

Soccer Video Analysis and Highlight Detection Using Deep Neural Networks

M. Bhanu Prakash Reddy¹; S. Mano Venkat²

¹Assistant Professor; ²M.Tech (Master of Technology) Final Semester

^{1,2}Sanketika Vidya Parishad Engineering College, Visakhapatnam, Andhra Pradesh, India

Publication Date: 2026/10/13

Abstract: The groundwork for event detection in football match analysis using audio. It describes the structure, annotated event categories and preparation of synchronized audio data for the Soccer Action Compilation (SADC) dataset. The chapter outlines the major steps in the pre-processing such as segmentation, noise reduction and feature extraction of matchsounds using Mel-Frequency Cepstral Coefficients (MFCCs), which are important for the temporal and spectral properties. The design and implementation of the audio event detection model are explained and the key event detection such as goal, goal attempt, penalty and yellow card is explained with the use of deep learning architectures. Training configurations, parameter selection and evaluation strategies are discussed to make sure that the models are reliable and reproducible. Results of the model are presented using quantitative measures and confusion matrices, and are followed by a discussion on important results, strengths and limitations that were observed. In general, this chapter provides the methodological and analytical framework for the subsequent visual and multimodal event detection in football videos, while showing the effectiveness of using audio cues in understanding the dynamics of a game and summarising the key moments. The topics are related to Multimodal Deep Learning, Skeletal Motion Analysis, Cross-Cultural Dance Recognition and Cultural Heritage Informatics and Explainable AI (XAI).

Keywords: Learning Architecture, MFCC, SADC, Skeletal Motion Analysis, Cross-Cultural Dance Recognition.

How to Cite: M. Bhanu Prakash Reddy; S. Mano Venkat (2026) Soccer Video Analysis and Highlight Detection Using Deep Neural Networks. *International Journal of Innovative Science and Research Technology*, 11(10), 429-435.
<https://doi.org/10.38124/ijisrt/26oct296>

I. INTRODUCTION

Performing arts studies are undergoing dramatic change as digital tools are increasingly used to document, analyze and preserve the culture. Dance is among these art forms, and is unique in that it is dependent on movement, rhythm, emotion and cultural symbolism. Dance is different from other traditions, such as writing and speech, in that it is represented by complex spatial and temporal patterns which are not easy to interpret systematically. The use of digital archives and motion capture has revealed more variation in styles across dances from all world areas, but furthermore advanced techniques are required to address the topic.

In comparison with Indian classical (nontheatrical) and western dance forms. The art forms of Bharatanatyam or Kathak are part of the Indian classical traditions, which are well established and codified. They feature exact hand movements (mudras), expressive storytelling style (abhinaya) and rhythm cycles (taalas) which help to lead the movement and the emotion. In contrast to the western styles like ballet and modern, the dance has a strong connection with the body. Ballet and modern dance, on the other hand, have more to do with the body.

Often dance will focus on moving the body in space, moving between one motion to another and in many instances improvisation of the dance interpretation. These differences at the root of the subject gives birth to a special artistic and cultural gap.

In the modern era, the choreographer is more inclined towards the experimentations and gives birth to a fusion performance, which is a combination of Indian classical and western styles. However, the analysis of such hybridization is quite significant. Rythm, stylistic borrowing and nuances of cultural exchange can sometimes play out in the subtleties of movement and expression. With the increasing speed of artistic interaction in the era of globalization, researchers, educators and cultural practitioners need to be able to identify and understand such moments of fusion in a systematic way.

Although there is a current interest among academics to analyse dance, the existing research is mainly focused on analyzing individual styles separately. Very little work has been done on the study of the dynamic interaction between different cultural systems in the formation of new expressive vocabularies. In this study, it is proposed that a

structured approach be developed to study classical dance fusion between Indian classical and Western dance styles to fill in the gap. The framework will focus on the identification of the opportunities for stylistic integration and the way cultural characteristics relate to performance using both motion-based inputs and rhythmic/symbolic inputs.

