[1] **Bijal U. Gadhia**

[2] **Shahid S. Modasiya**

LSTM-Based Soccer Video Summarization Via Event Classification for Highlighting Key Moments

Abstract: - Video summarization creates brief synopses of video content by selecting the most informative segments, either as key-frames or key-fragments. This study introduces an advanced method for event classification in soccer videos using a modified stacked Long Short-Term Memory (LSTM) model. By utilizing the Soccer Action Detection Compilation (SADC) dataset, which includes detailed annotations of football events, our method combines VGG16 for feature extraction with LSTM for event classification. The model efficiently identifies crucial events such as goals, goal attempts, and yellow cards, while also filtering out non-essential segments labelled as "No Events." When compared to the Bi-LSTM model, the modified LSTM demonstrates superior performance in terms of precision, recall, and F1-score for several key event classes. Specifically, at epoch 150, the modified LSTM achieves an F1-score of 0.87 for "No Event" segments and perfect precision for "Goal" events, surpassing the Bi-LSTM. These findings underscore the model's stability and accuracy, which ensure the production of high-quality highlight reels by emphasizing important events and minimizing irrelevant content. In the future, by removing non-essential segments recognized as "No Events" from any football match, one can generate perfect highlight reels that capture all crucial moments, providing viewers with a more focused and enjoyable experience.

Keywords: LSTM, Deep Learning, Video summarization, Event Classification, Highlight

1. Introduction:

Video summarization aims to generate a short synopsis that summarizes the video content by selecting its most informative and important parts [1, 2]. The produced summary is usually composed of a set of representative video frames (key-frames), or video fragments (key-fragments) that have been organized in chronological order to form a shorter video [2]. The former type of a video summary is known as video storyboard, and the latter type is known as video skim[3]. One advantage of video skims over static sets of frames is the ability to include audio and motion elements that offer a more natural story narration and potentially enhance the expressiveness and the amount of information conveyed by the video summary [4]. Furthermore, it is often more entertaining and interesting for the viewer to watch a skim rather than a slide show of frames [5].

From surveillance videos to sports matches, the need to efficiently skim through vast amounts of visual data has become increasingly prevalent across various domains. In the realm of video surveillance, the ability to distill hours of footage into succinct summaries facilitates efficient monitoring, anomaly detection, and forensic analysis [4-6]. Similarly, in the context of sports, where every moment holds significance, the ability to extract key events and highlights from matches is crucial for coaches, analysts, and fans [7]. Machine learning techniques have emerged as indispensable tools for automating the process of video summarization, leveraging algorithms to identify salient features, events, and patterns within the visual data [8, 9]. Deep learning, in particular, has revolutionized video analysis by enabling the development of sophisticated models capable of learning intricate representations directly from raw pixel data [10]. Within this background, Long Short-Term Memory (LSTM) networks have emerged as a powerful deep learning architecture for sequential data processing, offering the ability to capture temporal dependencies and contextual nuances within video sequences. Researchers have increasingly turned to LSTM models to enhance the effectiveness and efficiency of video summarization techniques, leveraging their ability to model long-range dependencies and extract meaningful insights from sequential data streams [11-14]. By integrating LSTM networks into video summarization pipelines, researchers aim to unlock new levels of automation, accuracy, and adaptability in summarizing diverse video content across a wide range of applications [15, 16]. Deep learning techniques excel in classification tasks when supported by a carefully crafted training dataset, surpassing alternative methodologies in effectiveness [23].

This paper presents our novel approach, a modified stacked LSTM model, for event classification in soccer test videos. We trained our model using the Soccer Action Compilation Dataset (SADC), incorporating VGG16 for feature extraction and LSTM for event classification. Our model effectively identifies crucial events such as goals, goal attempts, and yellow cards, while also distinguishing unimportant segments as "No Events." The Proposed Model section provides a comprehensive overview of our methodology. The Results and Discussion section offers a comparative analysis of our model's performance, juxtaposed with the Bi-LSTM model referenced in [18].

