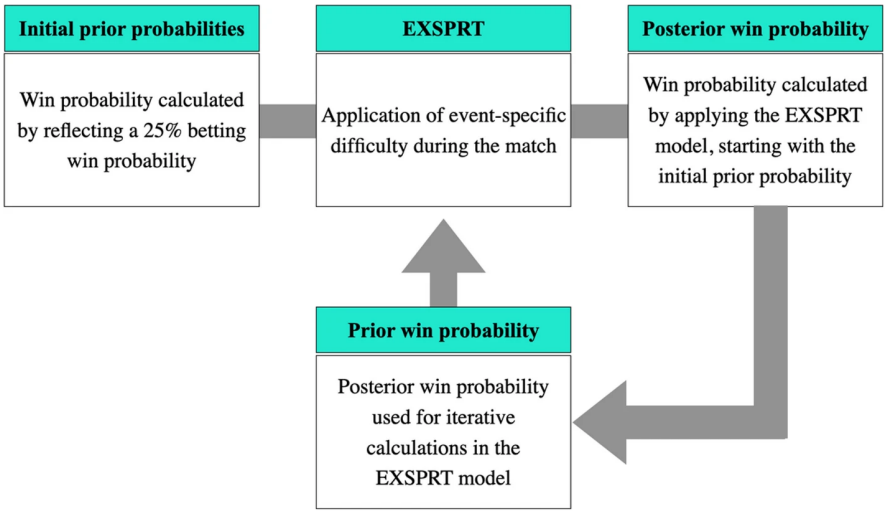

Figure 2.2: Sequential Win Prediction Framework using EXSPRT [30].

2.4 Pose Estimation in Sports and Badminton

Pose estimation tools such as OpenPose [7] and MMPose [17] extract 2D keypoints from videos and have been widely applied in sports analytics. These methods are particularly useful for detecting joint trajectories and player movement patterns.

Figure 2.3: Example of pose estimation in a badminton match using OpenPose [7].

In sports, pose-based analytics are used for performance evaluation, tactical modeling, and injury prevention [14]. Badiola-Bengoa et al. conducted a comprehensive survey of Human Pose Estimation (HPE) applications in sports and highlighted its potential for both elite and amateur training [9].

In the context of badminton, Jiang et al. proposed a lightweight pose estimation-based training system by improving OpenPose, specifically tailored to support technical skill evaluation and motion correction in badminton drills [29].

Figure 2.3 shows an example of OpenPose applied to a real-world badminton match, where it successfully detects body keypoints of both players and surrounding personnel. Such visualizations can aid in extracting pose features, tracking movement, and understanding player behavior for performance analysis.

2.5 Multi-modal Learning in Sports Analytics

The integration of multiple data modalities—such as visual, audio, textual, and positional information—has become increasingly common in sports analytics due to its potential to capture complementary aspects of player behavior and game context.

As shown in Figure 2.4, Goka et al. [21] proposed a model to predict shooting events in soccer by jointly processing visual and audio cues. The visual stream consisted of frame-level spatial features, while the audio stream captured crowd reactions and environmental sounds. Their model used a temporal convolutional network to encode each modality and then fused them via concatenation, leading to improved event recognition accuracy compared to unimodal baselines.

Takamido et al. [24] developed an interpretable AI system called PassAI to classify successful and failed passes in professional soccer. Their approach integrated player tracking trajectories with player statistics (e.g., pass accuracy, stamina) and applied SHapley Additive exPlanations (SHAP) values to identify which features influenced the model's predictions. This work highlighted the benefits of combining spatial and tabular data for tactical analysis and model explainability.

Multi-modal learning enables models to extract richer representations and improve generalization, especially in scenarios where single-modal inputs may be insufficient or ambiguous.

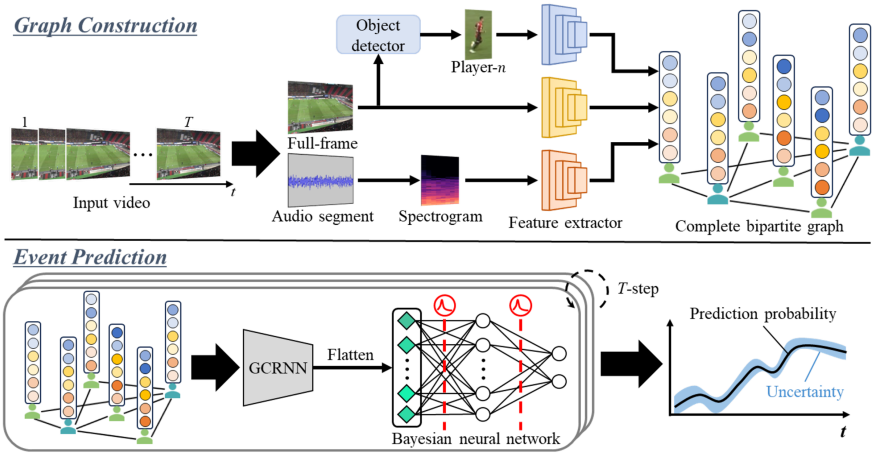

Figure 2.4: Overview of multi-modal shot prediction in soccer videos [21].

2.6 Fusion Strategies and Attention Mechanisms

Fusion methods in multi-modal systems are generally categorized into early fusion (feature-level), late fusion (decision-level), and hybrid strategies that combine both [20]. In sports analytics, early fusion is often used when temporal alignment is tight (e.g., synchronized video and audio), while late fusion allows for flexible integration across asynchronous modalities.

To dynamically weigh the relevance of different modalities, attention mechanisms have been increasingly adopted. For example, As shown in Figure 2.5, Li et al. [26] proposed a cross-modal attention framework for emotion recognition, where visual features guided the selection of salient audio segments, improving model robustness to noise. Their architecture included a dual-encoder with self-attention and a cross-modal fusion layer, leading to significant improvements over late fusion baselines.

Xv et al. [25] tackled highlight localization in long-form sports videos by aligning visual and textual fragments. They used a transformer-based architecture with query-guided attention to link semantic keywords with video snippets. Their results demonstrated that attention-based fusion outperforms naive alignment strategies in complex, multi-event scenarios.

These findings highlight the importance of context-aware fusion in handling heterogeneous modalities. In badminton, where the game context rapidly alternates between high- and low-activity segments, adaptive attention mechanisms can enable the model to focus on pose features during rallies and audio cues during breaks, motivating the design of our proposed Bad-

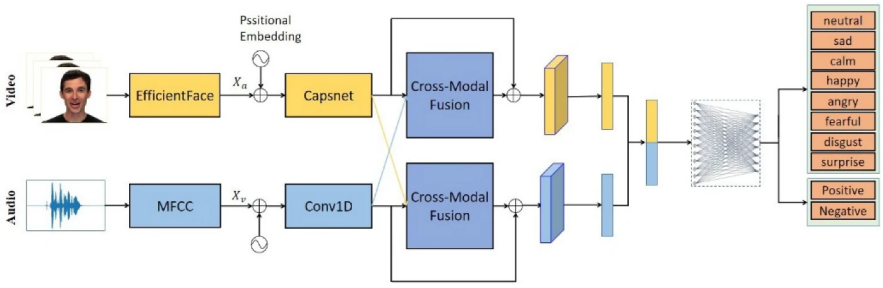

Figure 2.5: CCMA: CapsNet for audio–video sentiment analysis using cross-modal attention [26].

minton Winning Prediction Model framework.

