Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20260503008
Volume 5, Issue3 (September-December 2026), PP: 46-61.
www.indjcst.com
ISSN No: 2583-5300

Video-Based Handball Action Classification Using ResNet18-BiGRU with Temporal Attention

Prasad Raj Bingi[1], Penugonda Ravi Kumar[2]
[1]*PG Scholar, Department of CSE, RGUKT – RK VALLEY, Vempalli, Andhra Pradesh, India.*
[2]*Assistant Professor, Department of CSE, RGUKT – RK VALLEY, Vempalli, Andhra Pradesh, India.*

To Cite this Article: *Prasad Raj Bingi[1], Penugonda Ravi Kumar[2], "Video-Based Handball Action Classification Using ResNet18-BiGRU with Temporal Attention", Indian Journal of Computer Science and Technology, Volume 05, Issue 03 (September-December 2026), PP: 46-61.*

Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution License; Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: *This study investigates video-based handball action classification using the UNIRI Handball Dataset Version 1. The study focuses on seven handball actions: crossing, defence, dribbling, jump-shot, passing, running, and shot. The dataset was organized by associating scene videos, labelled action clips, and available player-detection information using scene identifiers. To reduce the possibility of information leakage, the data were divided at the scene level so that action clips originating from the same scene remained within a single subset. The study evaluates a sequence of feature representations and learning approaches, beginning with MediaPipe Pose-based pose and motion descriptors and conventional machine-learning classifiers, followed by temporal deep-learning models using recurrent, temporal convolutional, skeleton-based, and Transformer-based architectures. Based on these experiments, a video-based spatial-temporal architecture was developed using a pretrained ResNet18 network for frame-level spatial feature extraction, followed by feature projection, a bidirectional gated recurrent unit (BiGRU) for temporal modelling, and a temporal attention mechanism for aggregating informative features across the video sequence. A fixed number of RGB frames was sampled from each action clip to construct the temporal representation. The ResNet18 backbone was used as a frozen feature extractor, while the temporal and classification layers were trained using class-weighted loss with label smoothing. Model selection was performed using validation data, while the test subset was kept locked until final evaluation. The proposed framework provides a structured and leakage-aware approach for learning spatial and temporal representations from handball video and addresses the challenge of distinguishing visually and temporally similar handball actions.*

Key Word: *Handball Action Recognition; Video Classification; ResNet18; BiGRU; Temporal Attention; Deep Learning.*

I. INTRODUCTION

Video-based human action recognition has become an important area of computer vision because of its applications in activity understanding, sports analysis, surveillance, and human–computer interaction[4]. In sports, automatic recognition of player actions can assist in analysing gameplay, identifying important events, and supporting performance assessment. Compared with controlled action-recognition environments, sports videos contain substantial variation in player movement, camera viewpoint, background, occlusion, and interaction between multiple players. These characteristics make reliable recognition of short and visually similar actions a challenging problem.

Handball is particularly suitable for investigating these challenges because gameplay contains rapid movements and frequent interactions between players. Actions such as running, passing, dribbling, defending, crossing, and shooting may occur within a short sequence of frames and can have similar visual characteristics. The same action may also appear differently depending on the player's position, surrounding players, camera viewpoint, and movement of the scene. Consequently, a useful recognition system must consider not only the appearance of individual frames but also the temporal information contained across a sequence of frames.

The availability of the UNIRI Handball Dataset Version 1 (UNIRI-HBD) provides a specific benchmark for studying these problems. The dataset contains handball video clips representing seven action categories: crossing, defence, dribbling, jump-shot, passing, running, and shot[2]. Previous work associated with this dataset has investigated the organization of handball video data and the identification of active players in handball scenes[1]. These studies provide an important foundation for subsequent work on automated action understanding, while leaving scope for investigating learned spatiotemporal representations for direct video-based action classification.

Traditional action-recognition methods commonly rely on manually designed appearance, pose, or motion descriptors followed by conventional classifiers. Such representations can provide useful information about body configuration and movement, but they may not fully represent the visual context and temporal evolution of an action. Deep learning methods provide an alternative by learning feature representations directly from video data. Convolutional neural networks can extract spatial information from individual frames, whereas recurrent and temporal models can represent how visual information changes throughout an action sequence. Combining these complementary capabilities is therefore relevant to handball action recognition.

In this study, the research is conducted using a structured organization of the available scene videos, labelled action clips, and player-detection information. Because multiple action clips can originate from the same scene, the relationship between these data sources is established using scene identifiers rather than file ordering. The dataset is subsequently divided

at the scene level to avoid placing related clips from the same scene into different subsets. This procedure is important for obtaining an evaluation setting in which the validation and test samples are separated from the scenes used for model development.

The experimental investigation is carried out in several stages. First, pose and motion information is extracted using MediaPipe Pose, providing body-landmark-based representations for conventional machine-learning experiments. These representations are evaluated using different classical classifiers to establish a baseline. Temporal deep-learning models are then investigated to examine different ways of representing action sequences, including recurrent, temporal convolutional, skeleton-based, and Transformer-based models[11]. These experiments provide a basis for selecting a suitable representation and temporal modelling strategy for the final video-based classifier.

The proposed classification framework uses a pretrained ResNet18 network to extract spatial features from sampled RGB frames. The ResNet18 backbone is used as a frozen feature extractor, producing a 512-dimensional representation for each frame. The resulting sequence of frame-level features is projected into a lower-dimensional representation and processed using a bidirectional gated recurrent unit (BiGRU) to model temporal dependencies in both directions. A temporal attention mechanism is subsequently applied to assign greater importance to informative portions of the sequence. The resulting temporal representation is passed to a fully connected classifier for recognition of the seven handball actions.

The main objective of this research is therefore to develop and systematically evaluate a video-based handball action-classification framework that combines pretrained spatial feature extraction with bidirectional temporal modelling and temporal attention. The study also compares the proposed representation with alternative pose-based and temporal modelling strategies under the same structured dataset setting. In this way, the research examines how spatial visual information and temporal information can be combined for action recognition while maintaining scene-level separation between the data subsets.

The main contributions of this study are:

1. **A structured data preparation procedure** is established for associating scene videos, labelled action clips, and available player-detection information using scene identifiers.
2. **A scene-aware data-splitting strategy** is used to reduce the possibility of scene-level information leakage between training, validation, and test subsets.
3. **Pose- and motion-based conventional machine-learning experiments** are conducted using body-landmark information obtained through MediaPipe Pose.
4. **Multiple temporal modelling strategies** are investigated, including recurrent, temporal convolutional, skeleton-based, and Transformer-based models.
5. **A ResNet18–BiGRU architecture with temporal attention** is developed for learning spatial and temporal representations directly from sampled RGB video frames.
6. **A systematic evaluation procedure** is used to compare the investigated models using multiple classification measures while considering the class imbalance present in the handball action data.

II. RELATED WORK

Research on automated handball action recognition has developed from activity and player identification toward learning-based recognition of temporal action patterns. The availability of the UNIRI Handball Dataset (UNIRI-HBD) provided a common dataset for investigating these problems. The original UNIRI-HBD release contains 751 labelled videos representing seven handball actions: passing, shot, jump-shot, dribbling, running, crossing, and defence[2]. The videos were collected from handball training and match-related activities under real sporting conditions.