Finally, this study adds to the current work in digital dance studies by providing a methodology that can not only be used to capture the patterns of dance but can also also recognize the cultural dialogue that exists within the pattern. This knowledge is vital for the maintenance of artistic tradition, the development of new dances and greater appreciation of cultural exchange among dance communities around the world.

II. LITERATURE REVIEW

The study of dance through digital means has increasingly become important in the study of dance, in which the development of motion-capture systems, video-based analysis of movement and computational analysis of performance have all played significant roles. The first research in this field was mainly concerned with human motion tracking and gesture interpretation, which became the foundations for more exact tracking and assessment of gestures. By analyzing systematic body motion in a structured way, a series of studies with motion-sequence collections like the CMU Motion Capture and the NTU RGB+D collections have shown the potential of systematic observation of body trajectories for structured analysis of complex physical expression in dance and related performance traditions (Weichert, Rinkenauer, & Bachmann, 2018). It is a groundbreaking book that has contributed to the creation of frameworks that can identify various aspects of movement and categorize various dance genres taught as separate subjects.

In the course of time, researchers started to investigate more holistic approaches, combining several sources of information to encompass more aspects of the arts and culture. Research that integrates the context of performance, rhythm and symbolic gestures with posebased movement data has demonstrated that the addition of annotations and rhythmic cues significantly increases interpretative understanding. Kim et al. (2010), for example, emphasized that the use of the auditory (rhythm) cue in addition to body movement (feature) would help to recognize Emotional expression in a performance context. Similarly, Liu et al. (2023) demonstrated the potential of combining motion analysis with cultural descriptions to enhance the understanding of the meaning of culturally rich gestures of performance and stylistic markers.

In the context of traditional dance scholarship, Bharatanatyam and Kathak dance styles of India have been noted for their intricate expressive system of narration, codified vocabulary of movements and complex tala systems. Das, Paul and Rao (2025) provided an ontology-based representation to formally map symbolic gestures and narrative elements, which could be a structured model to

retain cultural semantics in digital archives responding to the structural depth of these traditions. These sorts of contributions highlight the importance of developing an analytical framework that is grounded in the culture to analyze performance traditions that have developed such long-standing aesthetics and pedagogies.

There has been a growing emphasis on stylistic crossfertilisation in intercultural and cross-genre dance research, evidenced in references to the “blending of styles” across performance contexts in the world today. Li (2024) mentioned that the fusion dance education is beneficial to the development of culture literacy and artistic innovation, but pointed out that there are no systematic analysis methods for blended dance. Jiao and Zhao (2024) further discussed key conceptual issues in the learning and teaching of cross-cultural choreography, highlighting the need for intercultural analysis that is attuned to the unique movement philosophies and expressiveness of different cultural contexts. This need for culturally diverse discussions of safeguarding intangible heritage has centered in wider discourses on safeguarding.

The scholarship and analytical tools that contextualized the analyses, such as those of Gîrbacia (2024) who analyzed digital tools for documenting traditional performance without compromising specificity and the sense of place, were also explored.

Although there has been While some strides have been made in the fields of movement research and studies of culture, there are places that still require research. The majority of previous studies are about single style movement classification or theoretical analysis of fusion, but lack analytical mechanisms. Furthermore, there is no specific labelled data that is representative of blended dance styles and this would help the empirical advances. Like “Dance Shala” (Mukherjee, 2022), most large datasets are devoted to documentation of individual styles of dancing, rather than hybrid dances. Finally, while much progress has been made in motion analysis and documentation of dance, few attempts have been made to look for and understand stylistic convergence of the movement traditions of India and the West.. In this study, the need to detect and analyze the fusion points in a choreography was addressed by providing a structured methodological framework, based on movement, rhythmic cues and symbolic descriptors. This work contributes to the continuing exploration and understanding of global dance heritage by enriching digital access to these resources and adding to scholarly tools and technologies for the understanding of cross-cultural performance practice.