2. Literature Review:

[1] Research scholar, Gujarat Technological University, Ahmedabad, bij.1988@gmail.com

[2] Assisstant Professor, Electronics and commnication Department, GEC Gandhinagar, shahid@gecg28.ac.in

Traditional approaches to video summarization have focused on key frame selection using motion analysis, aiming to enhance the understanding of video content by presenting a series of frames that summarize the intended video [2-5]. Recent advancements in video summarization have leveraged deep learning techniques, particularly deep neural networks, to extract and represent semantic information more effectively [6].In the domain of sports video analysis, video summarization plays a pivotal role in capturing key moments and highlights from extensive footage [7]. Below is a summary of recent review work focusing on sports video summarization, developed using machine learning or deep learning technique. In [8], Huawei Wei et al. has focused on semantic information and long-term temporal semantics which are crucial for capturing the essence of the video content. As, they introduced method which selects finest video shots by minimizing the gap between the description of the summarized video and original video text annotated by human. In [9], Shruti Jadon et al. combined deep learning techniques with clustering to extract key frames. Leveraging deep learning approaches, such as Convolutional Neural Networks (CNNs) or by applying transfer learning, researchers have developed sophisticated models for summarizing soccer match videos by recognizing and highlighting relevant actions, as demonstrated in [10] and [11]. By integrating deep learning models, such as 3D-CNNs and LSTM networks, researchers have achieved remarkable progress in summarizing complex videos, including soccer match footage[12].These models leverage the spatiotemporal learning capabilities of deep neural networks to identify and summarize key events with high precision and recall [13]. Furthermore, the integration of multimodal information, such as audio and video data, has emerged as a promising approach to enhancing the performance of video summarization systems just like in [14] Vanderplaetse et al. have achieved significant improvements in event detection and classification tasks, particularly in sports by combining audio features with visual cues extracted from videos. Advancements in video summarization methodologies have also addressed the challenges associated with event detection and localization in long football (soccer) videos Moreover, the utilization of hierarchical recurrent networks has enabled more robust understanding of team sports activities, leading to advancements in action recognition and summarization [15]. Long-range dependencies between video frames and correlations between mid-range and short-range are crucial for accurate event localization. To address this, Behzad et al. [16] proposed novel architectures, such as Dilated Recurrent Neural Networks (DilatedRNN), which leverage both long short-term memory (LSTM) units and two-stream convolutional neural network (Two-stream CNN) features. Soccer analytics present unique challenges due to the complex nature of the game, including the large pitch, numerous players, and sparse scoring events. Deep Reinforcement Learning (DRL) models have emerged as effective tools for comprehensively evaluating soccer actions. Guiliang et al. [17] have developed a model capable of fitting continuous game context signals and sequential features, enabling accurate assessment of player performance and overall team dynamics by learning action-value Q-functions using stacked LSTM towers. Addressing the issue of redundancy among key frames in user-created videos, recent research has proposed Graph Attention Networks (GAT) adjusted Bi-directional Long Short-term Memory (Bi-LSTM) models for unsupervised video summarization. By integrating GAT to transform visual features and Bi-LSTM to refine key frame selection, these models achieve superior performance in generating concise and representative video summaries [18]. The proposed techniques outperform state-of-the-art methods by effectively capturing salient areas and semantic features from video frames. Additionally, automated video skimming techniques offer efficient browsing and searching mechanisms for long videos. Alam et al. [19] introduced a video skimming approach which used unsupervised methods to extract motion-based features and shot boundary detection to extract important clips and generate concise summaries. By employing clustering and attention mechanisms, they have demonstrated the effectiveness of these approaches in summarizing diverse video content, as evidenced by superior performance on benchmark datasets represented in [20], [21]. Currently, research has concentrated on automatic sports summarization to highlight key actions and capture the full match narrative with an emotional impact similar to that produced by a human editor. In [22], Sanabria et al. proposed a method to detect match actions and determine which actions should be included in the summary, and generates multiple candidate summaries with relevant variability for final editing by introducing LSTM MIL pooling data.

In summary, the literature on video summarization highlights a shift towards using advanced deep learning techniques, particularly for sports video analysis. Methods now focus on semantic extraction, long-term temporal semantics, and multimodal information integration. Key advancements include the use of Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, Hierarchical LSTM and Bi-LSTM. These approaches significantly improve event detection, frame selection, and overall summarization performance, offering robust solutions for generating concise, representative video summaries in various domains.

3. Proposed Method:

In this section, we introduce an enhanced method for predicting labels from video data. We combine the strong features of the VGG16 model with the temporal modelling capabilities of our proposed stacked LSTM model. Our approach revolves around three key components: 1. Frame extraction, 2. Feature extraction, and 3. Model architecture. By leveraging the strengths of VGG16 and LSTM, we aim to enhance accuracy and robustness in label predictions. Our model distinctly classifies events such as goals, goal attempts, yellow cards, or penalty, while also identifying unimportant segments labelled as "No event." Ultimately, to generate a concise skimming or highlight reel, one can exclude all unimportant segments like "No events."

The Detailed block diagram is shown as per Figure no. 1.