2.1 UNIRI-HBD-Based Handball Research

Pobar and Ivašić-Kos investigated active-player detection in handball scenes using activity-related measures[1]. Their work considered scenes containing multiple players, where generally one player performs the action of interest. The study used 751 handball videos covering the action categories represented in the original dataset and investigated methods for identifying the active player from the surrounding players. This work is important for the UNIRI-HBD research line because it addresses the problem of locating the player associated with the action before higher-level action understanding is performed.

The UNIRI-HBD dataset was subsequently released as a labelled dataset for supervised machine learning. The dataset contains 751 videos corresponding to seven selected handball actions and was designed to support research in automated recognition of handball activities. The original dataset therefore provides a relatively focused setting in which different visual, pose, motion, and temporal representations can be investigated under the same seven-class action definition. More recently, a study using the original 751-clip UNIRI-HBD collection investigated player trajectories and kinematic information for action classification[3]. Rather than relying on an end-to-end visual representation, that work constructed motion-related descriptors from player trajectories and evaluated lightweight machine-learning classifiers. The study also emphasized the class imbalance and relatively limited size of the dataset as considerations when selecting modelling strategies.

These studies demonstrate that the original UNIRI-HBD dataset has been used for several related tasks, including active-player identification and motion-based action classification[5]. However, these approaches primarily emphasize player activity or engineered motion information. This leaves scope for investigating a learned video representation that directly captures spatial appearance from RGB frames together with temporal dependencies across an action sequence.

2.2 Pose and Motion Representations

Pose-based action recognition represents human activity through body landmarks or joint configurations rather than relying exclusively on raw image appearance. Such representations can describe changes in body posture and movement over time and can therefore provide useful information for distinguishing actions with different motion patterns. In handball, pose

and motion information can be particularly relevant because several actions are characterized by changes in the position of the arms, legs, and upper body.

The original UNIRI-HBD research also demonstrates the importance of identifying the player associated with the action in scenes containing multiple players. The active-player detection study reported that multiple players can appear within a video while only one or a small number of players generally perform the action of interest. This motivates the use of body-landmark information as one of the representations investigated in the present study.

In our work, pose information is obtained using MediaPipe Pose**,** and landmark-derived motion descriptors are used for conventional machine-learning experiments. These experiments provide a reference point for examining how far handcrafted pose and motion information can represent the seven handball actions before moving to learned visual and temporal representations.

2.3 Deep Learning and Temporal Modelling

Deep learning has enabled action-recognition systems to learn representations directly from visual data rather than relying entirely on manually designed descriptors[4]. Convolutional neural networks are particularly useful for extracting spatial information from individual video frames, while recurrent networks can model the temporal evolution of features across a sequence. More recent temporal architectures, including temporal convolutional networks and Transformer-based models, provide alternative mechanisms for representing relationships between frames.

For handball action recognition, temporal modelling is important because an action cannot always be identified reliably from a single frame. For example, the distinction between running, passing, dribbling, and shooting may depend on how the player's posture and movement develop over several consecutive frames. A model therefore needs to preserve relevant temporal information while reducing the influence of frames that contribute little to the recognition task.

The present study investigates several temporal modelling strategies before adopting the final ResNet18–BiGRU architecture with temporal attention. The use of a pretrained ResNet18 provides frame-level spatial representations, while the BiGRU models temporal dependencies in both forward and backward directions. The attention mechanism provides an additional temporal weighting stage so that the learned sequence representation can emphasize informative portions of an action.

2.4 Research Gap

The reviewed UNIRI-HBD Version 1 studies provide useful foundations for handball action understanding, particularly in dataset development, active-player identification, and engineered motion-based classification. However, there is still a need to examine a unified video-based representation that combines learned spatial features with temporal sequence modelling while retaining the original seven-class action definition.

The present study addresses this gap by investigating multiple representation and modelling strategies under a common experimental framework and by developing a ResNet18–BiGRU model with temporal attention for RGB video-based classification. In addition, the study explicitly organizes the scene, action, and associated information using scene identifiers and performs scene-level separation of the data. This provides a controlled basis for comparing pose/motion representations with learned spatial-temporal representations.

III. MATERIALS AND METHODS

3.1 Dataset

The experiments were conducted using the UNIRI Handball Dataset Version 1 (UNIRI-HBD)[2], which provides handball video data for action-recognition research. The dataset used in this study contains three related sources of information: scene videos, labelled action clips, and player-detection files. Scene videos represent the broader handball scenes, while action clips correspond to individual labelled activities extracted from those scenes. Player-detection information is provided separately in CSV format.

During the initial dataset inspection, 754 scene videos in MP4 format, 809 action clips in AVI format, and 751 player-detection CSV files were identified. The action clips are organized into seven target categories: crossing, defence, dribbling, jump-shot, passing, running, and shot. The numbers of files in the three sources are not identical; therefore, they were not treated as three independent datasets or matched according to their file positions. Instead, their relationships were maintained through the corresponding scene identifiers.

The original directory structure was retained during the initial inspection, with the data organized according to the seven action categories. This organization was used as the starting point for subsequent data association, preprocessing, feature extraction, model development, and evaluation.

Data component	File format	Number of files	Purpose
Scene videos	MP4	754	Broader game-scene information
Action clips	AVI	809	Labelled action samples for classification
Player-detection information	CSV	751	Associated player-detection information
Target action classes	—	7	Crossing, Defence, Dribbling, Jump-shot, Passing, Running, Shot

Table 1. Dataset components used in the study

3.2 Data Organization and Scene-Based Association

The three data sources were organized by establishing the relationship between each action clip, its corresponding scene video, and the available player-detection information. Since the numbers of files in these sources are different, file position or directory order was not used to establish correspondence. Instead, a common **scene identifier (scene_id)** was used as the primary reference for association.

The scene identifier was obtained from the filenames of the corresponding files. For an action clip, the identifier represents the source scene, while the action number was retained separately to distinguish multiple action clips originating from the same scene. Thus, more than one labelled action could be associated with the same parent scene without treating those clips as unrelated samples.

The association procedure can be represented as:

Action Clip → Scene ID → Corresponding Scene Video → Associated Player-Detection File

This mapping was particularly important because the dataset contains multiple action clips originating from individual scenes. Matching files according to their position in their respective directories could therefore associate an action with an incorrect scene or player-detection record. The implemented procedure explicitly used scene_id rather than positional or index-based matching.

The verified dataset relationship contained 809 action records associated with 583 unique scene identifiers. The dataset organization identified 362 single-action scenes, 220 multi-action scenes, and 169 scenes without associated action clips among the available scene videos. The action-to-scene and action-to-player associations were also checked during dataset verification.

Dataset relationship	Verified information
Action video clips	809
Unique parent scenes associated with actions	583
Single-action scenes	362
Multi-action scenes	220
Scenes without associated actions	169
Action → Scene association	808/809
Action → Player-detection association	808/809

Table 2. Verified relationship among dataset components

The identifier-based organization was retained for the subsequent data preparation and splitting procedures. This ensured that the relationship between an action sample and its originating scene was preserved throughout the experiments.

3.3 Action Classes

The action classification task considered seven handball action categories: crossing, defence, dribbling, jump-shot, passing, running, and shot. These labels were retained consistently throughout data preparation, model training, validation, and evaluation.

The classes represent different types of player activity observed in handball video sequences. Crossing refers to coordinated player movement in which players cross or exchange their running paths. Defence represents defensive movement performed to restrict or respond to an attacking player. Dribbling denotes movement while controlling the ball through repeated ball handling. Jump-shot represents a shooting action performed while jumping. Passing refers to transferring the ball from one player to another. Running represents running movement that is not assigned to a more specific action category. Shot represents an attempt to shoot the ball toward the goal without the additional jump-shot distinction.