III. METHODOLOGY

To preserve all the original information that may have been in the videos, all were downloaded in full, allowing for the inclusion of the original voiceover and crowd noise. The Video Dataset is as per the attached Table 3.1 below:

The manual annotation was carried out by 5 expert football fans who were trained to recognize and label

some of the key events in a football match. All the annotators marked each game independently and wrote the start and end times of any actions that they felt were important, on their own annotation tool. These events of interest were:

- Goal (including own goals)

Attempt to score with a goal, when there are clear shots on target or near misses.

- Penalty kick
- Free kick in dangerous areas
- Corner kick
- Yellow card

The annotators recorded the event type, start time, end time and the name of the audio file for each occurrence in a structured comma-separated value (.csv) file. In order to get rid of differences, a consensus pass was subsequently performed. Crowd noise was used to identify excitement but visual evidence for an event was hard to tell, so the segment was marked for a second review to ensure a high inter-annotator agreement. The original rich set of labels for the data was later simplified to a simple binary classification

of EVENT versus NO EVENT to make it consistent to machine learning. There is a clear distinction between high impact and routine play and the simplified representation can be used in both binary classifiers or more complex multi-class experiments.

➤ Event Annotation Logic

The rule to follow for the annotation was: crowd's reaction is closely linked with events in the game. The cheers and comments by the officials were thus an important cue for the annotators. If the cheering was above a specified level, this segment was identified as EVENT; if not, it was identified as NO EVENT. This human in the loop process is illustrated with audio cues being a real-time indicator of event boundaries..

The guideline for annotation was that the reaction of the crowd is correlated well with the key events of a game. Audible excitement during the cheers and excitement of the commentator was thus a significant cue by the annotators. The segment was classified as EVENT if the volume of cheers was above a certain threshold; otherwise it was classified as NO EVENT. Figure 1 shows this human in the loop, in which the audio cues provide a live indication of event boundaries. The extended-ResNet50 model is a proposed transfer learning model that is detailed below.

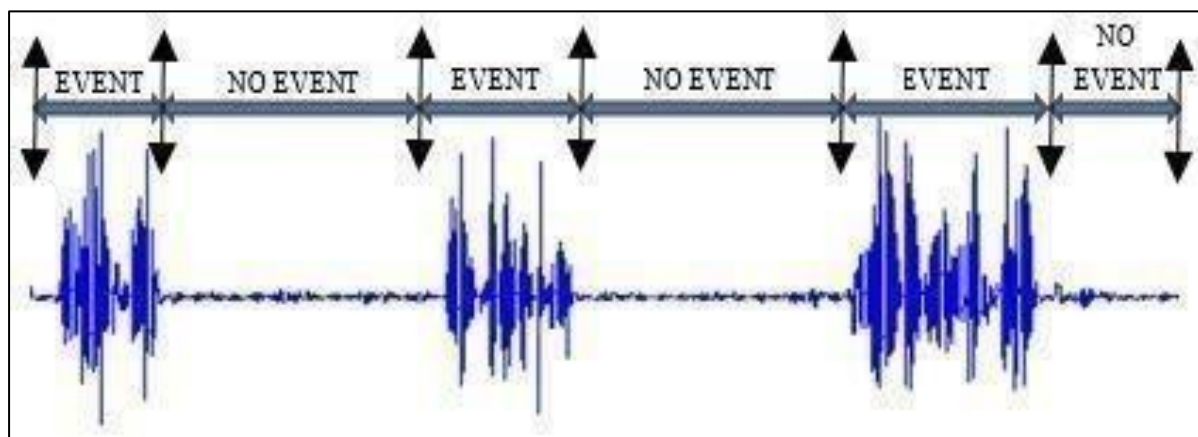


Fig 1 Process of Recording Events

➤ Time-Window Segmentation

For training stability, the raw events were further segmented into fixed-length windows of 40, 50, 60, and 90 seconds. This design captures not only the precise instant of an event but also the build-up and aftermath. For example, if a goal occurred at the 500 second mark, a

60-second segment spanning 470s to 530s was extracted. The same strategy was applied to NO EVENT intervals to maintain class balance. Further this events were classified into two different classes EVENT and NOEVENT for binary classification as shown in Table 1:

Table 1 Recorded Events in SADC

Name of Event	Start time (in Sec.)	End time (in Sec.)	File name
GOAL	0	40	Match1.mp3
NO EVENT	41	101	Match1.mp3
NO EVENT	102	457	Match1.mp3
PENALTY	458	466	Match1.mp3
.....			Match1.mp3
FREE KICK	6165	6214	Match1.mp3

Table 2 Processed Events in SADC

Name of Event	Start time (in Sec.)	End time (in Sec.)	File name
EVENT	0	40	Match1.mp3
NO EVENT	41	125	Match1.mp3
NO EVENT	126	185	Match1.mp3
.....			Match1.mp3
EVENT	5561	5650	Match1.mp3
EVENT	5651	5740	Match1.mp3

➤ Time-Window Segmentation

In order to enhance training stability, the raw events were additionally divided into fixed-length windows of 40, 50, 60 and 90 seconds. This design not only captures the exact moment of an event but also the buildup and aftermath. If the goal was at the 500 second mark, then a section of 60 seconds from 470s to 530s was removed. A similar approach was taken for NO EVENT intervals to ensure a balance of classes. Further this event was classified in to two different classes EVENT and NOEVENT for binary classification as shown in Table 2. This sliding-window approach allows taking account of the context in order to enhance the model performance with respect to using individual event frames.

This dataset has been carefully designed for both unimodal and multimodal learning tasks.

➤ Audio Extraction

The entire audio track was demultiplexed from each match video using MoviePy library to create .mp3 files. Audio was separated from video, to improve data loading time and to facilitate pre-processing steps, like generating a spectrogram. Each .mp3 file is named in a manner such that it always has the same name as its parent video, so alignment with the .csv based annotation is easy. Thus the final data set is made up of: Raw audio files for further processing (Training Sound Files) and SADC Dataset.