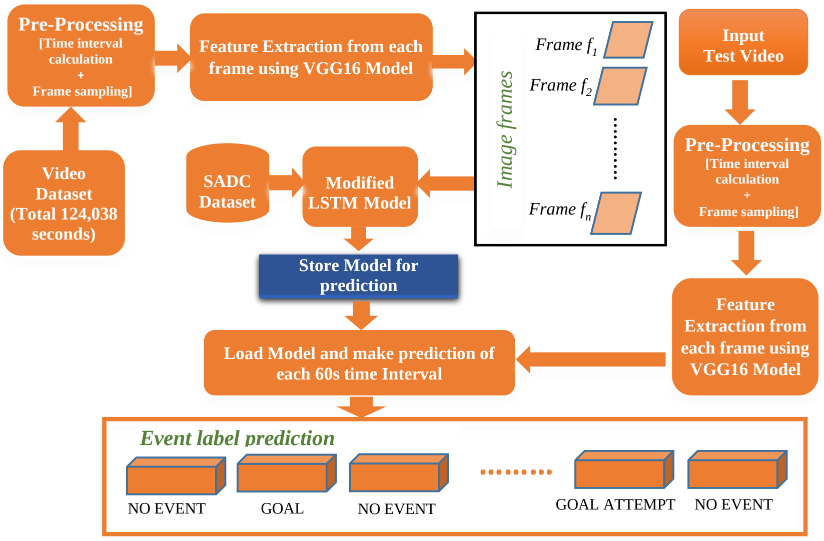

Pre-Processing
[Time interval calculation + Frame sampling]
Feature Extraction from each frame using VGG16 Model
Image frames
Frame f_1
Frame f_2
Frame f_n
Input Test Video
Video Dataset (Total 124,038 seconds)
SADC Dataset
Modified LSTM Model
Pre-Processing
[Time interval calculation + Frame sampling]
Store Model for prediction
Feature Extraction from each frame using VGG16 Model
Load Model and make prediction of each 60s time Interval
Event label prediction
NO EVENT
GOAL
NO EVENT
GOAL ATTEMPT
NO EVENT

Figure 1. Block Diagram of Proposed Model Architecture

3.1 Frame Extraction:

In the initial stage of our proposed method, extracting frames from the input video is crucial. This task is accomplished using the "extract_event_frames" function, which efficiently handles the process. Given parameters such as the video file path, start and end times, and a desired frame rate, this function extracts frames at regular intervals within the specified time range. It utilizes the OpenCV library to seamlessly and effectively read the video file and extract frames. The extracted frames are then saved as individual image files for further processing. The detailed procedure for this function is outlined below in algorithm-1.

```
Algorithm-1: Frame extraction of test input video
Input: Video dataset
Output: Image frames of video dataset
  1. Open video file at specified video_path
  2. Get total number of frames in the video
  3. Calculate start and end frame indices based on start and end times
  4. Calculate frame interval based on desired frame rate
  5. Initialize frame counter and frame index
  6. Set current frame position in the video to start frame
  7. Initialize empty list for image paths
  8. While frame counter <= end frame and frame counter < total frames:
     1. Read next frame from video
     2. If frame was read successfully:
        1. If current frame index is a multiple of frame interval:
           1. Save frame to output path with appropriate index
           2. Append frame output path to image list
           3. Increment frame index
     3. Increment frame counter
  9. Release video object
 10. Return image list
```

3.2 Feature Extraction:

In the feature extraction stage, we utilize a pre-trained convolutional neural network, specifically the VGG16 model, to extract visual features from the frames. The VGG16 model, initially trained on the ImageNet dataset[24], is employed to capture high-level representations of the visual content in each frame. To adapt the VGG16 model for our task, we remove its top classification layer, enabling us to obtain feature vectors directly from the penultimate layer. These feature vectors encapsulate discriminative information about the content of each frame and serve as input to the LSTM-based classification model. Leveraging these features, the model predicts the most probable event label (e.g., Goal, Goal Attempt, Penalty, Yellow Card) along with unimportant events labelled as “No Event” for each video segment.