Chapter 3

Badminton Winning Prediction Dataset Creation

3.1 Dataset Overview

To train and evaluate our badminton winning prediction model, we constructed a Badminton Win Prediction (BWP) dataset. The dataset consists of 12 full-length singles matches collected from the official YouTube channel of the Badminton World Federation (BWF TV) [33]. These include 6 men's singles and 6 women's singles matches, each approximately 50 minutes in duration, with clear video and audio quality suitable for multi-modal feature extraction.

Each match was manually segmented into alternating **rally** and **non-rally** stages. Each pair of consecutive non-rally and rally stages was grouped into a stage and labeled with a win/loss outcome depending on which player scored the point. After cleaning, filtering, and annotation, we obtained a total of **1702 labeled stages** (818 positive and 884 negative samples).

The total dataset duration is approximately **9 hours 21 minutes 49 seconds**. The detailed breakdown of match durations, scores, and sample counts is shown in Table 3.1. The average length of rally and non-rally stages is shown in Table 3.2.

3.2 Dataset Creation and Preprocess

The process of dataset creation and preprocessing is shown in the Figure 3.1.

Table 3.1: Sample counts per match (window size $L = 3$)

Match Name	Game Scores	Samples
BANSOD vs NGUYEN (QF)	21–15, 21–17	68
AN vs GAO (RO32)	21–16, 21–14	66
INTANON vs LI (RO32)	21–12, 21–6	54
MIYAZAKI vs NIDAIRA (RO32)	21–13, 21–17	66
SUNG vs BLICHFELDT (QF)	14–21, 14–21	64
WARDANI vs OHORI (RO32)	21–19, 21–18	73
Total of Women-Singles		**391**
SHETTY vs KOLJONEN (QF)	21–18, 21–18	72
CHOU vs GEMKE (QF)	13–21, 20–22	70
JAKOBSEN vs SANTHOSH (QF)	21–12, 21–17	65
POPOV vs LI (QF)	15–21, 18–21	69
WENG vs NISHIMOTO (QF)	21–10, 21–16	62
YANG vs GEMKE (RO32)	15–21, 11–21	62
Total of Man-Singles		**400**
Total		**791**

Table 3.2: Aggregate and average duration of rally and non-rally stages

Stage Type	Total Duration (s)	Average Duration (s)
Non-Rally	24009.93	27.82
Rally	9692.92	11.23

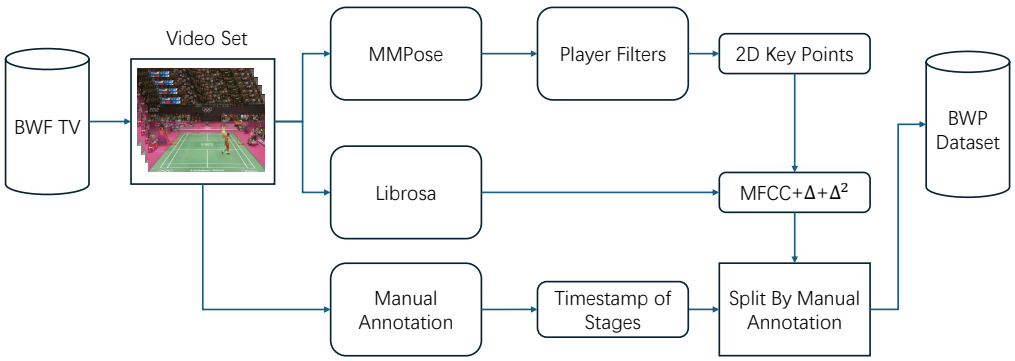

Figure 3.1: Overall process of The BWP Dataset creation.

3.2.1 Source and Format of Raw Data

All match videos were collected from the official BWF TV YouTube channel [33], which broadcasts full-length replays of professional badminton tour-

naments. To maintain continuity and minimize interference from camera switching, only matches filmed with a single fixed-angle camera were selected. All videos are in MP4 format with 1080p resolution, recorded at 30 frames per second (fps), and contain 44.1 kHz stereo audio.

3.3 Manual Annotation

3.3.1 Label Definition

Each complete stage is defined as a pair of consecutive stage: one **non-rally** stage followed by a **rally** stage. For each stage, we assign a binary label (1 or 0) indicating whether the player on the camera-facing side wins the upcoming rally. Specifically, label **1** denotes that the camera-side player wins the point, and label **0** indicates a loss. This labeling standard is consistently applied regardless of court switching between games. During dataset loading, stage sequences do not span across game boundaries, so the definition of the camera-side remains consistent.

In total, approximately 800 labeled stages were annotated across 12 matches. The distribution of win/loss labels is nearly balanced, as shown in Table 3.3.

3.3.2 Annotation Strategy and Tool

Manual annotation was conducted using the ELAN tool [34], developed by the Max Planck Institute for Psycholinguistics. ELAN is widely used in academic fields such as psychology, linguistics, education, and behavioral science. It supports multi-tier annotations for audio and video, and stores annotations in an XML-based format.

Using ELAN's segmentation interface, we manually marked the start and end times of both non-rally and rally intervals. Each pair of non-rally and rally stages together forms one annotated stage, and the corresponding label applies to the entire stage. Figure 3.2 shows a screenshot of the ELAN interface during the annotation process.

To ensure consistency, only the camera-facing player was labeled and tracked throughout the match. Even when players switched courts between games, we retained the labeling focus on the camera-side player without adjustment.

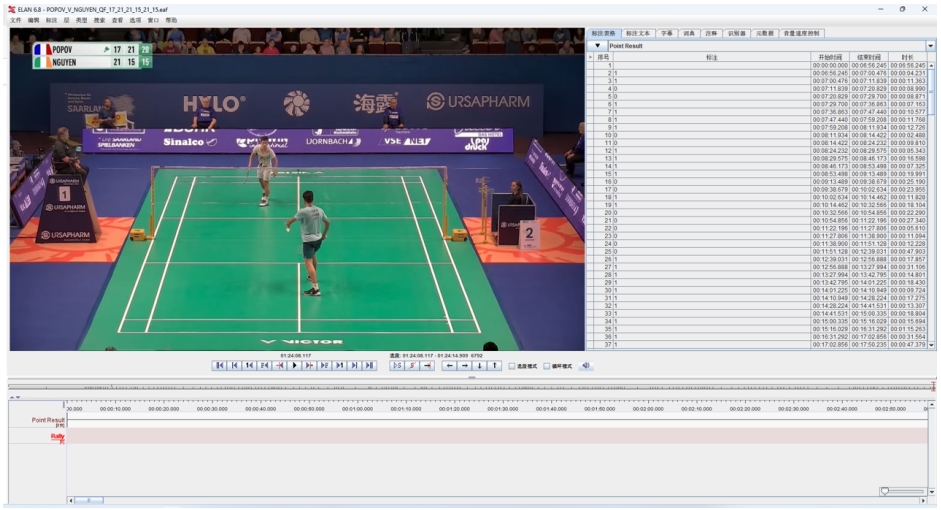

Figure 3.2: Manual annotation of rally and non-rally stages using ELAN [34].