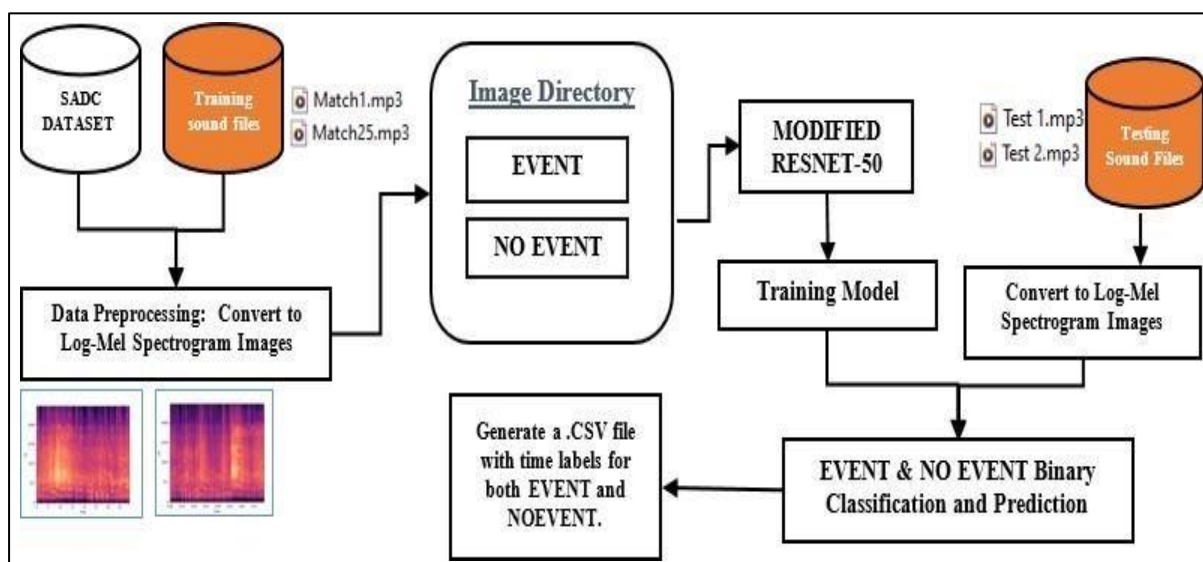


Fig 2 Process Diagram of Audio Event Recognition Framework and Prediction

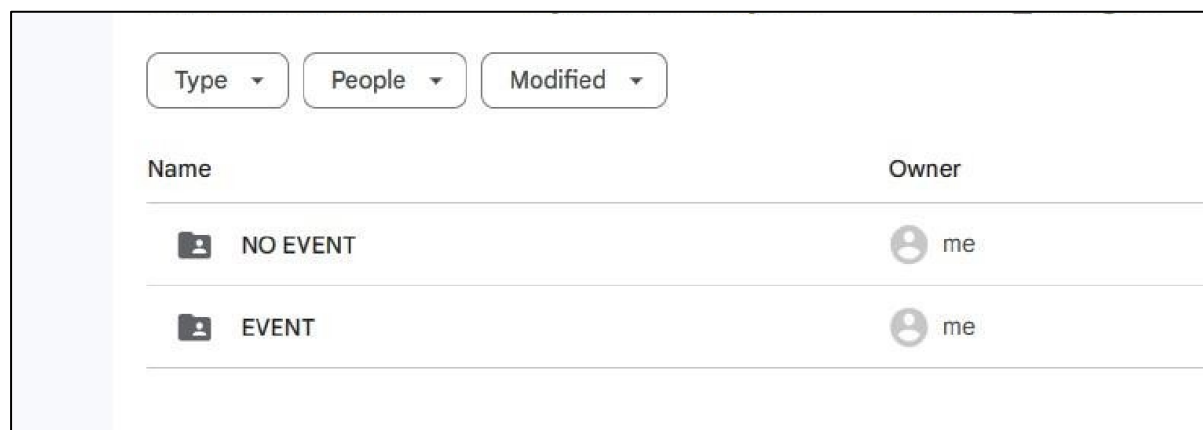


Fig 3 Screenshot of Separate Folders for Storing EVENT and NO EVENT MFCC Image

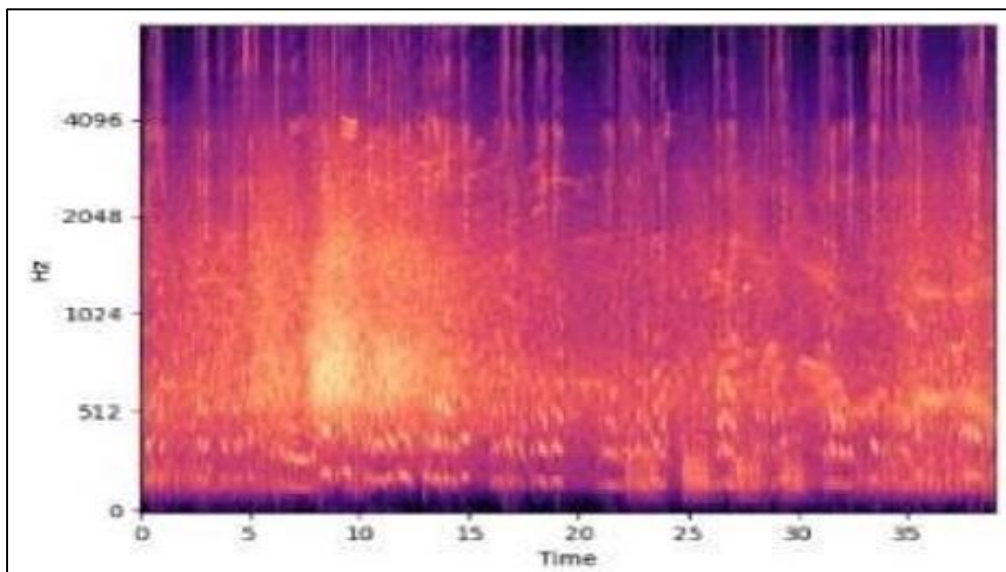


Fig 4 MFCC Image of “EVENT”