3.3 Model Architecture:

Our proposed model architecture uses Long Short-Term Memory (LSTM) networks, known for their ability to capture temporal dependencies in sequential data [12]. The model architecture comprises multiple stacked LSTM layers followed by dropout and batch normalization layers to enhance generalization and mitigate overfitting. Additionally, fully connected layers with activation functions are employed for final classification. For training the model, we use the SADC dataset. Detailed Explanation of this dataset is given in Section-4 and its representation is as mentioned in Table 1. This SADC dataset is split into training and testing sets using a standard train-test split approach. The Adam optimizer is utilized with a learning rate of 0.001, and the categorical cross-entropy loss function is chosen for its suitability in multi class classification tasks. The model is trained for a various epochs with a validation split to monitor performance and prevent over fitting. Once the model is trained it is being saved for Prediction. The detailed of Modified deep stack LSTM Model is shown as per Table 2:

Table 1: Recorded event of SADC

Event name	Start time (Sec.)	End time (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

Below Table Shows LSTM Model Architecture:

Table 2: LSTM Model Architecture

Start:		
Istm_input	input:	[(None, 80, 4096)]
InputLayer	output:	[(None, 80, 4096)]
Istm	input:	(None, 80, 4096)
LSTM	output:	(None, 80, 128)
Dropout	input:	(None, 80, 128)
Dropout	output:	(None, 80, 128)
batch_normalization	input:	(None, 80, 128)
BatchNormalization	output:	(None, 80, 128)
Istm_1	input:	(None, 80, 128)
LSTM	output:	(None, 80, 128)
batch_normalization_1	input:	(None, 80, 64)
BatchNormalization	output:	(None, 80, 64)
Istm_2	input:	(None, 80, 64)
LSTM	output:	(None, 80, 32)
dropout 1	input:	(None, 80, 32)
Dropout	output:	(None, 80, 32)
batch_normalization_2	input:	(None, 80, 32)
BatchNormalization	output:	(None, 80, 32)
Istm_3	input:	(None, 80, 32)
LSTM	output:	(None, 80, 16)
batch_normalization_3	input:	(None, 80, 16)
BatchNormalization	output:	(None, 80, 16)
Istm_4	input:	(None, 80, 16)
LSTM	output:	(None, 80, 8)
batch_normalization_4	input:	(None, 80, 8)
BatchNormalization	output:	(None, 80, 8)
Istm_5	input:	(None, 80, 8)
LSTM	output:	(None, 80, 4)
dropout 2	input:	(None, 80, 4)
Dropout	output:	(None, 80, 4)
batch_normalization_5	input:	(None, 80, 4)
BatchNormalization	output:	(None, 80, 4)
Istm_6	input:	(None, 80, 4)
LSTM	output:	(None, 2)
batch_normalization_4	input:	(None, 2)
BatchNormalization	output:	(None, 2)
End:		

4. Results and discussion:

In this section, we detail the experimental methodologies and datasets which used to develop and evaluate our proposed system. Our experiments were conducted in the Google Colab environment and implementation is written in Python. Data pre-processing computer vision tasks, is performed using optical flow and certain data processing operation and for the deep learning components of our experiments, we used TensorFlow explained in [25].

4.1 SADC Dataset:

We use Soccer Action Detection Compilation (SADC) for training the model. SADC consist a collection of 25 football video clips, sourced from YouTube. These clips collectively represent an extensive duration of 124,038 seconds during which a team of five football enthusiasts / player dedicatedly analysed each video frame by frame. Their analysis involved identifying and documenting various game-related events, including goals, penalty kicks, free kicks, penalty corners, and yellow cards. Each event's start and end times were specifically recorded in a structured .csv file, as shown in format of Table 1. The dataset provides accuracy not only in its duration but also in its detailed annotations, in labels of important football events which provide valuable insights into the dynamics of football matches. This carefully compiled dataset serves as the foundation for training our deep learning modified stacked Long Short-Term Memory (LSTM) model.

4.2 Event Prediction:

The proposed methodology outlined in Section 3 was implemented and trained over various epochs like 100, 150, 200 and 250. Throughout the duration of each video, game events and periods of inactivity ("NO EVENT") were classified at 60-second intervals. The primary goal of identifying "No events" is to filter out unimportant moments from a football game. By isolating these non-essential segments, we can merge all significant events to create a high-quality highlight reel, which will be valuable for football enthusiasts. Successfully recognizing "No events" ensures that the highlights focus only on key moments, resulting in a more engaging and informative summary of the game.

Subsequently, the algorithm's predicted events were matched with the manually observed events recorded by the football enthusiasts as shown in Table 3. The resulting outcomes were then collated and analysed to provide a comprehensive evaluation, as detailed in the Event Prediction Evaluation table presented below. Our proposed Modified LSTM model is compared with Bi-LSTM model discussed in [18].