The seven categories were treated as separate supervised classification targets. Their unequal representation was considered during the experimental design, particularly when selecting evaluation measures and applying class-weighted training. The final action-only experimental corpus contained 809 action clips, with the clips distributed across training, validation, and locked testing subsets according to their associated scene identifiers.

Action class	Training	Validation	Test	Total
Crossing	186	24	40	250
Defence	15	7	13	35
Dribbling	17	3	6	26
Jump-shot	164	43	41	248
Passing	94	23	24	141
Running	10	2	1	13
Shot	68	13	15	96

Action class	Training	Validation	Test	Total
Total	**554**	**115**	**140**	**809**

Table 3. Action-class distribution in the final action-only split

The distribution shows a considerable difference in the number of samples available for individual classes. Crossing and jump-shot constitute the largest groups, whereas running, dribbling, and defence have substantially fewer samples. This imbalance is important when interpreting classification performance because accuracy alone can be dominated by the more frequently represented categories. The later experiments therefore use macro-level and balanced measures in addition to overall accuracy.

3.4 Scene-Level Data Splitting and Leakage Prevention

A scene-level splitting strategy was used to reduce information leakage because multiple labelled action clips can originate from the same scene. The scene identifier was therefore used for splitting, and all clips from the same scene were assigned to a single subset.

The final action-only corpus contained 809 action clips from 583 unique scenes, divided into 554 clips from 398 scenes for training, 115 clips from 85 scenes for validation, and 140 clips from 100 scenes for testing.

Subset	Action Clips	Unique Scenes
Training	554	398
Validation	115	85
Testing	140	100
Total	809	583

Table 4. Scene-level distribution of the action-only dataset

Verification confirmed zero scene overlap among the training, validation, and test subsets. The test subset was kept locked and was not used for training, model selection, hyperparameter tuning, or early stopping.

3.5 Player Detection Information and Pose Extraction

The dataset contains player-detection information in the form of CSV files associated with the corresponding scene identifiers. These files were treated as an additional source of player-related information during dataset organization and preparation. The association was performed using the scene identifier rather than file position, as described in Section 3.2.
For pose-based experiments, MediaPipe Pose was used to obtain human body landmark information from the video frames[13]. The pose representation follows the 33-landmark body structure provided by MediaPipe Pose[13]. The extracted landmarks were used to describe the spatial position of body joints over time and to derive motion-related descriptors for the conventional machine-learning experiments.

The pose extraction stage was separate from the supplied player-detection information. The available player-detection CSV files were inspected and associated with their corresponding scenes, but the experimental notebook does **not** contain a separately trained YOLO, SSD, Faster R-CNN, or similar player-detection network. Similarly, no SORT, DeepSORT, or ByteTrack tracker was trained as part of this work. Therefore, the methodology does not claim an independently trained player detector or multi-object tracking system.

The extracted pose information was subsequently used to construct numerical body-motion representations. These representations formed the input for the conventional machine-learning experiments and for the temporal sequence experiments described in the following subsections. The final proposed RGB-video model did not depend on the pose landmarks as its primary input; instead, it used RGB frames and a pretrained ResNet18 feature extractor.

This separation between the player-related dataset information, MediaPipe-based pose extraction, and the RGB-video classification pipeline was maintained throughout the experiments to distinguish the different representations investigated in the study.

3.6 Pose and Motion Feature Representation

Two forms of pose-based representation were prepared for the experiments: aggregate pose and motion descriptors for conventional machine-learning models, and temporal pose sequences for deep-learning models.

For the conventional machine-learning experiments, each video was represented using **18 numerical features**. These features were designed to capture video characteristics, pose availability, body position, visibility, and movement. The feature set consisted of total frames, processed frames, pose-detected frames, pose detection rate, mean and standard deviation of landmark coordinates in the horizontal and vertical directions, mean and minimum landmark visibility, mean, standard deviation and maximum motion, mean horizontal and vertical motion, mean wrist motion, mean ankle motion, and video duration.

The resulting feature matrices contained 18 numerical variables for each sample. The notebook verified that the features were numeric and that no missing or infinite values were present before classification. These aggregate descriptors were used as inputs to the Dummy Classifier, Logistic Regression, Random Forest, RBF-SVM, and HistGradientBoosting baseline models.

For the temporal pose experiments, the landmark information was retained as a sequence rather than reduced to summary statistics. Each sequence contained 32 temporal frames, with 33 MediaPipe landmarks per frame and four values for each landmark: x-coordinate, y-coordinate, z-coordinate, and visibility. Thus, each frame was represented by 132 values, giving an input representation of 32×132, which can also be viewed as $32 \times 33 \times 4$.

This representation preserved the temporal changes in body configuration and provided the input for the recurrent, temporal convolutional, Transformer-based, and skeleton-based experiments. The pose-based representation was therefore used specifically to investigate how body landmarks and motion descriptors perform for action recognition. The final RGB-video model used a different representation based on frame-level visual features extracted by the pretrained ResNet18 backbone.

3.7 Baseline Machine-Learning Models

The first modelling stage used the 18 handcrafted pose and motion features described in Section 3.6. These features provided a conventional machine-learning baseline against which the subsequent temporal and video-based deep-learning models could be compared.

Five conventional classifiers were evaluated:

1. Dummy Classifier – used as a simple reference baseline.
2. Logistic Regression – used to evaluate a linear decision boundary in the feature space.
3. Random Forest – used to model nonlinear relationships between the extracted pose and motion descriptors.
4. Radial Basis Function Support Vector Machine (RBF-SVM) – used to evaluate nonlinear classification using an RBF kernel.
5. HistGradientBoosting – used as a gradient-boosting model capable of learning nonlinear relationships from the numerical descriptors.

The models received the same 18-dimensional feature representation so that their performance could be compared under a common input representation. The notebook verified the resulting training, validation, and testing matrices and confirmed that the feature values were numerical and contained no missing or infinite values.

Because the seven action classes were not evenly distributed, class-balanced weighting was applied where supported by the individual classifier implementations. This was intended to reduce the influence of the more frequently represented classes and provide a fairer evaluation of the less frequent actions.

The conventional models were used primarily as baseline methods, rather than as the final classification system. Their results were evaluated on the validation subset using accuracy, balanced accuracy, macro-precision, macro-recall, macro-F1, and weighted F1. The validation-based comparison provided a reference for determining whether retaining temporal information and learning visual representations directly from RGB video could provide a more effective representation of the handball actions.

3.8 Temporal Deep-Learning Experiments

After evaluating the handcrafted pose and motion descriptors with conventional machine-learning classifiers, temporal deep-learning experiments were conducted using the prepared 32-step pose sequences[12]. Unlike the aggregate feature representation, the temporal representation retained the changes in body landmarks across successive frames. This allowed the experiments to examine whether sequential modelling could capture movement patterns associated with the seven handball actions.

Several architectures were investigated to provide a systematic comparison of different temporal modelling strategies. The experiments included recurrent, convolutional, hybrid, skeleton-based, and Transformer-based models[10]. The evaluated approaches were:

- BiLSTM + Attention
- Improved BiLSTM + Attention
- Temporal CNN
- Temporal CNN + BiLSTM
- GRU + Attention
- TCN + BiLSTM
- TCN + BiLSTM + Attention
- Transformer Encoder
- ST-GCN-style skeleton model
- HandballPoseNet v4
- HandballPoseNet v4 variant
- BiLSTM + Transformer Fusion
- BiLSTM + Transformer Fusion + Attention
- BiLSTM + Transformer + Attention

These experiments were intended to investigate different mechanisms for modelling temporal changes in the extracted player-pose sequences before moving to the final RGB-video architecture.