3.3.3 Label Distribution

After manually annotating each stage of the 12 matches with a binary win/loss label, we obtained a total of 1,702 labeled samples. A complete stage—consisting of both the Rally and Non-Rally segments—shares the same label. Table 3.3 shows the distribution of positive (win) and negative (loss) labelswith separate counts for each non-rally and rally label. The dataset remains reasonably balanced, with a slight skew toward negative samples.

Table 3.3: Label distribution in the annotated dataset

Label	**Count**
Positive (Win)	818
Negative (Loss)	884
Total	**1702**

3.3.4 Audio Feature Extraction

The audio environment in professional badminton matches mainly consists of player shouts, racket hitting sounds, and intermittent audience reactions. These elements contribute informative acoustic cues that correlate with player actions and rally intensity. To capture this information, we extracted MFCC, a widely used feature representation in speech and sound analysis.

Audio was extracted from the original video and convert stereo to mono. We then computed 13-dimensional MFCCs along with their first-order (Δ) and second-order (Δ^2) temporal derivatives, resulting in a 39-dimensional audio feature vector for each frame. The features were extracted using the `librosa` library.

The MFCCs are computed by applying the Dscrete Cosine Transform (DCT) to the log power of Mel-scaled filter bank energies:

$$MFCC_n = \sum_{k=1}^{K} \log(E_k) \cdot \cos\left[\frac{\pi n}{K}(k - 0.5)\right], \tag{3.1}$$

where E_k is the energy in the k-th Mel filter, K is the total number of Mel filters, and n is the MFCC coefficient index.

The first- and second-order deltas are computed as:

$$\Delta_t = \frac{\sum_{n=1}^{N} n \cdot (c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}, \quad \Delta_t^2 = \frac{\sum_{n=1}^{N} n \cdot (\Delta_{t+n} - \Delta_{t-n})}{2\sum_{n=1}^{N} n^2}, \tag{3.2}$$

where c_t is the MFCC at time t, and N is the window size for delta computation.

To align audio frames with video frames (30 fps), we computed the `hop_length` based on the sampling rate (44,100 Hz) and video frame rate (30 fps), as follows:

$$\texttt{hop\_length} = \frac{\texttt{sampling\_rate}}{\texttt{fps}} = \frac{44100}{30} \approx 1470. \tag{3.3}$$

This ensures that the number of audio frames extracted per second matches the number of video frames, enabling frame-level alignment between audio and pose modalities.

The MFCC extraction code is as follows:

```
mfcc = librosa.feature.mfcc(y=y, sr=44100, n_mfcc=13,
    hop_length=1470)
delta = librosa.feature.delta(mfcc)
delta2 = librosa.feature.delta(mfcc, order=2)
```

3.3.5 Posture Feature Extraction

We used the `MMPose` [17] library to extract 2D body keypoints from each frame of the match videos. Specifically, the `human` preset was used, which

employs the RTMPose-m pose estimation model with RTMDet-m for person detection. The extracted output includes 17 keypoints per person in (x, y) format, along with confidence scores and bounding boxes.

Since each frame may contain multiple human detections, we designed a tracking-based filtering algorithm to extract only the two main players across the full match. The core idea is to select the two players closest to the court center in a stable frame and then track them forward and backward using bounding box Intersection over Union (IoU). If tracking fails, a fallback selection mechanism is applied.

Initial Frame Selection We set the initial reference frame to frame 9000 (corresponding to approximately 5 minutes in a 30 fps video). Empirically, this frame always falls within the rally stage when both players are clearly visible and positioned near the center of the court. This ensures stable initialization of player tracking across different videos.

The overall filtering process is described in Algorithm 1.

Algorithm 1: Player Filtering via Stable Tracking

Input: Keypoint JSON file containing all detected instances per frame
Output: Filtered JSON file with two main players per frame
Set `start_frame` $\leftarrow$ 9000;
Select two players closest to image center at `start_frame` as `stable_players`;
for *each frame f from* `start_frame` *to first frame (in reverse)* **do**
    Match instances in f with `stable_players` using IoU;
    if *match fails for k consecutive frames* **then**
        Select two closest instances to screen center (fallback);
        Reset failure counter;
    else
        Update `stable_players` with matched instances;
    Save matched/fallback players for frame f;
Repeat the same process forward from `start_frame` + 1 to last frame;
Sort all frames and write filtered JSON output;

After this filtering process, only the two main players are retained per frame. The final posture feature used for modeling is a tensor of shape $[2, 17, 2]$, representing 17 keypoints for each of the two players.

3.3.6 Feature Synchronization and Normalization

To ensure consistent input lengths, we predefined the input duration for both non-rally and rally stages. Initially, we set the input length to 500 frames, and extracted the first 500 frames from each stage. However, we later revised this strategy based on empirical observations. Specifically, we updated the fixed input length to 600 frames per stage, which better matched the average duration statistics reported earlier (see Table 3.2).

In addition, we changed the truncation strategy from head-cut (selecting the first 500 frames) to tail-cut (selecting the last 600 frames). This adjustment ensured that the end of each stage—often containing decisive movements or acoustic signals—was retained for learning.

Chapter 4

Badminton Winning Prediction Model

4.1 Model Overview

This chapter presents our proposed model for point-by-point badminton win prediction, which is designed to leverage multi-modal information—specifically, environmental audio and player posture—across different stages of each point. The model incorporates two key components to capture both intra-stage and inter-stage dependencies: the Cross-Modal Fusion Module and the Cross-Stage Fusion Module. The full architecture is illustrated in Figure 4.1.

For each point, we partition the gameplay into non-rally and rally stages and extract two types of features from each stage: (1) audio features (MFCC, Δ, and Δ^2), and (2) 2D keypoint-based posture features from both players. The audio features are encoded by a Transformer-based encoder, while the posture features are processed through a bi-directional cross-attention mechanism to capture the interaction dynamics between the two players.

The encoded audio and posture representations are then integrated by the Cross-Modal Fusion Module. Although this module is built upon a bi-directional attention framework, it is specifically tailored to reconcile the heterogeneous nature of spatial posture data and temporal audio signals. This stage-level fusion forms a strong semantic foundation for subsequent reasoning.

To model the dependencies between the non-rally and rally stages within the same point, we introduce the Cross-Stage Fusion Module. Instead of a mere attention block, this module performs a dedicated fusion of the stage-wise representations, allowing the model to explicitly combine and reason

over inter-stage information. This cross-stage fusion is the central innovation of our approach, enabling the capture of subtle dynamics across stages.

Finally, to incorporate temporal context across multiple points, we apply a sliding window of length $L = 3$ over consecutive points. Fused representations from each window are concatenated and passed to a linear classifier, which predicts the outcome of the final stage in the window. Through this hierarchical pipeline—from modality-specific encoding and fusion to inter-stage fusion and temporal aggregation—the model learns both fine-grained interactions and broader momentum trends throughout a match.