Table 3. Event Prediction Table

Observed Event	**Predicted Event**	**Start time (sec.)**	**End time (sec.)**	**Prediction Outcome**	**Class Label**
NO EVENT	NO EVENT	0	60	MATCH	TN
GOAL ATTEMPT	GOAL ATTEMPT	61	120	MATCH	TP
GOAL	NO EVENT	121	180	MATCH	FN
NO EVENT	GOAL	181	240	NO MATCH	FP
………	………	……	……	……..	…..
YELLOW CARD	GOAL ATTEMPT	5400	5460	NO MATCH	FP

Below Tables 4 and 5 present the performance metrics of the Modified LSTM model in comparison with the Bi-LSTM model. The following are key observations derived from these performance tables. The Modified LSTM model demonstrates higher precision in predicting "GOAL" events at epoch 150, making it more reliable than the Bi-LSTM model. It’s slightly higher F1-score indicates better overall performance for this class. In classifying "GOAL ATTEMPT" events, the Modified LSTM surpasses the Bi-LSTM model. At epoch 150, it achieves an F1-score of 0.64 compared to the Bi-LSTM's 0.48, showing a better balance between precision and recall. For "YELLOW CARD" events, the Modified LSTM performs exceptionally well at epoch 150, achieving perfect scores for precision, recall, and F1 (1.00). In contrast, the Bi-LSTM model's performance drops significantly after epoch 100.

Table 4. Performance Metrics of Modified LSTM Model

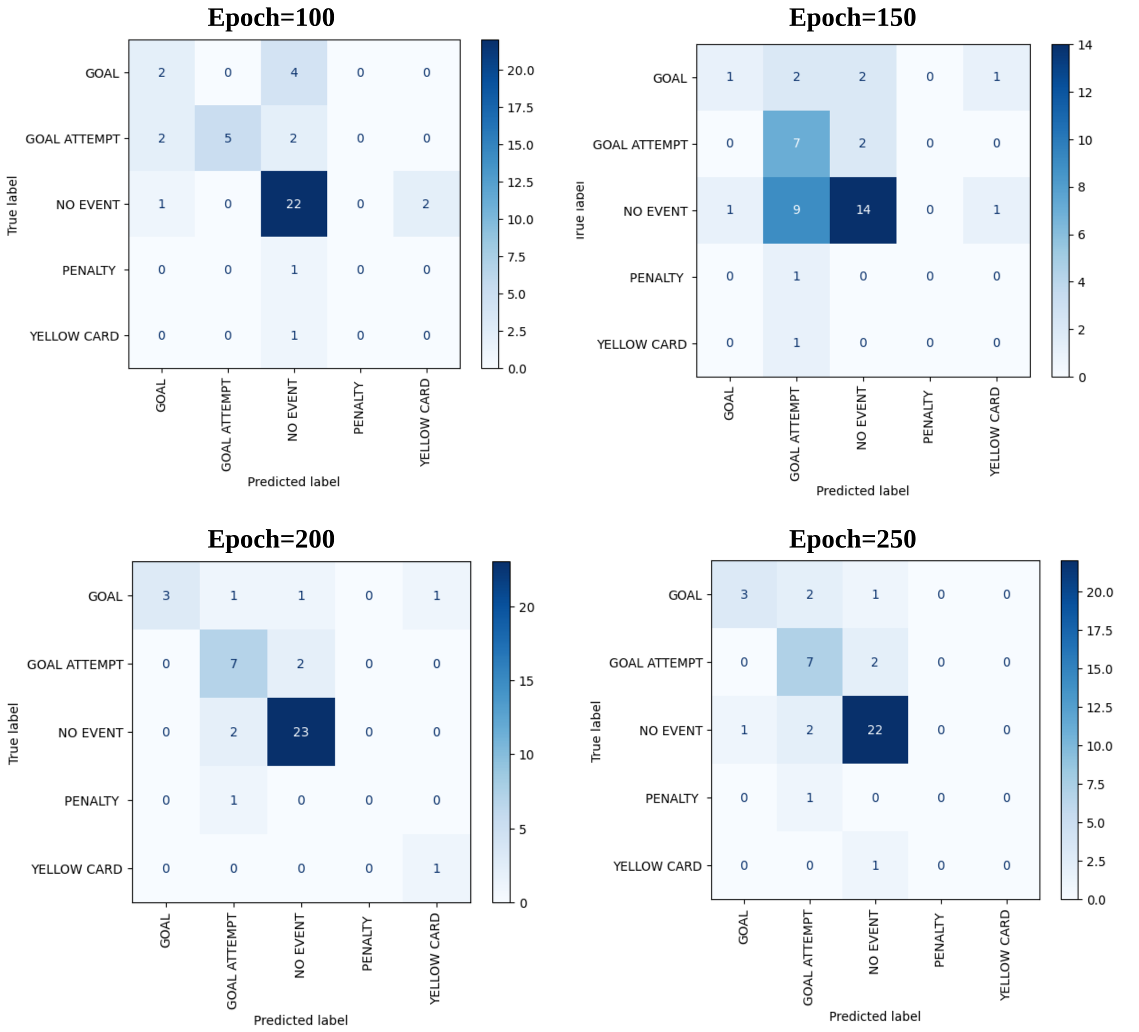
Modified LSTM Model		**Precision**	**Recall**	**F1-Score**
EPOCH=100	**NO EVENT**	**0.81**	**0.84**	**0.82**
	GOAL	0.67	0.33	0.44
	GOAL ATTEMPT	0.55	0.67	0.60
	YELLOW CARD	0.50	1.00	0.67
	PENALTY CORNER	0.00	0.00	0.00
EPOCH=150	**NO EVENT**	**0.82**	**0.92**	**0.87**
	GOAL	1.00	0.17	0.29
	GOAL ATTEMPT	0.54	0.78	0.64
	YELLOW CARD	1.00	1.00	1.00
	PENALTY CORNER	0.00	0.00	0.00
EPOCH=200	**NO EVENT**	**0.80**	**0.80**	**0.80**
	GOAL	0.50	0.33	0.40
	GOAL ATTEMPT	0.56	0.56	0.56