The temporal experiments were evaluated using the validation subset. The evaluation considered accuracy, balanced accuracy, macro-precision, macro-recall, macro-F1, weighted F1, and validation loss where these measures were available in the corresponding notebook output. Macro-F1 was treated as the primary comparison measure, since it gives equal importance

to each of the seven action classes and is therefore informative when the class distribution is unequal.
The results from these experiments were used as an intermediate stage in model development. Rather than directly assuming that a more complex temporal architecture would provide the best representation, the experiments allowed recurrent, convolutional, skeleton-based, and Transformer-based approaches to be compared under the prepared pose-sequence representation including BiLSTM-based approaches[8]. The strongest temporal configuration then served as a reference when developing the final RGB-based ResNet18–BiGRU architecture with temporal attention.

3.9 Proposed ResNet18–BiGRU with Temporal Attention

The final video-based classification model was developed using a pretrained ResNet18 backbone followed by a bidirectional GRU (BiGRU) and temporal attention[6]. The purpose of this architecture was to separate spatial feature extraction from temporal modelling. ResNet18 extracts visual information from individual RGB frames, while the BiGRU models the temporal evolution of those features across the action sequence.

For each action clip, 32 RGB frames were sampled and used as the input sequence. The sampled frames were prepared at 112 × 112 pixels before entering the video-processing pipeline. Each frame was passed through the pretrained ResNet18 network after its original classification layer was removed[6]. The resulting representation for each frame was 512-dimensional.

Therefore, every action clip was represented as a sequence of 32 × 512 visual features.

The ResNet18 backbone was kept frozen during feature extraction. No gradient computation was performed for this backbone, and the extracted frame features were cached for the training and validation samples. This reduced repeated computation during temporal-model training and maintained a clear separation between spatial feature extraction and temporal classification.

Before entering the recurrent network, the 512-dimensional frame features were transformed using an input projection consisting of Layer Normalization, a linear layer from 512 to 256 dimensions, GELU activation, and dropout with a rate of 0.15. The resulting 256-dimensional sequence was then provided to a two-layer bidirectional GRU. The GRU used an input size of 256 and a hidden size of 128 in each direction. Since the network operates bidirectionally[7], its output at each temporal position had 256 dimensions. A dropout rate of 0.25 was used within the recurrent component.

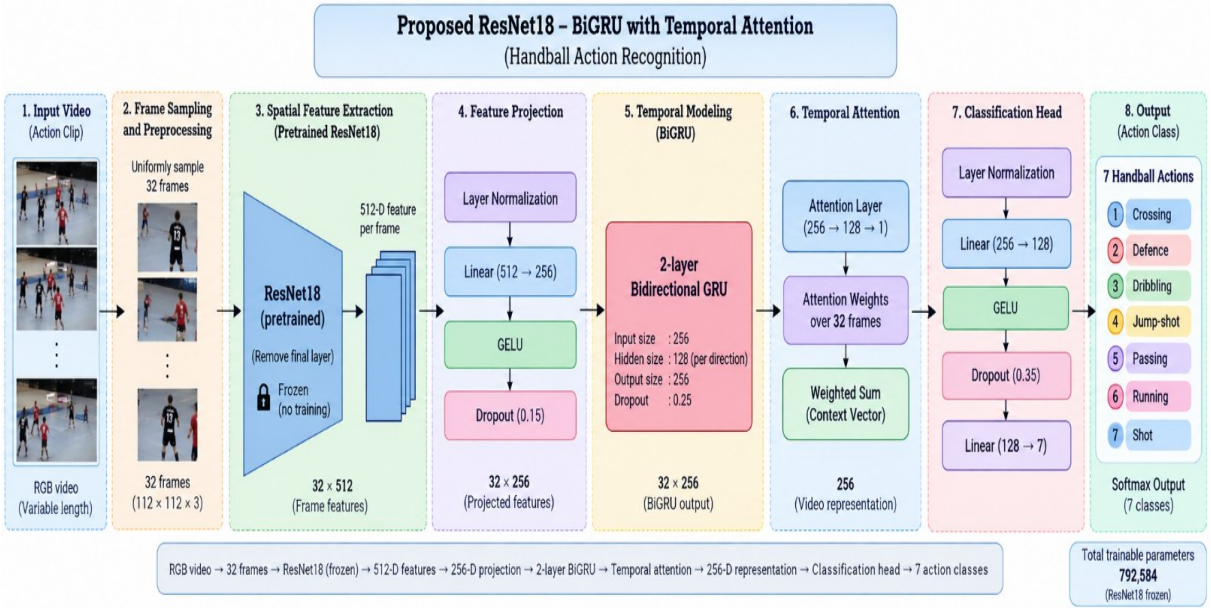
A temporal attention mechanism was applied to the BiGRU output[9]. The attention module calculated a score for each temporal position and used these scores to assign different importance to the feature representations across the sequence[9]. The weighted representations were combined to obtain a 256-dimensional context representation of the action clip. Thus, the model did not rely on treating all sampled frames equally during the final temporal aggregation.

The resulting context representation was passed to the classification head. The classifier consisted of Layer Normalization, a fully connected layer reducing 256 dimensions to 128, GELU activation, dropout with a rate of 0.35, and a final linear layer producing seven class outputs. The seven outputs corresponded to crossing, defence, dribbling, jump-shot, passing, running, and shot.

The complete processing sequence can therefore be expressed as:

RGB video → 32 sampled frames → pretrained ResNet18 → 512-D frame features → 256-D projection → two-layer BiGRU → temporal attention → 256-D context representation → classification head → seven action classes.
The temporal and classification components contained 792,584 trainable parameters, while the ResNet18 backbone remained frozen as the pretrained spatial feature extractor.

An important distinction is that this final model used RGB visual features as its primary input. The MediaPipe Pose representation and the supplied player-detection information described in the earlier subsections were not concatenated with the ResNet18 features. They were used in separate experimental stages to evaluate pose- and motion-based representations.
Proposed ResNet18–BiGRU Architecture



3.10 Training Configuration and Evaluation Metrics

The proposed ResNet18–BiGRU model was trained using 554 action clips, with 115 clips for validation and 140 clips reserved as a locked test set. The test set was not used for training, tuning, or checkpoint selection.

Training used a batch size of 2 for a maximum of 30 epochs. The temporal and classification components used a learning rate of 2×10^{-4}, with the ResNet18 backbone kept frozen during feature extraction. Weight decay was 1×10^{-4}. The loss function used inverse-frequency class weighting with label smoothing of 0.05 to address class imbalance.

A ReduceLROnPlateau scheduler was used, together with early stopping (patience = 7). The checkpoint with the best validation Macro-F1 was retained for final evaluation. The test set remained completely isolated from model development.

Performance was assessed using accuracy, balanced accuracy, precision, recall, Macro-F1, weighted F1, and confusion matrices. Macro-F1 was treated as the primary comparison metric because it gives equal importance to all seven action classes, while the remaining metrics were used to provide complementary overall and class-sensitive evaluation.