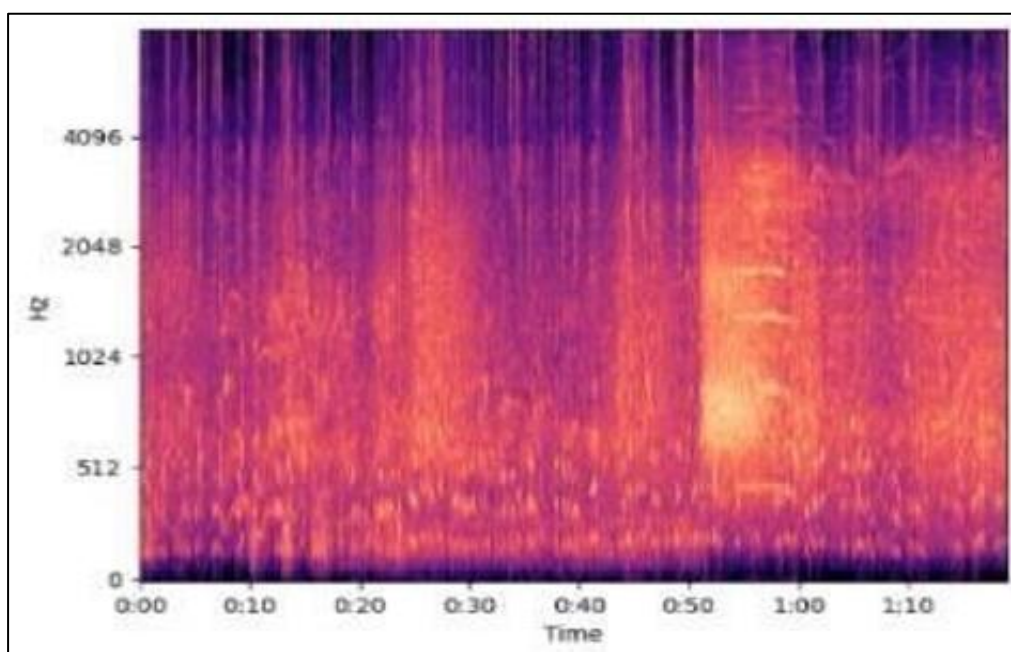


Fig 5 MFCC Image of “NO EVENT”

➤ Proposed Audio Framework

The audio model aims to label every time segment of a match as EVENT or NO EVENT, based solely on the sounds. This section outlines the raw audio loading to prediction pipeline. Complete workflow is illustrated. The first step of the pipeline is to read the .mp3 clips and annotation tables. This is done using AudioFileClip function from the MoviePy library that calculates the duration and allows for exact alignment with the start and end time labels. Each of 60 seconds is segmented to preserve the context needed to identify a rise in the crowd and applause after the event has ended.

➤ Audio Pre-Processing

Every segment is transformed into a Mel-Frequency Cepstral Coefficient (MFCC) spectrogram which represents both frequency and temporal evolution [11]. The Librosa

library transforms the waveform into a matrix, and displays this matrix as a picture, as shown in Figure 3.5 and in Figure 3.6. This is a wonderful way to convert an acoustic problem into a visual one, and use powerful computer-vision models. This spectrogram images were saved in separate library. The images are resized to 256×256 size and stored in different directories of two classes. This structure of the directory is similar to standard image classification datasets, making it easier to train downstream.

The proposed transfer learning based ResNet-50 Classification Framework is shown as 1.2.3 in the figure below.

The proposed framework uses the deep residual convolutional neural network with 50 layers (ResNet50) for classification of audio spectrogram images to belong to one of

the event classes. ResNet-50 is renowned for its impressive capability to extract hierarchical visual features with its residual learning mechanism, which is able to overcome the vanishing gradient problem in very deep networks. This is because it is a property that makes it particularly suitable for.