	YELLOW CARD	0.00	0.00	0.00
	PENALTY CORNER	0.00	0.00	0.00

Table 5. Performance Metrics of Bi-LSTM Model

Bi-LSTM Model		**Precision**	**Recall**	**F1-Score**
EPOCH=100	**NO EVENT**	**0.74**	**0.87**	**0.76**
	GOAL	0.50	0.33	0.40
	GOAL ATTEMPT	0.67	0.44	0.53
	YELLOW CARD	1.00	1.00	1.00
	PENALTY CORNER	0.00	0.00	0.00
EPOCH=150	**NO EVENT**	**0.78**	**0.56**	**0.65**
	GOAL	0.50	0.17	0.25
	GOAL ATTEMPT	0.35	0.78	0.48
	YELLOW CARD	0.00	0.00	0.00
	PENALTY CORNER	0.00	0.00	0.00
EPOCH=200	**NO EVENT**	**0.73**	**0.88**	**0.80**
	GOAL	0.33	0.17	0.22
	GOAL ATTEMPT	0.62	0.56	0.59
	YELLOW CARD	0.00	0.00	0.00
	PENALTY CORNER	0.00	0.00	0.00

Overall, the Modified LSTM consistently achieves higher precision, recall, and F1-scores for several key event classes compared to the Bi-LSTM model. For instance, at epoch 150, the Modified LSTM's "NO EVENT" class has an F1-score of 0.87, outperforming the Bi-LSTM's 0.65 at the same epoch