IV.RESULTS AND DISCUSSION

4.1 Conventional Machine-Learning Baseline Results

The first experimental stage evaluated the 18-dimensional pose and motion representation using conventional machine-learning classifiers. These experiments were conducted to establish a reference point before evaluating the temporal deep-learning and RGB-video models.

Five classifiers were considered: Dummy Classifier, Logistic Regression, Random Forest, RBF-SVM, and Hist Gradient Boosting. Their performance was evaluated using accuracy, balanced accuracy, macro-precision, macro-recall, Macro-F1, and weighted F1. Since the seven action classes are not equally represented, particular attention was given to the macro-level measures.

Model	Accuracy	Balanced Accuracy	Macro-Precision	Macro-Recall	Macro-F1	Weighted F1
Dummy Classifier	42.53%	14.29%	6.08%	14.29%	8.53%	25.38%
Logistic Regression	40.23%	39.82%	28.71%	39.82%	28.04%	42.04%
Random Forest	55.17%	43.67%	31.83%	43.67%	32.90%	53.48%
RBF-SVM	39.08%	38.17%	29.88%	38.17%	27.04%	42.28%
Hist Gradient Boosting	51.72%	44.09%	42.45%	44.09%	42.96%	50.16%

Table 5. Performance of Conventional Machine-Learning Models

Among the conventional models, HistGradientBoosting produced the highest Macro-F1 of 42.96%, followed by Random Forest with 32.90%. HistGradientBoosting also obtained the highest balanced accuracy among the conventional classifiers. This indicates that the nonlinear boosting model was better able to use the available aggregate pose and motion descriptors than the simpler linear and kernel-based alternatives.

The Random Forest model achieved the highest overall accuracy among the conventional baselines, but its Macro-F1 was lower than that of HistGradientBoosting. This difference is important because overall accuracy can be influenced by the more frequently represented action classes. The lower macro-level performance of most models suggests that the 18 aggregate descriptors did not provide sufficiently discriminative information for all seven actions.

The Dummy Classifier provides the reference lower bound and produced a particularly low Macro-F1. Logistic Regression and RBF-SVM also remained below the stronger nonlinear baselines. Overall, these results motivated the investigation of temporal representations, where the sequential changes in body configuration could be retained instead of reducing each video to a small set of summary statistics.

4.2 Temporal Deep-Learning Model Results

Following the conventional machine-learning experiments, the next stage evaluated temporal deep-learning models using the 32-step pose sequences described in Section 3.6. These experiments were intended to determine whether retaining the temporal evolution of body landmarks could provide a better representation of the seven handball actions than the aggregate 18-feature representation.

A range of architectures was evaluated, including recurrent, convolutional, hybrid, skeleton-based, and Transformer-based models. All results in this stage were obtained on the validation subset, and Macro-F1 was used as the primary comparison measure because it gives equal importance to all seven action classes.

Model	Accuracy (%)	Balanced Accuracy (%)	Macro-Precision (%)	Macro-Recall (%)	Macro-F1 (%)	Weighted F1 (%)	Loss
BiLSTM + Attention	54.02	55.70	55.37	55.70	53.57	55.57	1.9379

Model	Accuracy (%)	Balanced Accuracy (%)	Macro-Precision (%)	Macro-Recall (%)	Macro-F1 (%)	Weighted F1 (%)	Loss
Improved BiLSTM + Attention	48.28	56.18	47.49	56.18	46.57	48.31	2.0119
Temporal CNN + BiLSTM	45.98	58.07	—	—	48.35	50.69	—
Temporal CNN	40.23	52.91	—	—	47.87	40.08	—
Transformer Encoder	42.53	55.60	—	—	48.11	46.85	—
GRU + Attention	41.38	45.22	34.93	45.22	35.45	41.61	1.9481
TCN + BiLSTM	26.44	41.78	40.28	41.78	35.22	30.18	1.8955
ST-GCN-style model	34.48	44.11	37.66	44.11	31.59	34.41	1.8396
TCN + BiLSTM + Attention	34.48	42.17	26.28	42.17	25.97	31.30	2.0260
HandballPoseNet v4	42.53	52.83	47.55	52.83	44.92	46.91	2.0469
HandballPoseNet v4 variant	37.93	52.79	46.13	52.79	41.86	40.91	1.8558
BiLSTM + Transformer Fusion	50.57	43.77	53.87	43.77	45.63	48.41	1.4762
BiLSTM + Transformer Fusion + Attention	45.98	55.45	53.99	55.45	50.38	50.39	1.8810
BiLSTM + Transformer + Attention	33.33	45.73	43.36	45.73	39.10	—	2.0079

Table 6. Validation Results of Temporal Deep-Learning Models

Note: "—" indicates that the corresponding metric was not reported in the final notebook output and has not been inferred.

The BiLSTM + Attention model achieved the highest Macro-F1 among the temporal experiments, with a value of 53.57%. It also obtained 54.02% accuracy and 55.70% balanced accuracy. This result indicates that retaining the temporal sequence of pose information provided a more useful representation than the aggregate handcrafted features used in the conventional baseline stage.

The Temporal CNN + BiLSTM model obtained the highest balanced accuracy at 58.07%, but its Macro-F1 of 48.35% remained below the BiLSTM + Attention result. Similarly, the Transformer Encoder achieved a Macro-F1 of 48.11%, while the Temporal CNN achieved 47.87%. Therefore, balanced accuracy and Macro-F1 did not always rank the models in the same order, supporting the use of multiple evaluation measures.

The BiLSTM + Transformer Fusion + Attention model produced the second-highest Macro-F1 at 50.38%, while the remaining recurrent, convolutional, skeleton-based, and HandballPoseNet variants produced lower Macro-F1 values. The ST-GCN-style model achieved 31.59%, and the TCN + BiLSTM + Attention configuration obtained 25.97%.

Overall, these experiments showed that temporal modelling of pose sequences was useful, but the performance remained limited when compared with the subsequent RGB-based ResNet18–BiGRU with temporal attention model. This motivated the investigation of a representation capable of retaining richer visual information from the video frames rather than relying only on extracted body landmarks.

4.3 Proposed ResNet18–BiGRU with Temporal Attention Results

The final experiment combined frame-level spatial feature extraction with bidirectional temporal modelling and temporal attention. A pretrained ResNet18 network was used to obtain a 512-dimensional representation from each sampled

RGB frame. These frame-level representations were then processed as a temporal sequence by the two-layer BiGRU, followed by temporal attention and the classification layer.

The final model was selected using the validation subset, with Macro-F1 as the checkpoint-selection criterion. The best validation checkpoint was obtained at epoch 16.

Metric	Validation Result
Accuracy	93.91%
Balanced Accuracy	82.75%
Macro-Precision	93.86%
Macro-Recall	82.75%
Macro-F1	86.50%
Weighted F1	93.65%
Validation Loss	0.9709
Best Epoch	16

Table 7. Validation Performance of the Proposed ResNet18–BiGRU with Temporal Attention Model

The proposed model achieved a validation accuracy of 93.91%, with a weighted F1-score of 93.65%. The macro-level results were lower, with a Macro-F1 of 86.50% and macro-recall of 82.75%. This difference reflects the unequal representation of the seven action classes, where performance on more frequent classes contributes more strongly to weighted measures.

The balanced accuracy was 82.75%, equal to the reported macro-recall because both represent the mean recall across the individual classes in this evaluation. The result therefore provides a more class-sensitive view than overall accuracy.