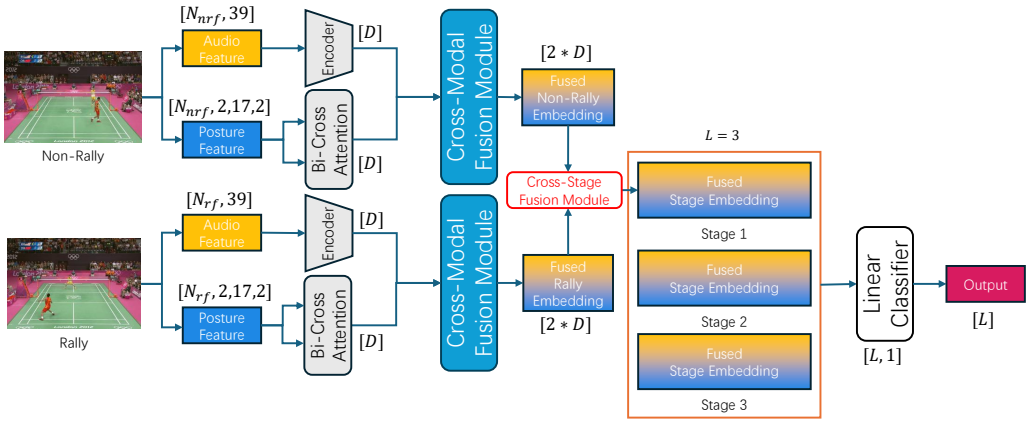

Figure 4.1: Proposed Badminton Winning Prediction Model framework.

4.2 Cross-Modal Fusion Module

The Cross-Modal Fusion Module is designed to integrate two heterogeneous modalities—environmental audio and players' posture—into a unified representation for each stage (rally or non-rally). This fusion process captures both global scene information and fine-grained body dynamics, enabling the model to learn semantic correlations across modalities.

Architecture Overview

As shown in Figure 4.2, the module takes audio and posture features as input. The audio features, extracted from environmental sound and encoded through a Transformer-based encoder, yield the Audio Embedding. Posture features, which represent the 2D coordinates of keypoints for both players, are processed through a bi-directional cross-attention mechanism to generate the Posture Embedding.

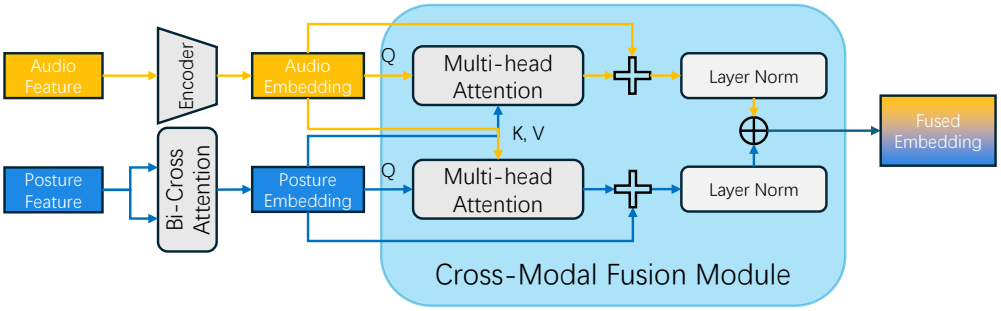

Figure 4.2: Cross-Modal Fusion Module architecture.

These embeddings are then exchanged through a bi-directional attention structure:

- **Audio-guided Posture Attention:** The posture embedding serves as the query (Q), while the audio embedding provides the keys and values (K, V).
- **Posture-guided Audio Attention:** The audio embedding serves as the query (Q), while the posture embedding provides the keys and values (K, V).

Each attention output is followed by a residual connection and a Layer Normalization step. These are represented by the operator $+$, indicating residual addition, and the box labeled "LayerNorm" in the diagram.

Finally, the attention-enhanced posture and audio embeddings are concatenated along the feature dimension (denoted by $\oplus$) to produce the final fused representation.

Multi-Head Attention

The multi-head attention mechanism allows the model to jointly attend to information from different representation subspaces. Given a query $Q \in R^{L\times D}$, key $K \in R^{L\times D}$, and value $V \in R^{L\times D}$, the attention output is computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{4.1}$$

For multi-head attention, this is applied in parallel h times with different learnable projections:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots, head_h)W^O \tag{4.2}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{4.3}$$

where $W_i^Q, W_i^K, W_i^V \in R^{D \times d_k}$ are learnable projection matrices and $W^O \in R^{hd_k \times D}$ is the output projection.

Layer Normalization and Residual Addition

Each attention output is followed by a residual connection and a LayerNorm operation:

$$\hat{X} = LayerNorm(X + Attention(Q, K, V)) \tag{4.4}$$

LayerNorm normalizes the feature across dimensions to stabilize training:

$$LayerNorm(x) = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \cdot \gamma + \beta \tag{4.5}$$

where μ and σ^2 are the mean and variance of input x, and γ, β are learnable affine parameters.

Fusion Strategy

The final step in the module concatenates the enhanced audio and posture representations:

$$F_{fused} = \hat{P} \oplus \hat{A} \tag{4.6}$$

Here, $\hat{P}$ and $\hat{A}$ denote the attention-enhanced posture and audio features, respectively, after residual addition and layer normalization:
$\hat{P} = LayerNorm(Pose + Attention_{pose \leftarrow audio})$
$\hat{A} = LayerNorm(Audio + Attention_{audio \leftarrow pose})$

The operator $\oplus$ represents feature concatenation along the last dimension. As a result, the fused feature $F_{fused} \in R^{L \times 2D}$ integrates modality-aware information from both input streams without further projection layers. This fused embedding serves as the stage-level representation for either the rally or non-rally stage, and is subsequently passed to the Cross-Stage Fusion Module.

4.3 Cross-Stage Fusion Module

The Cross-Stage Fusion Module integrates the fused embeddings from the non-rally and rally stages within each point. Unlike simple concatenation

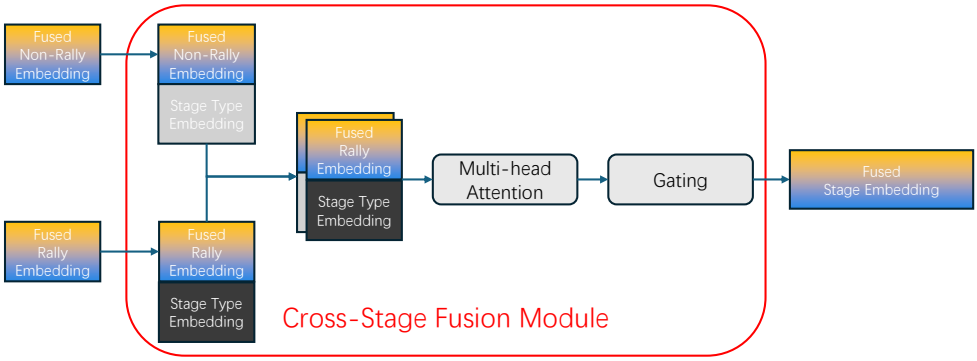

Figure 4.3: Cross-Stage Fusion Module architecture.

