Event detection in soccer matches through audio classification using transfer learning

Bijal Utsav Gadhia[1], Shahid S. Modasiya[2]
[1]Department of Computer Engineering, Gujarat Technological University, Ahmedabad, India
[2]Department of Electronics and Communication, Government Engineering College, Gandhinagar, India

Article Info

Article history:

Received Mar 26, 2024
Revised Oct 31, 2024
Accepted Nov 14, 2024

Keywords:

Audio classification
Deep learning
ResNet-50
Soccer summarization
Transfer learning

ABSTRACT

Addressing the complexities of generating sports summaries through machine learning, our research aims to bridge the gap in audio-based event detection, particularly in soccer games. We introduce an extended ResNet-50 deep learning approach for soccer audio, emphasizing key moments from large soccer content archives through the use of transfer learning. The proposed model accurately classifies soccer audio segments into two categories: i) events, representing crucial in-game occurrences and ii) no events, denoting less impactful moments. The model involves complete audio preprocessing, the implementation of proposed model using transfer learning and the classification of events. The model's reliability is validated using the dataset soccer action dataset compilation (SADC), involves dataset creation by football fans. Comparative analysis with pre-trained models such as VGG19, DesNet121, and EfficientNetB7 demonstrates the superior performance of the extended ResNet-50 based approach. Results across different epochs reveal consistently high accuracy, precision, recall, and F1-score, emphasizing the proposed model's effectiveness in event detection through audio classification. The paper concludes that the proposed model offers a robust solution for detecting an event from audio of soccer sports providing valuable insights for fans, analysts, and content creators to identify interested moments from soccer game with low failure.

Corresponding Author:

Bijal Utsav Gadhia
Department of Computer Engineering, Gujarat Technological University
Ahmedabad, Gujarat, India
Email: bij.1988@gmail.com

1. INTRODUCTION

The expansion of multimedia content, including videos utilized for both entertainment and professional purposes, has experienced unprecedented growth in recent years [1], [2]. The commercial potential of automatic sports video summarization techniques has gathered significant attention, sparking interest in various approaches to address this aspect [3], [4]. Soccer, often referred to as the world's most popular sport, fascinates millions of fans worldwide with its thrilling matches and iconic moments [5]. In the age of digital media, the availability of vast archives of any sports content, including videos and audio recordings, has created a treasure trove of information waiting to be explored [6]. Soccer summarization, is a growing field at the intersection of sports analytics and artificial intelligence, activities to unlock the full potential of this rich multimedia data.

Khan and Pawar [7] reviews recent work on key frame-based and dynamic video summarization techniques, discussing challenges and future directions in the field of sports. Jadon and Jasim [8] attempted to address video summarization using an unsupervised learning paradigm which was achieved by applying conventional vision-based algorithms for precise feature extraction from video frames. Above methods for video summarization, including key frame-based and dynamic techniques, have made strides, they often lack the ability to efficiently differentiate between significant and less impactful moments in soccer [7], [8]. Soccer summaries are essential because they can reduce hours of video into concise and informative highlights. These highlights are not only valuable for fans seeking to relive the most exciting moments but also for analysts, coaches, and players striving to gain deeper insights into team strategies and player performance. Rongved *et al.* [9] introduces a 3D convolutional neural network (3D-CNN) algorithm for automated event detection in soccer videos. Pablos *et al.* [10] proposed 3D-CNN based deep neural network addressing the challenge of unedited user-generated kendo sport content. Emon *et al.* [11] suggested deep cricket summarization network (DCSN) approach to provide concise synopses of long cricket matches by using CNN long short-term memory (LSTM) approach. The proposed system, evaluated on the new cricsum dataset using mean opinion score (MOS). A few researchers have delved into audio processing to predict precise events in diverse domain.

Sound plays a pivotal role in capturing attention and can proficiently discern saliency to extract out important occurrences from video [12]–[15]. Sanabria *et al.* [12] devised an architectural framework that employs a multiple instance learning (MIL) approach to consider the sequential interdependence among events. Additionally, it incorporates a hierarchical multimodal attention layer with audio features designed to discern the significance of each event within an action context. Evangelopoulos *et al.* [13] has integrated audio feature through waveform modulation with visual to identify saliency from movie video streams and concluded that multimodal saliency producing subjectively high quality summaries. Vanderplaetse and Dupont [14] detailed an experimental investigation to explore the integration of audio and video information within various stages of deep neural network architectures. Ilse *et al.* [15] addresses MIL by formulating the problem as learning the Bernoulli distribution of bag labels using neural networks. It introduces an attention-based operator, providing insights into the contribution of each instance to label.