The selected checkpoint had a validation loss of 0.9709. Importantly, the checkpoint was selected according to validation Macro-F1 rather than accuracy, so that model selection considered the performance of the individual action categories rather than relying only on the overall proportion of correct predictions.

4.3.1 Confusion Matrix Analysis

The confusion matrix of the proposed ResNet18–BiGRU model with temporal attention was examined to understand the distribution of correct and incorrect predictions across the seven action classes. Rows represent the actual classes, while columns represent the predicted classes. The analysis is based on the 115 samples in the validation subset.

Actual \ Predicted	Crossing	Defence	Dribbling	Jump-shot	Passing	Running	Shot
Crossing	23	0	0	1	0	0	0
Defence	0	5	0	0	0	0	2
Dribbling	0	0	2	0	1	0	0
Jump-shot	1	1	0	41	0	0	0
Passing	0	0	0	0	23	0	0
Running	0	0	0	1	0	1	0
Shot	0	0	0	0	0	0	13

Table 8. Confusion Matrix of the Proposed Model on the Validation Subset

The model correctly classified 108 of the 115 validation samples, with seven misclassified samples. The crossing class had 23 correct predictions out of 24, with one sample classified as jump-shot. The jump-shot class had 41 correct predictions out of 43, with one prediction assigned to crossing and another to defence.

The passing and shot classes were classified correctly for all their validation samples. Passing had 23 correct predictions, while shot had 13 correct predictions.

The remaining errors were concentrated in the less represented classes. Two defence samples were predicted as shot, one dribbling sample was predicted as passing, and one running sample was predicted as jump-shot. Because these classes contain only a small number of validation samples, individual errors have a relatively large effect on their class-wise performance.

Overall, the confusion matrix indicates that the proposed model produced a limited number of classification errors on the validation subset. The errors were concentrated in a small number of class pairs rather than being distributed broadly across all seven actions. Nevertheless, the results for classes with very few validation samples should be interpreted cautiously, and the locked test subset was subsequently evaluated independently after completion of model selection, as described in Section 4.6.

4.4 Training Behaviour of the Proposed Model

The training behaviour of the proposed ResNet18–BiGRU model with temporal attention was evaluated using the recorded training and validation history. The model was trained for a maximum of 30 epochs, with early stopping applied using

a patience of seven epochs. Model selection was based on validation Macro-F1, rather than overall accuracy, to provide a more balanced evaluation across the seven action classes.

The classifier and temporal components were initialized with a learning rate of $\mathbf{2 \times 10^{-4}}$, while the ResNet18 backbone was kept frozen during training. A ReduceLROnPlateau learning-rate scheduler was applied based on validation Macro-F1, with a reduction factor of 0.5 and a scheduler patience of three epochs. The recorded learning-rate schedule was 2×10^{-4} for epochs 1–6, 1×10^{-4} for epochs 7–14, 5×10^{-5} for epochs 15–19, and 2.5×10^{-5} for epochs 20–23. The weight-decay coefficient was set to 1×10^{-4}.

The best validation performance was obtained at epoch 16, with a validation accuracy of 93.91% and Macro-F1 of 86.50%. Training continued after this point to allow further improvement, but no subsequent epoch produced a higher validation Macro-F1. Consequently, early stopping was triggered at epoch 23, and the checkpoint from epoch 16 was restored as the final selected model.

The training history therefore demonstrates that the learning-rate reduction helped refine the model after the initial rapid learning phase. The increasing training performance together with comparatively fluctuating validation performance also indicates the importance of validation-based checkpoint selection. The final reported model corresponds to the best validation checkpoint at epoch 16, rather than the final epoch of training.

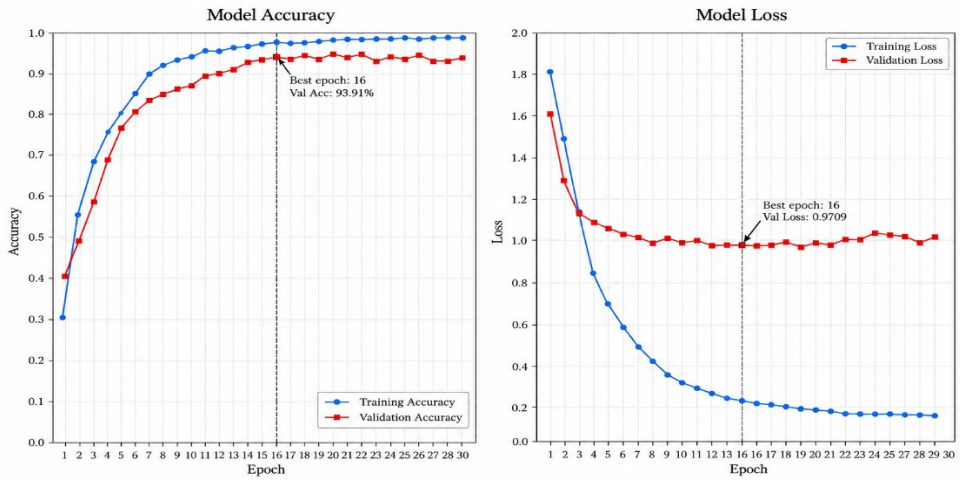


Figure 2. Training and validation performance of the proposed ResNet18–BiGRU with temporal attention across training epochs.

The learning-curve analysis should be interpreted together with the checkpoint-selection procedure rather than using the final training epoch as the model's reported performance. The reported model corresponds to the best validation checkpoint, while the independent test subset remained locked during model development.

4.5 Class-Wise Performance of the Proposed Model

The class-wise performance of the proposed ResNet18–BiGRU model with temporal attention was examined using the predictions obtained on the validation subset. Since the seven action categories are not equally represented, class-wise measures provide additional information beyond overall accuracy. The analysis considers the number of validation samples, correct predictions, misclassified samples, and recall for each action class.

Action Class	Validation Samples	Correct Predictions	Misclassified	Class-wise Recall
Crossing	24	23	1	95.83%
Defence	7	5	2	71.43%
Dribbling	3	2	1	66.67%
Jump-shot	43	41	2	95.35%
Passing	23	23	0	100.00%
Running	2	1	1	50.00%
Shot	13	13	0	100.00%
Total	**115**	**108**	**7**	**93.91% overall accuracy**

Table 9. Class-Wise erformance of the Proposed Model on the Validation Set

The proposed model correctly classified 108 of the 115 validation samples. Passing and shot achieved complete classification within the validation subset, with all 23 passing samples and all 13 shot samples correctly identified. Crossing and jump-shot also showed high recall, at 95.83% and 95.35%, respectively.

The lower recall values occurred for defence, dribbling, and running. Defence achieved 71.43% recall, while dribbling achieved 66.67%. Running had the lowest recall at 50.00%; however, only two validation samples belonged to this class. Consequently, a single incorrect prediction has a substantial effect on the reported percentage for this class.

These results demonstrate why the evaluation should not rely on accuracy alone. The class-wise results should be considered together with Macro-F1 and balanced accuracy, particularly because the validation classes contain substantially different numbers of samples.

4.6 Final Independent Test Evaluation

After completion of model development and checkpoint selection, the locked test subset was evaluated once using the best ResNet18–BiGRU with temporal attention checkpoint selected at epoch 16. The test subset contained 140 action clips originating from 100 scenes, with no overlap between the training and test scenes. The test subset was not used for training, model selection, hyperparameter tuning, or early stopping.