Architecture Overview

As shown in Figure 4.3, the module takes as input the fused non-rally and rally embeddings from the previous stage. A learnable stage type embedding (0 for non-rally, 1 for rally) is added to each input to explicitly encode stage identity.

The two enriched embeddings are then passed through a multi-head self-attention layer to model interactions between the stages. The resulting attention-enhanced features are subsequently fused using a gating mechanism.

Gated Fusion Mechanism

To dynamically control the contribution of each stage, the module employs a learnable gate. Given the attention-enhanced non-rally and rally features, denoted as f_{non} and f_{rally}, the gate vector g is computed as:

$$g = \sigma(W_g[f_{non} \oplus f_{rally}]) \tag{4.7}$$

where W_g is a learnable linear projection, $\oplus$ denotes feature concatenation, and σ is the sigmoid activation function.

The final fused stage representation is given by a weighted combination:

$$F_{stage} = g \odot f_{non} + (1 - g) \odot f_{rally} \tag{4.8}$$

where $\odot$ denotes element-wise multiplication. This allows the model to adaptively emphasize either stage based on the context.

Output Representation

The output of this module is the stage-level embedding $F_{stage} \in R^{B\times 2D}$, which encodes both gameplay phases and their semantic relationships. These representations are later aggregated across multiple points and passed to the final classification head for point outcome prediction.

4.4 Baseline Methods

We implemented several baseline models to evaluate the effectiveness of our proposed attention-based fusion and sequence modeling strategy:

4.4.1 Concatenation-based Fusion

To evaluate the effectiveness of the proposed Cross-Stage Fusion Module, we design an ablation experiment by replacing it with a simple concatenation-based fusion strategy. In this baseline, the modality-specific features from the non-rally and rally segments are directly concatenated without any cross-modal or cross-stage interaction. Specifically, given the non-rally features $F_{non} \in R^{B\times L\times D}$ and rally features $F_{rally} \in R^{B\times L\times D}$, the fused representation is obtained as:

$$F_{concat} = Concat(F_{non}, F_{rally}) \in R^{B\times L\times 2D}$$

This fused feature sequence is then passed through a temporal modeling backbone such as a Transformer encoder or LSTM to capture sequential dependencies and make stage-wise predictions. Since no explicit fusion or interaction is performed between the modalities, this setup serves as a strong baseline to assess the contribution of the proposed fusion mechanism.

4.4.2 CNN-based Cross-Stage Fusion

As another ablation setting, we replace the proposed Cross-Stage Fusion Module with a lightweight convolution-based fusion approach. Instead of modeling cross-stage interactions via attention, this method applies a temporal convolution over the concatenated non-rally and rally features to capture local correlations.

Given non-rally features $F_{non} \in R^{B\times L\times D}$ and rally features $F_{rally} \in R^{B\times L\times D}$, we first concatenate them along a new stage dimension to form:

$$F_{stacked} = Stack(F_{non}, F_{rally}) \in R^{B\times L\times 2\times D}$$

We then apply a 1-dimensional convolution across the stage axis:

$$F_{fused} = Conv1D(F_{stacked},\ kernelsize = 2) \in R^{B \times L \times D}$$

This fused feature sequence is subsequently passed to the temporal modeling module (e.g., Transformer encoder or LSTM) for final stage-wise classification. By comparing this simple Convolutional Neural Network (CNN)-based fusion to the full attention-based Cross-Stage Fusion Module, we can evaluate the contribution of Cross-Stage Fusion Module in capturing more informative interactions.

4.5 Classification Head

To investigate the impact of different sequence modeling strategies on stage-wise outcome prediction, we compare three types of classification heads: a simple Linear classifier, an LSTM-based model, and a Transformer-based model. Each classification head takes as input the fused stage-level features over a sliding window and outputs binary predictions for each stage.

4.5.1 Transformer Block

The Transformer-based classification head leverages self-attention to model temporal dependencies across multiple stages. Given an input sequence $\mathbf{X} \in R^{B \times L \times D}$, where L is the number of stages in the sliding window and D is the feature dimension, the Transformer Block captures global contextual information using multi-head self-attention and feed-forward layers. The architecture is shown in Figure 4.4.

$$\mathbf{H} = TransformerBlock(\mathbf{X}) \in R^{B \times L \times D}$$

$$\hat{y} = Linear(\mathbf{H}) \in R^{B \times L \times 1}$$

4.5.2 Long Short-Term Memory

The LSTM-based head uses a unidirectional Long Short-Term Memory [1] network to sequentially encode the temporal dynamics of stage-level features. The LSTM captures short- and long-range dependencies between stages:

$$\mathbf{H} = LSTM(\mathbf{X}) \in R^{B \times L \times D}$$

$$\hat{y} = Linear(\mathbf{H}) \in R^{B \times L \times 1}$$

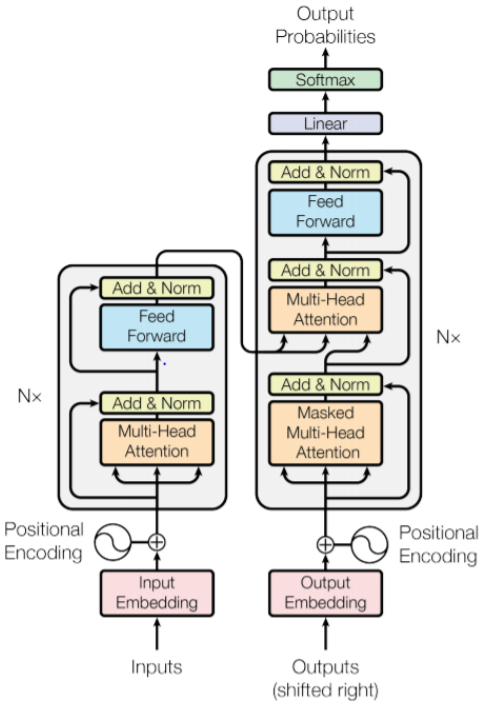

Figure 4.4: Transformer Blocks architecture.[6]

4.5.3 Linear Classifier

As a lightweight classification head, we apply a linear projection over the fused feature vector $\mathbf{X}$:

$$\hat{y} = Linear(\mathbf{X}) = \mathbf{W}\mathbf{X} + \mathbf{b}$$

where $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters of the linear layer.

This linear classification head is used to evaluate the effectiveness of the preceding feature fusion method.

The experimental results demonstrate that, thanks to the strong and effective multi-modal fusion, even a simple linear classifier can achieve the best prediction performance among various classification heads tested.

Chapter 5

Experiment and Results

5.1 Experiment Settings

In this section, we describe the experimental setup for evaluating our proposed model. All experiments are conducted on the BWP dataset introduced in Chapter 4. The dataset is split into training, validation, and test sets in an 8:1:1 ratio. For each match, stage-level samples are randomly partitioned and merged across the splits, ensuring every match appears in all three subsets. This preserves diversity across gender, camera angles, and player styles, thereby enhancing generalization and robustness.