Deep learning techniques intricately extract feature representations [16]–[18]. Sanabria *et al.* [16] solely relied on the energy of the audio signal, which, in other contexts, have proven beneficial for enhancing classifications in soccer games. Agyeman *et al.* [17] presented deep learning for summarizing lengthy soccer videos, utilizing a 3D-CNN and LSTM recurrent neural network (RNN). Ji *et al.* [18] proposed a deep learning framework for video summarization, which uses GoogleNet with BiLSTM framework to address challenges of relation discovery and semantic loss by integrating encoder-decoder attention and semantic preserving loss.

A recent breakthrough in this field, as emphasized in [19]–[23] robustly underscores the effectiveness of these techniques in discriminating a wide array of key events within the context of sports summarization. Rafiq *et al.* [19] worked on scene classification in cricket sports by applying transfer learning on AlexNet CNN to prevent model from over fitting. Deliege *et al.* [20] a large-scale annotated dataset of 500 untrimmed soccer broadcast videos is introduced, which is used by many reserchers for action spotting, camera shot segmentation, and replay grounding Liu *et al.* [21] also used visual and audio data to conduct an analysis which involves unsupervised shot clustering and supervised audio classification to capture mid-level patterns. Raventós *et. al.* [22] proposed methodology relies on segmenting the video sequence into shots and places particular emphasis on leveraging audio information to enhance the overall robustness of the summarization system. Shih [23] extensively explored content-aware techniques for analyzing and summarizing sports videos across a broad spectrum of sports, challenges, approaches, datasets, and evaluation metrics.

The above studies have indicated that the experiments ultimately illustrate how the use of audio features enhances the performance of event detection for event classification. This paper addresses a critical gap by incorporating audio classification into the summarization process. Our innovation extends to addressing no events, enabling the exclusion of irrelevant sections. This improvement fine-tunes the summarization process, leading to a more effective utilization of audio classification. Comprehensive methodology details are provided in the following section. In this paper, we focus on soccer summarization through the exploration of a deep learning-based audio classification method. We employ our extended ResNet-50 based proposed model to analyze audio files from soccer matches, predicting the seconds that encompass significant in-game events using transfer learning. Our approach effectively categorizes audio segments into two classes: i) events, representing crucial and thrilling moments and ii) no events, indicating less impactful parts. These identified crucial and thrilling moments can subsequently be utilized to generate highlights. To ensure accuracy, we carefully compiled our own dataset, the soccer action dataset compilation (SADC), as described in section 2 is the proposed method. We conduct a comparative analysis of our proposed approach with pre-trained deep learning models, including VGG19, DesNet121, and EfficientNetB7, presenting the results in section 3 is the results and discussion.

2. PROPOSED METHOD

Our goal is to detect significant events within soccer audio. Specifically, we target audio segments encompassing elements such as enthusiastic crowd cheering or heightened pitch in commentators' voices, which often correspond to key occurrences as suggested in [17]. Our proposed approach involves categorizing the most important and non-important parts of input soccer game audio in terms of seconds. By organizing these significant segments sequentially, we can create highlights. To achieve this, our methodology is divided into two sections, namely dataset compilation and event recognition framework. Dataset compilation explains how our own dataset named SADC, was formed. Event-recognition framework illustrates the technique used to predict and classify important moments in seconds. The identified moments can subsequently be visually arranged in a sequential manner to create highlights, as proposed in [8].

2.1. Dataset compilation

As indicated in [20], an optimal dataset is required to explore innovative tasks and approaches in the domain of soccer summarization. SADC, a dataset we created on our own, comprises 25 football video films downloaded from YouTube with a cumulative runtime of 34 hours, 33 minutes and 58 seconds (124,038 seconds). A group of five football fans carefully examined these videos. The start and end times of a variety of game related events were carefully recorded in this dataset as .csv file. The table format of it as per Table 1. Important occurrences including goals, goal attempts, penalty kicks, free kicks, penalty corners, and yellow cards are among the events that were recorded. Figure 1 illustrates the process of event recording by football fans. It marks an “event” when the audience cheering reaches a certain volume while watching a football match; otherwise, it is considered as “no event”.