Complex feature learning tasks, like audio-visual event classification. In this study, transfer learning is chosen to improve the generalization of the model and to cut the training time. The first layers use a pretrained ResNet-50 model (trained on ImageNet), which is used as a fixed feature extractor. The main part of these blocks are frozen convolutional layers, which extract general low-level features like edges and textures, while the last layers are fine-tuned to learn dataset-specific high-level features. A custom classification head is added on, which is made up of:

- A Global Average Pooling (GAP) layer to reduce the spatial dimensions, and
- Add two Dense layers using ReLU activation functions to add non-linearity, The dropout layers (rate = 0.4, 0.5) for regularization, and A final Softmax layer for binary classification (event / non-event).

The model is built with ADAM optimizer with learning rate = 0.001 and categorical cross-entropy as the loss function. The training is done in mini batch format, where number of epochs and mini batch size is chosen empirically to optimise the performance. For stable convergence, the early stopping and learning rate reduction are included to prevent overfitting. Detailed procedure is below described.

The Applied Transfer learning on ResNet-50 is implemented in Algorithm 1. Algorithm 1: Applied Transfer learning on ResNet-50.

Set the output tensor of the ResNet-50 model to x. Add a Global Average Pooling layer on top of x. Add a Global Average Pooling layer to x.

- STEP 2: Add a Dense layer with 128 units and ReLU activation to x.

Consequently, an additional Dropout layer (with a dropout rate of 0.5) is added to x.

- STEP 4: Feed a Dense layer of 64 units and ReLU to x.

Repeat the process with another Dropout layer (0.5 dropout rate) added to x.

- STEP 6: Add the last Dense layer - 2 units - Softmax as predictions.
- STEP 7: Set the final model where the inputs are resnet_50.input and the output are predictions.

This extended ResNet-50 model is then saved for prediction and testing, to efficiently evaluate on unseen spectrogram samples.

IV. CONCLUSION

With the changing nature of choreography in the world, where more and more the Indian classical and western style are being mixed, the need for computational tools to capture the subtleties of stylistic hybridity also increases. To address this need, this research has tried to establish a conceptual framework of “Multimodal Deep Learning” that is designed to automatically learn and classify the fusion patterns between these cultural traditions. The proposed system combines the advantages of the three components of skeletal motion, rhythmic audio attributes, and semantic gesture annotations, and is designed into an explainable AI structure, thus showing a potential way to computational crosscultural dance analysis.

This work's major contribution is the development of an automated detection of stylistic convergence, which is said to be a gap in dance informatics. While most previous research has focused on classification as part of a single style paradigm, the continuum between traditional and contemporary movement styles is directly modeled in this paradigm. The use of features like attention-guided fusion scoring and an interpretable Fusion Index not only ensures the accuracy of the model's classifications but also their transparency and auditability, reflecting the need for culturally sensitive and ethical AI in the arts. Although it is still a theoretical conceptual model, it is founded on existing data analysis and architectural principles and provides a foundation for empirical implementation in the future. The results suggest that the need for a well-controlled, multi-modal solution to enhance technical recognition performance and to preserve the cultural integrity of the data of the movement being recognised.

This work may be used as a springboard for more research to create this conceptual model as a practical tool. The next steps are urgent, and involve curating a large-scale annotated dataset of cross-cultural dance fusion to be used for strong training and validation. Future projects will involve the development of realtime, interactive systems that can be used in live performance, pedagogical and digital archiving environments and that provide novel tools to choreographers, teachers and scholars to document, analyse and innovate in the complex and constantly changing landscape of global dance.

REFERENCES

- [1]. Singh, K., Yadav, M., Singh, Y., & Barak, D. (2024). Finding security gaps and vulnerabilities in IoT devices. In *Revolutionizing Automated Waste Treatment Systems: IoT and Bioelectronics* (pp. 379-395). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-6016-3.ch023>
- [2]. Hajimahmud, V. A., Singh, Y., & Yadav, M. (2024). Using a smart trash can sensor for trash disposal. In *Revolutionizing Automated Waste Treatment Systems: IoT and Bioelectronics* (pp. 311-319). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-6016-3.ch020>