5.1.1 Evaluation Metrics

To quantitatively evaluate the performance of our classification models, we use the following metrics:

- **Accuracy** measures the proportion of correctly predicted samples among all samples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** evaluates the accuracy of positive predictions:

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

- **Recall** (also known as Sensitivity) measures the ability to identify positive samples:

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

- **F1-score** is the harmonic mean of Precision and Recall:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Here, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

These metrics provide a comprehensive evaluation of classification performance from multiple perspectives.

5.1.2 Window Size and Stage-wise Settings

We set the sliding window size $L = 3$ by default. Each window contains 3 stages, and prediction can be performed on either all 3 stages (multi-label) or the last stage only (single-label). Padding is used when insufficient context is available.

Preliminary Analysis and Input Strategy

We first conducted a pilot study to investigate how different input configurations and output strategies affect model performance. Specifically, we used a simple linear classification head and fused MFCC and posture features as input. Each stage input was fixed to 500 frames using a *head-cut* padding strategy—retaining the earlier part of the stage and truncating the tail.

To model short-term temporal dynamics, we adopted a Cross-Stage Fusion Module with a sliding window of size 3. That is, each prediction considers the current stage and the two preceding ones.

We also evaluated two output label strategies:

- **Label 1**: Only the outcome of the latest (third) stage in the window is predicted.
- **Label 3**: Outcomes for all three stages in the sliding window are predicted in parallel.

The performance of these configurations is summarized in Table 5.1. Regardless of the output setting, models trained on head-cut inputs consistently hovered around 50% validation accuracy—close to a random baseline. This suggests that early-stage frames lack strong predictive signals for outcome classification.

As shown, even the best-performing configuration only marginally outperforms random guessing. These results confirm that crucial decision-making cues may lie later in the stage.

Table 5.1: Results on head-cut 500-frame inputs using Cross-Stage Fusion Module and Linear Classifier.

Output	Accuracy	Precision	Recall	F1-score
3	0.5525	0.5321	0.5512	0.6125
1	0.5323	0.6247	0.6791	0.5972

Revised Input Strategy

To address this limitation, we revised the input configuration in two ways: (1) we increased the stage input length from 500 to 600 frames, and (2) we switched to a *tail-cut* strategy, preserving the latter part of the stage while discarding the early frames. This adjustment led to substantially better training convergence and improved model accuracy. It also supports the intuition that outcome-relevant behaviors—such as smashes, drop shots, or unforced errors—tend to occur near the end of a rally. This revised setting serves as the default configuration in the remaining experiments, unless stated otherwise.

5.2 Main Results

We now present the main experimental results using the revised input strategy (tail-cut, 600 frames) under two output configurations: a multi-label setting that predicts all three stages within each window (3-label output), and a single-label setting that predicts only the latest stage (1-label output).

Tables 5.2 and 5.3 report the performance across different classifier heads, input types, and fusion strategies. All models use a sliding window size of $L = 3$ and the default bimodal input (MFCC + posture) unless otherwise noted.

In the 3-label setting (Table 5.2), the best performance is achieved using a simple linear classifier with the Cross-Stage Fusion Module, reaching an accuracy of 87.5%. Linear models slightly outperform Transformer and LSTM heads in this task, possibly due to the limited data size or the relative simplicity of multi-label prediction.

In the 1-label setting (Table 5.3), a similar trend holds, where the linear classifier again achieves the highest accuracy (88.75%), outperforming both LSTM and Transformer-based alternatives. The linear model also attains the best F1-score of 0.9011, suggesting more consistent predictions on the most recent stage.

Across both tasks, several consistent patterns emerge:

Table 5.2: Main experimental results (3-label classification). Best accuracy is bolded.

Classification Head	Input	Fusion Method	Slide Winodw Size	Accuracy	Precision	Recall	F1-score
Linear Classifier	MFCC, Posture	Cross-Stage Fusion Module	3	0.875	0.8364	0.9328	0.881
Transformer	MFCC, Posture	Cross-Stage Fusion Module	3	0.8208	0.8654	0.7563	0.8072
LSTM	MFCC, Posture	Cross-Stage Fusion Module	3	0.7625	0.7385	0.8067	0.7711
Transformer	Posture Only	Cross-Stage Fusion Module	3	0.7458	0.7544	0.7227	0.7382
Transformer	MFCC, Posture	Concat	3	0.7125	0.7404	0.6471	0.6906
Transformer	MFCC, Posture	Cross-Stage CNN	3	0.6875	0.6897	0.6723	0.6809
Linear Classifier	MFCC, Posture (Rally Stage Only)	/	3	0.672	0.6324	0.625	0.6205
Transformer	MFCC, Posture (Rally Stage Only)	/	3	0.6	0.6117	0.5294	0.5676
Linear Classifier	MFCC, Posture (Padding 500 head cut)	Cross-Stage Fusion Module	3	0.5525	0.5321	0.5512	0.6125

Table 5.3: Main experimental results (1-label classification). Best accuracy is bolded.

Classification Head	Input	Fusion Method	Slide Winodw Size	Accuracy	Precision	Recall	F1-score
Linear Classifier	MFCC, Posture	Cross-Stage Fusion Module	3	0.8875	0.8723	0.9318	0.9011
LSTM	MFCC, Posture	Cross-Stage Fusion Module	3	0.85	0.9211	0.7955	0.8537
Transformer	MFCC, Posture	Cross-Stage Fusion Module	3	0.775	0.8611	0.7045	0.775
Transformer	MFCC, Posture	Concat	3	0.725	0.8056	0.6591	0.725
Transformer	Posture Only	Cross-Stage Fusion Module	3	0.7	0.8846	0.5227	0.6571
Linear Classifier	MFCC, Posture (Rally Stage Only)	/	3	0.6385	0.6948	0.5861	0.6582
Transformer	MFCC, Posture (Rally Stage Only)	/	3	0.5875	0.6667	0.5	0.5714
Transformer	MFCC, Posture	Cross-Stage CNN	3	0.5625	0.6286	0.5	0.557
Linear Classifier	MFCC, Posture (Padding 500 head cut)	Cross-Stage Fusion Module	3	0.5323	0.6247	0.6791	0.5972

- **Multi-Modal inputs (MFCC + posture)** consistently outperform unimodal ones (e.g., posture only), confirming the complementary nature of audio and visual cues.
- **Cross-Stage Fusion** provides clear improvements over simpler alternatives such as feature concatenation or CNN-based fusion, demonstrating its effectiveness in leveraging temporal dependencies across stages.
- Models trained on **rally-stage-only inputs** suffer substantial performance drops, indicating the necessity of modeling both rally and non-rally stages for accurate prediction.
- The **head-cut input strategy** (removing early frames) consistently results in the worst performance, validating our use of tail-cut segments as a more informative temporal window.

These observations reinforce the importance of full-stage temporal modeling and multi-modal integration for both stage-wise and latest-stage behavior recognition. Further analyses on generalization and ablations are presented in Section 5.3.