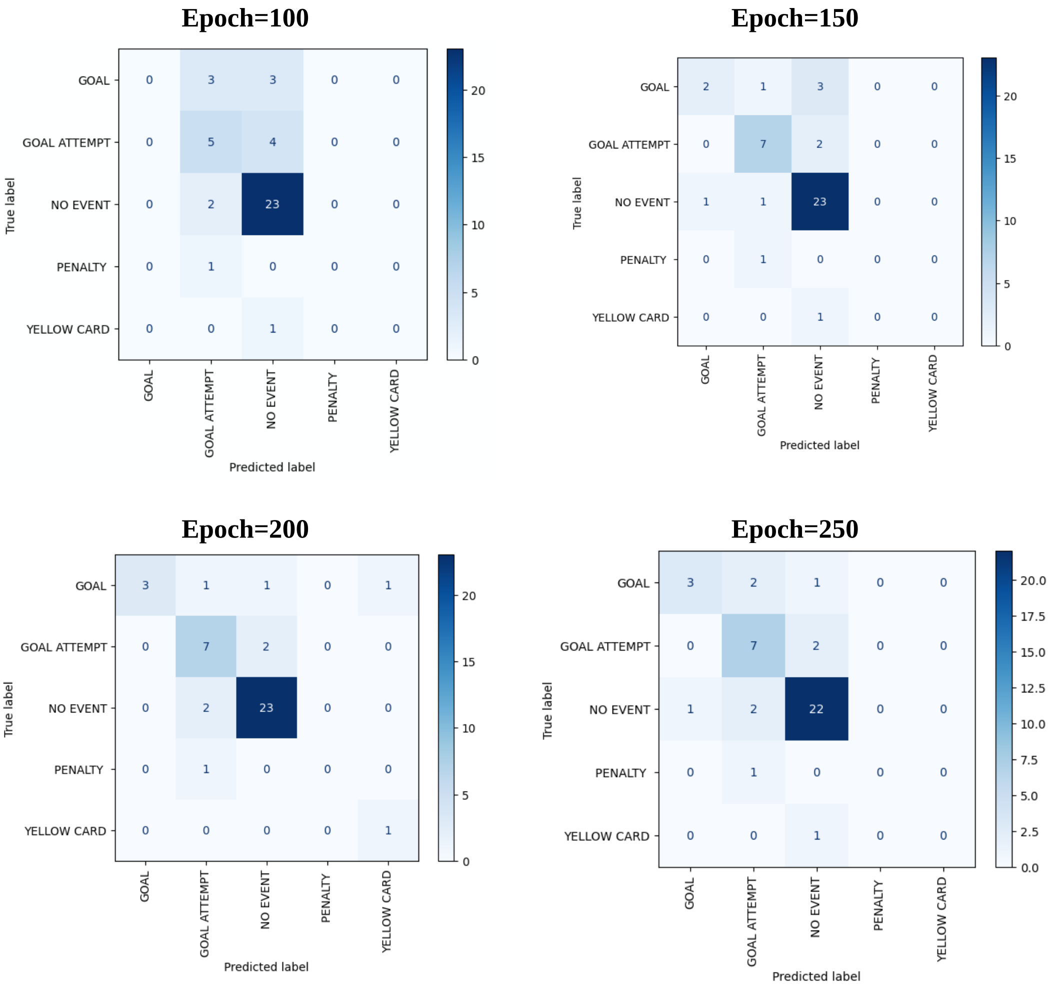

Epoch=100

Epoch=150

Epoch=200

Epoch=250

Figure 2: Confusion matrix of Bi-LSTM for various epoch


Epoch=100

Epoch=150

Epoch=200

Epoch=250

Figure 3: Confusion matrix of Modified LSTM for various epochs

From the figure depicting the accuracy comparison of modified LSTM with Bi-LSTM. Modified LSTM is more effective at learning and generalizing from the data, making it a more reliable choice for tasks requiring high accuracy in event classification.

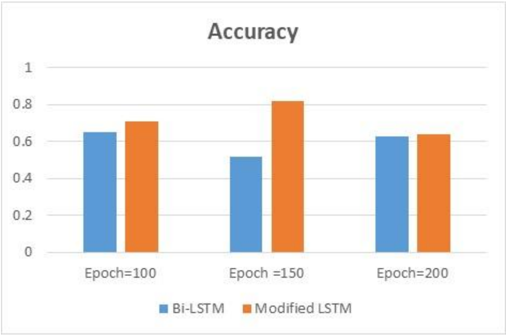


Figure 4: Accuracy comparison of Modified LSTM with Bi-LSTM

5. Conclusion:

Our experimental results highlight the superior performance of the Modified LSTM model, particularly in classifying both "GOAL" and "NO EVENT" segments, crucial for generating high-quality highlight reels. The model consistently demonstrates remarkable stability and accuracy in identifying "NO EVENT" segments, achieving its highest F1-score of 0.87 at epoch 150, outperforming the Bi-LSTM model's peak F1-score of 0.80 at epoch 200. This consistent performance ensures effective filtration of unimportant segments, enhancing the relevance and engagement of the highlight reel.

For "GOAL" events, the Modified LSTM model maintains higher precision, especially at epoch 150, where it achieves perfect precision (1.00). Although the recall for "GOAL" events is lower, the high precision indicates reliable identification of true goal events, minimizing false positives. This is crucial for creating concise and accurate highlights that capture the most critical moments of the game.

In conclusion, the Modified LSTM model's ability to accurately classify both significant events like goals and unimportant segments like "NO EVENTS" significantly improves the quality of generated highlight reels. This capability is essential for providing football enthusiasts with engaging and informative summaries, thereby enhancing their viewing experience. The results validate the effectiveness of our approach in leveraging deep learning techniques and curated datasets to advance sports video analysis, promising better automated summarization solutions across various domains.

References:

[1] Y. Li, T. Zhang, and D. Tretter, “An overview of video abstraction techniques,” Hewlett Packard, Technical Reports, 01 2001.

[2] J. Calic, D. P. Gibson, and N. W. Campbell, “Efficient Layout of Comic-Like Video Summaries,” IEEE Trans. on Circuits and Systems for Video Technology, vol. 17, no. 7, pp. 931–936, July 2007.