The proposed model achieved 75.00% accuracy and 75.80% balanced accuracy on the independent test subset. The model obtained 73.51% macro-precision, 75.80% macro-recall, 66.19% Macro-F1, and 76.38% weighted F1, with a test loss of 1.3526. These results represent the final independent evaluation of the selected model.

Metric	Test Result
Accuracy	75.00%
Balanced Accuracy	75.80%
Macro-Precision	73.51%
Macro-Recall	75.80%
Macro-F1	66.19%
Weighted F1	76.38%
Test Loss	1.3526
Test Samples	140

Table 10. Final Independent Test Performance of the Proposed Model

The independent test results are lower than the validation results reported in Table 6. Validation accuracy was 93.91%, whereas independent test accuracy was 75.00%. Macro-F1 decreased from 86.50% on validation data to 66.19% on the test data. This difference indicates that the selected model generalizes less effectively to previously unseen scenes than suggested by the validation performance. The test result should therefore be considered the more conservative estimate of the model's generalization capability.

The test-set class distribution was also substantially imbalanced. The test subset contained 40 crossing, 13 defence, 6 dribbling, 41 jump-shot, 24 passing, 1 running, and 15 shot samples. Therefore, class-wise results for the running category should be interpreted cautiously because they are based on only one test sample.

Actual / Predicted	Crossing	Defence	Dribbling	Jump-shot	Passing	Running	Shot
Crossing	40	0	0	0	0	0	0
Defence	6	6	0	0	0	1	0
Dribbling	1	0	4	0	0	1	0
Jump-shot	3	0	0	23	0	5	10
Passing	1	0	1	2	18	2	0
Running	0	0	0	0	0	1	0
Shot	1	0	0	0	0	1	13

Table 11. Confusion Matrix of the Proposed Model on the Independent Test Set

The confusion matrix shows that the model correctly classified 105 of the 140 test clips, resulting in 35 errors. The largest confusion occurred between jump-shot and shot, where 10 jump-shot samples were classified as shot. Defence was also frequently classified as crossing, accounting for six errors. Five jump-shot samples were classified as running, while smaller numbers of errors occurred among passing, dribbling, and shot.

Action Class	Test Samples	Correct	Misclassified	Recall
Crossing	40	40	0	100.00%
Defence	13	6	7	46.15%
Dribbling	6	4	2	66.67%
Jump-shot	41	23	18	56.10%
Passing	24	18	6	75.00%
Running	1	1	0	100.00%

Action Class	Test Samples	Correct	Misclassified	Recall
Shot	15	13	2	86.67%
Total	**140**	**105**	**35**	**75.00%**

Table 12. Class-Wise Performance on the Independent Test Set

The class-wise results show that crossing achieved 100.00% recall, while shot and passing achieved 86.67% and 75.00%, respectively. Dribbling achieved 66.67% recall. The lowest recall among the classes with multiple test samples was observed for jump-shot at 56.10%, followed by defence at 46.15%. The dominant jump-shot-to-shot confusion suggests that these two shooting-related actions can be visually similar when the temporal evidence for the jumping phase is insufficient or difficult to distinguish.

Overall, the independent test evaluation demonstrates that the proposed architecture retains useful classification capability on unseen scenes, while also revealing a larger generalization gap than indicated by validation performance. The results support the use of scene-level separation and independent testing for obtaining a more realistic assessment of handball action-recognition performance.

4.7 Overall Comparison of the Evaluated Models

The experiments were conducted in successive stages to examine different representations and temporal modelling strategies for seven-class handball action classification. The first stage evaluated conventional machine-learning classifiers using aggregate pose and motion descriptors. This was followed by temporal deep-learning experiments using 32-frame pose sequences. The final stage investigated a video-based architecture using a pretrained ResNet18 feature extractor followed by a bidirectional GRU and temporal attention.

The conventional machine-learning experiments used the original scene-level partition containing 407 training scenes, 87 validation scenes, and 88 test scenes. The temporal pose experiments used the same scene-level partition. In both stages, the test subset remained locked and was not used for model selection or tuning.

The final action-only ResNet18–BiGRU experiment was conducted after the action-to-scene association and dataset verification stages. This resulted in 809 labelled action clips associated with 583 unique scenes. The final partition contained 554 training actions, 115 validation actions, and 140 test actions, with no scene overlap between the three subsets. The test portion was kept locked throughout model development and was evaluated only after the final model-selection decisions had been completed.

Model / Method	Accuracy (%)	Balanced Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro-F1 (%)	Weighted F1 (%)
Dummy Classifier	42.53	14.29	6.08	14.29	8.53	25.38
Logistic Regression	40.23	39.82	28.71	39.82	28.04	42.04
Random Forest	55.17	43.67	31.83	43.67	32.90	53.48
RBF-SVM	39.08	38.17	29.88	38.17	27.04	42.28
HistGradientBoosting	51.72	44.09	42.45	44.09	**42.96**	50.16
BiLSTM + Attention	54.02	55.70	55.37	55.70	**53.57**	55.57
Improved BiLSTM + Attention	48.28	56.18	47.49	56.18	46.57	48.31
Temporal CNN + BiLSTM	45.98	58.07	—	—	48.35	50.69
Temporal CNN	40.23	52.91	—	—	47.87	40.08
Transformer Encoder	42.53	55.60	—	—	48.11	46.85
GRU + Attention	41.38	45.22	34.93	45.22	35.45	41.61
TCN + BiLSTM	26.44	41.78	40.28	41.78	35.22	30.18
ST-GCN-style Model	34.48	44.11	37.66	44.11	31.59	34.41

Model / Method	Accuracy (%)	Balanced Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro-F1 (%)	Weighted F1 (%)
TCN + BiLSTM + Attention	34.48	42.17	26.28	42.17	25.97	31.30
HandballPoseNet v4	42.53	52.83	47.55	52.83	44.92	46.91
HandballPoseNet v4 Variant	37.93	52.79	46.13	52.79	41.86	40.91
BiLSTM + Transformer Fusion	50.57	43.77	53.87	43.77	45.63	48.41
BiLSTM + Transformer Fusion + Attention	45.98	55.45	53.99	55.45	50.38	50.39
ResNet18 + BiGRU + Temporal Attention	**93.91**	**82.75**	**93.86**	**82.75**	**86.50**	**93.65**

Table 13. Overall Validation Performance Across the Experimental Stages

Note: "—" indicates that the corresponding metric was not available in the notebook output and has not been reconstructed.

The conventional models provided the initial reference point for the classification task. Among these models, Random Forest obtained the highest validation accuracy at 55.17%, while HistGradientBoosting produced the highest Macro-F1 of 42.96%. This indicates that the aggregate pose and motion descriptors contained useful information for distinguishing the action categories, although their ability to separate the classes was limited. The corresponding baseline experiments were performed using 407 training samples and 87 validation samples from the original scene-level split.

The temporal pose experiments examined whether retaining the sequence of body landmarks could provide a stronger representation than aggregate descriptors. Among the evaluated temporal architectures, BiLSTM + Attention achieved the highest Macro-F1 of 53.57%, with a validation accuracy of 54.02%. Other architectures, including Transformer, temporal convolutional, skeleton-based, and hybrid recurrent models, produced lower Macro-F1 values in the recorded experiments.

The final ResNet18–BiGRU with temporal attention model produced substantially higher validation performance than the earlier experimental configurations. The model achieved 93.91% validation accuracy, 82.75% balanced accuracy, 93.86% macro-precision, 82.75% macro-recall, 86.50% Macro-F1, and 93.65% weighted F1. These values were obtained at the selected validation checkpoint and should therefore be described as validation performance rather than test performance.