Table 5.4: Comparison of Fusion Methods (Classification Head: Transformer, $L = 3$, 3-label classification)

Fusion Method	Accuracy	Precision	Recall	F1-score
Cross-Stage Fusion Module (Ours)	0.8208	0.8654	0.7563	0.8072
Concatenation	0.7125	0.7404	0.6471	0.6906
CNN Fusion	0.6875	0.6897	0.6723	0.6809

Table 5.5: Comparison of Fusion Methods (Classification Head: Transformer, $L = 3$, 1-label classification)

Fusion Method	Accuracy	Precision	Recall	F1-score
Cross-Stage Fusion Module (Ours)	0.775	0.8611	0.7045	0.775
Concatenation	0.725	0.8056	0.6591	0.725
CNN Fusion	0.5625	0.6286	0.5	0.557

5.3 Ablation Studies

5.3.1 Fusion Method Comparison

Table 5.4 and Table 5.5 compares the performance of different fusion strategies with the same backbone classifier (Transformer). Our proposed attention-based fusion outperforms early fusion and CNN-based fusion in all metrics.

5.3.2 Classification Head Comparison

Table 5.6 and Table 5.7shows results of using different sequence models while keeping the fusion method fixed (Cross-Stage Fusion Module). The linear classifier achieves surprisingly strong overall performance, outperforming more complex sequence models in both accuracy and F1-score. However, Transformer and LSTM architectures still offer competitive recall and may provide better temporal interpretability.

Table 5.6: Comparison of Sequence Models (Fusion: Cross-Stage Fusion Module, $L = 3$, 3-label classification)

Classification Head	Accuracy	Precision	Recall	F1-score
Linear Classifier	0.8750	0.8364	0.9328	0.8810
Transformer	0.8208	0.8654	0.7563	0.8072
LSTM	0.7625	0.7385	0.8067	0.7709

Table 5.7: Comparison of Sequence Models (Fusion: Cross-Stage Fusion Module, $L = 3$, 1-label classification)

Classification Head	Accuracy	Precision	Recall	F1-score
Linear Classifier	0.8875	0.8723	0.9318	0.9011
Transformer	0.85	0.9211	0.7955	0.8537
LSTM	0.775	0.8611	0.7045	0.775

Table 5.8: Modality Comparison (Fusion: Cross-Stage Fusion Module, Classification Head: Transformer, $L = 3$, 3-label classification)

Input Modality	Accuracy	Precision	Recall	F1-score
Pose + Audio	0.8208	0.8654	0.7563	0.8072
Pose Only	0.7458	0.7544	0.7227	0.7382

5.3.3 Modality Comparison

We conducted an ablation experiment to verify the effectiveness of multimodal learning. Specifically, we compare the full model using both audio and posture features with a reduced version using only posture. Tables 5.8 and 5.9 show that the inclusion of audio features improves overall performance, especially in precision and F1-score.

5.3.4 Stage Inputs Comparison

To evaluate the importance of non-rally stages in badminton matches, we conducted experiments under two settings: using only rally stages (**R**) and using both rally and non-rally stages (**R+NR**). We tested two classification heads—a linear classifier and a transformer-based classifier—to ensure the generality of the findings. Notably, stage fusion is applicable only in the **R+NR** setting, as it requires the presence of both rally and non-rally features for fusion. Therefore, in the **R** setting, the model directly classifies the rally feature of each stage without fusion.

As shown in Table 5.10 and Table 5.11, using both rally and non-rally

Table 5.9: Modality Comparison (Fusion: Cross-Stage Fusion Module, Classification Head: Transformer, $L = 3$, 1-label classification)

Input Modality	Accuracy	Precision	Recall	F1-score
Pose + Audio	0.775	0.8611	0.7045	0.775
Pose Only	0.7	0.8846	0.5227	0.6571

Table 5.10: Stage Inputs Comparison (Stage Input: R vs. R+NR, Fusion: Cross-Stage Fusion Module, Classification Head: Transformer, $L = 3$, 3-label classification)

Classification Head	Input Stage	Accuracy	Precision	Recall	F1-score
Linear Classifier	R+NR	0.875	0.8364	0.9328	0.881
Transformer	R+NR	0.8208	0.8654	0.7563	0.8072
Linear Classifier	R	0.672	0.6324	0.625	0.6205
Transformer	R	0.6	0.6117	0.5294	0.5676

Table 5.11: Stage Inputs Comparison (Stage Input: R vs. R+NR, Fusion: Cross-Stage Fusion Module, Classification Head: Transformer, $L = 3$, 1-label classification)

Classification Head	Input Stage	Accuracy	Precision	Recall	F1-score
Linear Classifier	R+NR	0.8875	0.8723	0.9318	0.9011
Transformer	R+NR	0.775	0.8611	0.7045	0.775
Linear Classifier	R	0.6385	0.6948	0.5861	0.6582
Transformer	R	0.5875	0.6667	0.5	0.5714

stages (**R+NR**) consistently outperforms using only rally stages (**R**) across both classification heads. This result highlights the significance of the non-rally stage in badminton match prediction. Due to the nature of the sport, the non-rally period often contains valuable information—such as player movement patterns, emotional cues, and audio context—which are not present during fast-paced rallies. These additional cues contribute to a more comprehensive representation of each game stage, thereby enhancing the model's ability to predict match outcomes.

These results emphasize the necessity of utilizing full-stage information when modeling competitive dynamics in badminton. Ignoring non-rally segments leads to a notable loss in contextual understanding, underlining the unique temporal characteristics of the sport.

5.4 Discussion

The main key findings of this study are:

(1) Our multi-modal fusion approach effectively captures predictive information in badminton matches.

(2) Even a simple linear classifier achieves excellent performance, demonstrating the strength of the fused features.

(3) Audio cues and non-rally segments significantly contribute to improving prediction accuracy.

Our experiments reveal that badminton matches exhibit notable predictability at the point level. The relatively high accuracy across settings indicates that historical stage-level features encapsulate meaningful patterns that contribute to the upcoming point outcome.

In particular, our proposed Cross-Stage Fusion Module shows strong fitting ability, as evidenced by the fact that even a simple linear classification head can achieve comparable or superior performance to more complex sequence models. This suggests that the fused features themselves are highly discriminative.

The inclusion of audio features, such as environmental cheering, proved beneficial. Performance dropped when the audio modality was removed, indicating that auditory cues play a non-negligible role in competitive matches. Interestingly, the presence of crowd noise may influence even professional players, subtly affecting their performance.

Another observation is that non-rally segments, which occupy a much longer duration than rally segments in badminton, still hold valuable predictive information. Our findings show that incorporating both rally and non-rally stages leads to better prediction results, justifying the need to model non-rally contexts instead of discarding them as irrelevant.

Chapter 6

Conclusion

6.1 Summary of Contributions

This study proposed a novel multi-modal approach for point-by-point prediction of badminton match outcomes by leveraging both human posture and environmental audio signals. We introduced a Cross-Stage Fusion Module that effectively integrates rally and non-rally stage features. Extensive experiments demonstrate the following key contributions:

- We confirmed the predictability of badminton match outcomes based on historical stage-level information, achieving high classification performance. This implies practical potential in aiding match analysis for players, coaches, and audiences.
- We proposed the Cross-Stage Fusion Module, which showed strong feature modeling capacity. Even with a simple linear classifier, the fused representations yielded highly competitive results, surpassing more complex architectures in some settings.
- The inclusion of audio features, particularly audience cheers and environmental sound, significantly improved prediction performance. This highlights that even professional players are influenced by auditory context, which provides additional cues for win/loss estimation.
- Our analysis revealed that non-rally stages, which occupy a larger portion of match time compared to rally stages, contain valuable information for prediction. This justifies the importance of modeling both stages instead of focusing solely on rallies.