[3] T. Wang, T. Mei, X. Hua, X. Liu, and H. Zhou, “Video Collage: A Novel Presentation of Video Sequence,” in 2007 IEEE Int. Conf. on Multimedia and Expo, July 2007, pp. 1479–1482

[4] Arthur G. Money and Harry Agius. 2008. “Video summarisation: A conceptual framework and survey of the state of the art”. J. Vis. Comun. Image Represent. 19, 2 ,pp. 121–143 February, 2008.

[5] Vivekraj V. K., Debashis Sen, and Balasubramanian Raman. 2019. “Video Skimming: Taxonomy and Comprehensive Survey”. ACM Comput. Surv. 52, 5, Article 106 , October 2019.

[6] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris and I. Patras, "Video Summarization Using Deep Neural Networks: A Survey," in Proceedings of the IEEE, vol. 109, no. 11, pp. 1838-1863, Nov. 2021

[7] Mendi, Engin & Clemente, Hélio & Bayrak, Coskun. “Sports video summarization based on motion analysis. Computers & Electrical Engineering”. 39. 790–796, 2013.

[8] Wei, H., Ni, B., Yan, Y., Yu, H., Yang, X., & Yao, C. (2018). Video Summarization via Semantic Attended Networks. AAAI Conference on Artificial Intelligence.

[9] S. Jadon and M. Jasim, "Unsupervised video summarization framework using key frame extraction and video skimming," 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2020, pp. 140-145.

[10] Gaikwad, D., Sarap, S., & Dhande, D. Y. (2022). Video Summarization Using Deep Learning for Cricket Highlights Generation. Journal of Scientific Research, 14(2), 533–544.

[11] Rafiq, Muhammad & Rafiq, Ghazala & Agyeman, Rockson & Jin, Seong-Il & Choi, Gyu Sang. (2020). Scene Classification for Sports Video Summarization Using Transfer Learning. Sensors. 20. 1702.

[12] R. Agyeman, R. Muhammad and G. S. Choi, "Soccer Video Summarization Using Deep Learning," 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019, pp. 270-273.

[13] Olav A. Nergard Rongved, Steven A. Hicks, Vajira Tham- bawita, Hakon K. Stensland, Evi Zouganeli, Dag Johansen, Michael A. Riegler, and Pal Halvorsen. “Real-time detection of events in soccer videos using 3D convolutional neural networks”. In IEEE International Symposium on Multimedia (ISM), December 2020.

[14] B. Vanderplaetse and S. Dupont, "Improved Soccer Action Spotting using both Audio and Video Streams," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 3921-3931.

[15] T. Tsunoda, Y. Komori, M. Matsugu and T. Harada, "Football Action Recognition Using Hierarchical LSTM," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 2017, pp. 155-163.

[16] Mahaseni, B., Faizal, E.R., & Raj, R.G. (2021). Spotting Football Events Using Two-Stream Convolutional Neural Network and Dilated Recurrent Neural Network. IEEE Access, 9, 61929-61942.

[17] Liu, G., Luo, Y., Schulte, O. et al. Deep soccer analytics: learning an action-value function for evaluating soccer players. Data Min Knowl Disc 34, 1531–1559 (2020).

[18] R. Zhong, R. Wang, Y. Zou, Z. Hong and M. Hu, "Graph Attention Networks Adjusted Bi-LSTM for Video Summarization," in IEEE Signal Processing Letters, vol. 28, pp. 663-667, 2021.

[19] Alam, I., Jalan, D., Shaw, P., & Mohanta, P.P. (2020). Motion Based Video Skimming. 2020 IEEE Calcutta Conference (CALCON), 407-411.

[20] Sanabria, M., Precioso, F., Mattei, P., & Menguy, T. (2022). A Multi-stage deep architecture for summary generation of soccer videos. ArXiv, abs/2205.00694.

[21] Song, Yale, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes,”TVSum: Summarizing web videos using titles,” In Proceedings of the IEEE Conference onComputer Vision and Pattern Recognition, 5179- 5187 (2015).

[22] Gygli, Michael and Grabner, Helmut and Riemenschneider, Hayko and Van Gool, Luc,”Creating Summaries from User Videos,” European conference on computer vision, Zurich, 505-520 (2014).

[23] Gadhia, Bijal & Modasiya, Shahid. (2023). An Evaluation-based Analysis of Video Summarising Methods for Diverse Domains. Journal of Innovative Image Processing. 5. 127-139.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ‘‘ImageNet: A large-scale hierarchical image database,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2009, pp. 248–255.

[25] Abadi et al. ‘‘TensorFlow: Large-scale machine learning on heterogeneous distributed systems,’’ 2016, arXiv: 1603.04467. [Online]. Available: http://arxiv.org/abs/1603.04467.