The improvement in Macro-F1 is particularly notable when considering the progression from the conventional and temporal pose representations to the RGB-video representation. HistGradientBoosting achieved the highest Macro-F1 among the conventional classifiers at 42.96%, while BiLSTM + Attention achieved 53.57% among the earlier temporal models. The proposed RGB-video model reached 86.50% Macro-F1.

The experimental progression suggests that the representation used by the proposed model provides information that was not sufficiently captured by the aggregate pose descriptors or the temporal pose representation. Instead of reducing the video to manually constructed motion statistics or body landmarks, the proposed pipeline extracts frame-level visual features using the pretrained ResNet18 and subsequently models their temporal evolution using the BiGRU and temporal attention mechanism.

However, the comparison should be interpreted with the split difference described above. The conventional and temporal pose experiments were performed using the original 407/87/88 scene-level partition, whereas the final action-only model used the subsequently verified 554/115/140 action-level partition. Consequently, the table represents the experimental development and validation progression of the study, rather than a strictly controlled same-validation-set benchmark between every model.

4.8 Comparison with Existing UNIRI-HBD-Based Studies

To place the present results in context, the proposed method was compared with studies using the original UNIRI-HBD Version 1 dataset. Studies based on the later UNIRI-HBD_v2 dataset were excluded because of differences in dataset construction and action configuration. The original release contains 751 labelled videos covering seven handball action categories.

Study	Dataset / Version	Main Representation	Main Method	Evaluation / Result	Relation to Present Work
Pobar and Ivašić-Kos (2020)	UNIRI-HBD-related handball video data	Player activity / visual activity measures	Active-player detection based on activity measures	Focused on identifying the active player rather than seven-class action recognition	Provides background for player and activity analysis
Ivašić-Kos and Pobar (2021)	UNIRI-HBD Version 1	Labelled RGB video clips	Dataset release	751 labelled videos across 7 action classes	Establishes the dataset used in the present study
Katsioulas and Maglogiannis (2026)	UNIRI-HBD original 751-clip release	Trajectory-derived kinematic and scene features	Random Forest, Gradient Boosting, Extra Trees, XGBoost and other classical classifiers	Best test accuracy 80.5%; Random Forest Macro-F1 74.5%	Most directly comparable published action-recognition study, but uses different representations and evaluation protocol
Present study	UNIRI-HBD Version 1; 809 action clips	RGB frames → ResNet18 features → BiGRU → temporal attention	ResNet18–BiGRU with temporal attention	Best validation accuracy 93.91%; validation Macro-F1 86.50%; independent test accuracy 75.00%; test Macro-F1 66.19%	Learns spatial-temporal representations directly from RGB video; independent test results provide a stricter estimate of generalization

Table 14. Comparison with Existing Studies Using the Original UNIRI-HBD Dataset

4.9 Discussion of Results

The experimental results show a clear improvement from aggregate pose and motion descriptors to temporal and RGB-based representations. The conventional machine-learning models achieved limited performance, while temporal pose modelling provided moderate improvement. The proposed ResNet18–BiGRU with temporal attention achieved the strongest validation performance, reaching 93.91% accuracy and 86.50% Macro-F1. This improvement can be attributed to the combination of frame-level visual features, bidirectional temporal modelling, and attention-based aggregation.

However, evaluation on the independent locked test set provided a more conservative estimate of generalization. The model achieved 75.00% accuracy, 75.80% balanced accuracy, and 66.19% Macro-F1. The reduction from validation to test performance indicates a noticeable generalization gap when the model is applied to previously unseen scenes.

Class-wise results showed that crossing and shot were recognized relatively well, whereas defence and jump-shot were more challenging. The most frequent error was confusion between jump-shot and shot, while defence was frequently confused with crossing. These errors suggest that actions with similar visual and temporal characteristics remain difficult to distinguish. The very small number of test samples for some classes, particularly running, also limits the reliability of class-wise conclusions.

The comparison with earlier experimental configurations should be interpreted cautiously because the conventional and temporal pose experiments used the original scene-level partition, whereas the final action-only model used the subsequently verified 554/115/140 action split. Therefore, the results demonstrate the progression of the investigated approaches but do not constitute a strictly controlled same-split comparison.

Overall, the findings indicate that combining pretrained spatial features with bidirectional temporal modelling and attention is a promising approach for handball action recognition. Nevertheless, the independent test results demonstrate that generalization to unseen scenes remains the main challenge, motivating future work on larger datasets, underrepresented classes, improved temporal representations, and multimodal information.

V. CONCLUSION AND FUTURE SCOPE

This study investigated video-based handball action classification across seven action categories of the UNIRI-HBD Version 1 dataset. Among the evaluated approaches, the proposed ResNet18–BiGRU model with temporal attention achieved the strongest validation performance, with 93.91% accuracy and 86.50% Macro-F1.

The selected model was further evaluated on an independent locked test set of 140 action clips, achieving 75.00% accuracy, 75.80% balanced accuracy, and 66.19% Macro-F1. The test results provide a more conservative measure of generalization. Class-wise evaluation showed strong performance for crossing and shot, while defence and jump-shot remained more challenging. The main confusion occurred between visually and temporally similar actions, particularly jump-shot and shot.

The study also highlights the importance of scene-based data separation and class-sensitive metrics for evaluating imbalanced action-recognition datasets. However, the limited number of samples in some classes and the use of a single dataset restrict the generalizability of the findings.

Future work will focus on increasing underrepresented samples, fine-tuning stronger spatial-temporal feature extractors, improving temporal modelling, and combining RGB features with pose, player trajectories, ball movement, and contextual information. Evaluation on larger datasets and cross-dataset testing will also be considered to assess the robustness and transferability of the proposed approach.

REFERENCES

1. M. Pobar and M. Ivašić-Kos, "Active Player Detection in Handball Scenes Based on Activity Measures," Sensors, vol. 20, no. 5, p. 1475, 2020, doi: 10.3390/s20051475.
2. M. Ivašić-Kos and M. Pobar, "Handball Action Dataset – UNIRI-HBD," IEEE DataPort, 2021, doi: 10.21227/0g0a-fe06.
3. K. Host, M. Ivašić-Kos, and M. Pobar, "Action Recognition in Handball Scenes," Lecture Notes in Networks and Systems, vol. 283, pp. 645–656, 2022, doi: 10.1007/978-3-030-80119-9_41.
4. K. Host and M. Ivašić-Kos, "An overview of Human Action Recognition in sports based on Computer Vision," Heliyon, vol. 8, no. 6, p. e09633, 2022, doi: 10.1016/j.heliyon.2022.e09633.
5. K. Host, M. Pobar, and M. Ivašić-Kos, "Analysis of Movement and Activities of Handball Players Using Deep Neural Networks," Journal of Imaging, vol. 9, no. 4, p. 80, 2023, doi: 10.3390/jimaging9040080.
6. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
7. K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.
8. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
9. D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in International Conference on Learning Representations (ICLR), 2015.
10. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems, vol. 30, 2017.
11. S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018, pp. 7444–7452, doi: 10.1609/aaai.v32i1.12328.
12. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal Convolutional Networks for Action Segmentation and Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1003–1012, doi: 10.1109/CVPR.2017.113.
13. V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose Tracking," arXiv:2006.10204, 2020.