6.2 Limitations

Despite promising results, several limitations remain in this work. First, the dataset size was limited and confined to a specific set of matches, which may impact the generalizability of the model. Second, we relied on pre-extracted posture features and MFCC-based audio descriptors; more sophisticated or learned representations could potentially yield better results. Additionally, the classification task was treated as binary (win/loss), ignoring the broader match dynamics such as confidence, fatigue, or momentum shifts.

6.3 Future Work

In future research, we plan to expand the dataset to cover more diverse matches, including different player levels and tournament types. We also aim to explore end-to-end feature extraction and modeling pipelines, potentially integrating vision transformers or audio transformers for better modality encoding. Further, introducing continuous outcome prediction (e.g., win probability curves) or interpretability modules could enhance the usability of the model in real-time coaching or broadcasting scenarios.

In addition, we plan to extend this predictive framework to other racket sports, enabling outcome prediction and match analysis using only single-camera broadcast footage. Ultimately, we aim to make multi-modal sports analysis a central focus, improving the accuracy and applicability of AI-driven insights across various competitive settings.

Bibliography

[1] Hochreiter, Sepp and Jürgen Schmidhuber. Long short-term memory, *Neural Computation*, vol.9, no.8, pp.1735–1780 (1997)

[2] Brahms, Bernd-Volker. *Badminton handbook*, Meyer Meyer Sport, (2014)

[3] Kovalchik, Stephanie Ann. Searching for the GOAT of tennis win prediction, *Journal of Quantitative Analysis in Sports*, vol.12, no.3, pp.127–138 (2016)

[4] Mack, G. S. B. *The Game of Badminton-The Rules and Tactics of a Singles Match*, Read Books Ltd, (2016)

[5] Tan, Daniel Yong Wen, Huong Yong Ting, and Simon Boung Yew Lau. "A review on badminton motion analysis." *2016 International Conference on Robotics, Automation and Sciences (ICORAS). IEEE*, (2016)

[6] Vaswani, Ashish, et al. Attention is all you need, *Advances in Neural Information Processing Systems*, vol.30 (2017)

[7] Cao, Zhe, et al. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.43, no.1, pp.172–186 (2019)

[8] Zhu, Luyang, et al. Reconstructing NBA players, *European Conference on Computer Vision* (2020)

[9] Badiola-Bengoa, Aritz and Amaia Mendez-Zorrilla. A systematic review of the application of camera-based human pose estimation in the field of sport and physical exercise, *Sensors*, vol.21, no.18, pp.5996 (2021)

[10] Gao, Zijian and Amanda Kowalczyk. Random forest model identifies serve strength as a key predictor of tennis match outcome, *Journal of Sports Analytics*, vol.7, no.4, pp.255–262 (2021)

[11] Brady, C., Tuyls, K., and Omidshafiei, S. *AI for Sports*, CRC Press (2021)

[12] Sharma, Manoj, et al. Badminton match outcome prediction model using Naïve Bayes and Feature Weighting technique, *Journal of Ambient Intelligence and Humanized Computing*, vol.12, no.8, pp.8441–8455 (2021)

[13] Nahavandi, Darius, et al. Application of artificial intelligence in wearable devices: Opportunities and challenges, *Computer Methods and Programs in Biomedicine*, vol.213, pp.106541 (2022)

[14] Kitamura, Takumi, et al. Refining OpenPose with a new sports dataset for robust 2D pose estimation, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2022)

[15] Vatolik, I., et al. Development of a multi-modal sensor network to detect and monitor knee joint condition, *Measurement: Sensors*, vol.24, pp.100483 (2022)

[16] Komitova, Rumena, et al. Time series data mining for sport data: A review, *Journal of the International Association of Computer Science in Sport*, vol.21, no.2 (2022)

[17] Jiang, Tao, et al. Rtmpose: Real-time multi-person pose estimation based on mmpose, *arXiv preprint arXiv:2303.07399* (2023)

[18] Bodemer, Oliver. Enhancing individual sports training through artificial intelligence: A comprehensive review, *Authorea Preprints* (2023)

[19] Nilesh, Nitin, et al. Towards real-time analysis of broadcast badminton videos, *arXiv preprint arXiv:2308.12199* (2023)

[20] Jiao, Tianzhe, et al. A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications, *Computers, Materials Continua*, vol.80, no.1 (2024)

[21] Goka, Ryota, et al. Multimodal shot prediction based on spatial-temporal interaction between players in soccer videos, *Applied Sciences*, vol.14, no.11, pp.4847 (2024)

[22] Yuan, Hanguang, et al. Prediction model and technical and tactical decision analysis of women's badminton singles based on machine learning, *PloS One*, vol.19, no.11 (2024)

[23] Yu, Han, Luyun Jia, and Yumo Miao. CLSTM: A Dynamic Model for Predicting Winning Percentages in Tennis Matches, *Proceedings of the 2024 5th International Conference on Computing, Networks and Internet of Things* (2024)

[24] Takamido, Ryota, Jun Ota, and Hiroki Nakamoto. PassAI: explainable artificial intelligence algorithm for soccer pass analysis using multimodal information resources, *arXiv preprint arXiv:2503.08945* (2025)

[25] Xv, Guang and Xingchen Wu. Temporal event localization in sports videos via self-supervised proposal generation and cross-modal fusion, *Intelligent Systems with Applications*, pp.200539 (2025)

[26] Li, Haibin, Aodi Guo, and Yaqian Li. CCMA: CapsNet for audio–video sentiment analysis using cross-modal attention, *The Visual Computer*, vol.41, no.3, pp.1609–1620 (2025)

[27] Sheng, Yi, et al. Predicting badminton outcomes through machine learning and technical action frequencies, *Scientific Reports*, vol.15, no.1, pp.10575 (2025)

[28] Majeed, Fahad, et al. Real-time analysis of soccer ball–player interactions using graph convolutional networks for enhanced game insights, *Scientific Reports*, vol.15, no.1, pp.21859 (2025)

[29] Jiang, Kailai. Optimization study of badminton sports training system based on MoileNet OpenPose lightweight human posture estimation model, *Entertainment Computing*, pp.100975 (2025)

[30] Jo, Eunhye. Development of sequential winning-percentage prediction model for badminton competitions: applying the expert system sequential probability ratio test, *BMC Sports Science, Medicine and Rehabilitation*, vol.17, no.1, pp.48 (2025)

[31] Zhao, Zhonghan, et al. "A survey of deep learning in sports applications: Perception, comprehension, and decision." *IEEE Transactions on Visualization and Computer Graphics*, (2025)

[32] BWF Corporate. Available at: https://corporate.bwfbadminton.com/ (Accessed: 03 August 2025)

[33] BWF TV - youtube. Available at: https://www.youtube.com/user/bwf (Accessed: 13 April 2025)

[34] ELAN — The Language Archive. Available at: https://archive.mpi.nl/tla/elan (Accessed: 03 August 2025)