- [3]. Singh, K., Yadav, M., & Yadav, R. K. (2024). IoT-Based automated dust bins and improved waste optimization techniques for smart City. In *Revolutionizing Automated Waste Treatment Systems: IoT and Bioelectronics* (pp. 167-194). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-6016-3.ch012>
- [4]. Yadav, M., Hajimahmud, V. A., Singh, K., & Singh, Y. (2024). Convert waste into energy using a low capacity igniter. In *Revolutionizing Automated Waste Treatment Systems: IoT and Bioelectronics* (pp. 301-310). IGI Global. <https://doi.org/10.4018/979-8-3693-6016-3.ch019>
- [5]. Khwaldeh, S., Mohit, Y., & Khushwant, S. (2024, May). Defensive auto-updatable and adaptable bot recommender system (DAABRS): a new architecture approach in cloud computing systems. In *2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-6). IEEE. <https://doi.org/10.1109/HORA61326.2024.10550519>
- [6]. Singh, K., Yadav, M., Singh, Y., & Moreira, F. (2025). Techniques in reliability of internet of things (IoT). *Procedia Computer Science*, 256, 55-62. <https://doi.org/10.1016/j.procs.2025.02.095>
- [7]. Singh, K., Yadav, M., Barak, D., Bansal, S., & Moreira, F. (2025). Machine-Learning-Based Frameworks for Reliable and Sustainable Crop Forecasting. *Sustainability*, 17(10), 4711. <https://doi.org/10.3390/su17104711>
- [8]. Singh, K., & Yadav, M. (2025). Prognosis of artificial intelligence in education. *LatIA*, 3, 107-107. <https://doi.org/10.62486/latia2025107>
- [9]. Yadav, M., Singh, K., Thukral, K., Kwatra, S., & Barak, D. (2025). Intelligent Electronic Ticketing Platform in Smart Transportation Ecosystem. In *Driving Green Transportation System Through Artificial Intelligence and Automation: Approaches, Technologies and Applications* (pp. 581-601). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72617-0_30
- [10]. Singh, K., Yadav, M., & Abdullayev, V. H. (2024). Prediction of Flight Areas using Machine Learning Algorithm. *LatIA*, (2), 13. <https://doi.org/10.62486/latia202493>
- [11]. Singh, K., Yadav, M., Singh, Y., Malik, P. S., Siwach, V., Khurana, D., ... & Elngar, A. A. (2024). IoT Networks: Integrated Learning for Privacy-Preserving Machine Learning. In *Artificial Intelligence Using Federated Learning* (pp. 250-275). CRC Press. https://doi.org/10.1201/978100348200013?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle
- [12]. Singh, K., Yadav, M., Yadav, R. K., Naib, B. B., Bhatia, S., & Panda, D. K. (2024, December). Prognosis of Relocation Disease in Animals using Aggregation Method with Optimization Techniques. In *2024 Eighth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (pp. 15-20). IEEE.
- [13]. Khang, A., Singh, K., & Yadav, M. (2026). Anomaly Detection Driven by Machine Learning: Enhancing SIEM Tools for Sturdy Cyber Defense. In *AI-Powered Cybersecurity for Banking and Finance* (pp. 338-361). Productivity Press. https://doi.org/10.4324/978100359687514?urlappend=%3Futm_source%3Dresearchgate.net%26utm_medium%3Darticle
- [14]. Ngoc, A. P. T., Abdullayev, V., Khang, A., Ragimova, N., & Yadav, M. (2026). Identify and classify brain tumors using deep learning techniques and magnetic resonance imaging. In *Revolutionizing Digital Healthcare Through Artificial Intelligence and Automation* (pp. 333- 346). Academic Press. <https://doi.org/10.1016/C2024-0-02016-18>
- [15]. Ngoc, A. P. T., Singh, K., & Yadav, M. (2026). New neural network methods for the detection of heart diseases. *Revolutionizing Digital Healthcare Through Artificial Intelligence and Automation*, 289-298. <https://doi.org/10.1016/B978-0-443-36434-1.00034-3>