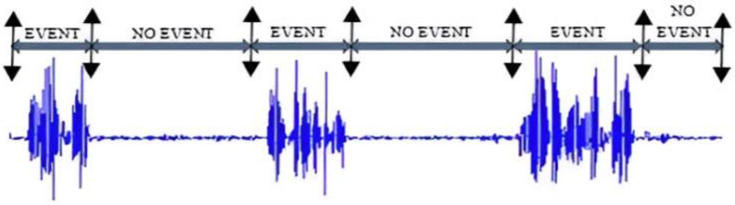

Figure 1. Process of recording events

Our focus is on identifying crucial occurrences in order to synthesize significant insights. As a result, we divided the recorded cases into two separate categories: i) events, which represents important occurrences and ii) no events, which represents all other instances as per Table 1. To enhance the stability of our model and ensure accurate prediction of all moments, we recorded all events within specific time frames, like 40, 50, 60, and 90 seconds as per the format shown in Table 2. To facilitate the training of the model, the video files have been transformed into .mp3 audio files. These audio files were then made available alongside the generated .csv file to ensure a comprehensive training approach.

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

Table 2. Processed event of SADC

Event name	Start time (sec.)	End time (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

2.2. Event recognition framework

This section presents the systematic methodology employed to achieve accurate audio-based classification by using SADC dataset. The suggested approach includes a number of steps that provide the prediction class labels "event" and "no event" for the training audio data provided. A range of libraries, including Librosa, Pandas, TensorFlow, Keras, and PIL, are imported to facilitate the tasks at hand. The objective is to create a systematic approach to classify important events from soccer audio. The process diagram of the proposed approach is illustrated in Figure 2.

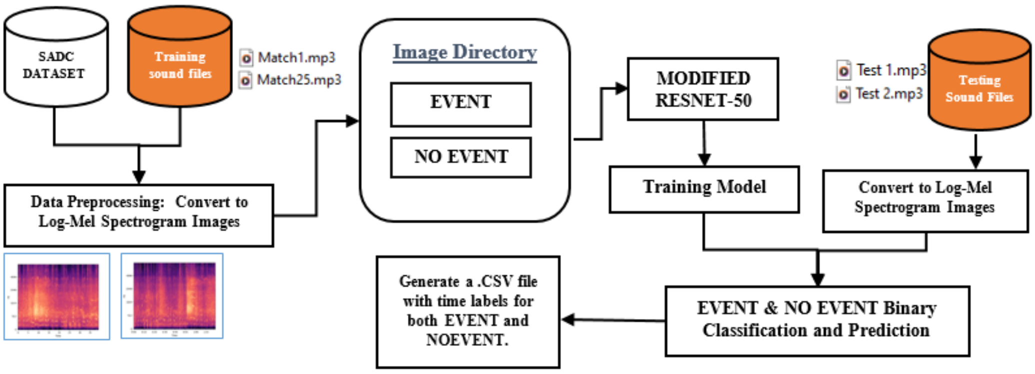
SADC DATASET
Training sound files
Match1.mp3
Match25.mp3
Data Preprocessing: Convert to Log-Mel Spectrogram Images
Image Directory
EVENT
NO EVENT
MODIFIED RESNET-50
Training Model
Test 1.mp3
Test 2.mp3
Testing Sound Files
Convert to Log-Mel Spectrogram Images
EVENT & NO EVENT Binary Classification and Prediction
Generate a .CSV file with time labels for both EVENT and NOEVENT.

Figure 2. Process diagram of event recognition framework and prediction

2.2.1. Input audio with soccer action dataset compilation dataset

In this section the provided dataset SADC is loaded, forming the foundations for subsequent operations. The data is manipulated, organized, and also rectifies discrepancies and standardizes parameter values for accurate analysis. After that all .mp3 audio files efficiently loaded using the "AudioFileClip" function from the "MoviePy" library which calculates the audio's duration in seconds, labeled as "duration," which is a significant parameter. For effective analysis, subsets of the dataset are extracted based on specific audio files and event types which is then given as an input to the data preprocessing stage.

2.2.2. Data-preprocessing

Sound features rely on psychoacoustic sound properties like loudness, pitch, and timbre. It commonly used cepstral features, such as mel-frequency cepstral coefficients (MFCC) and their derivatives [24]. In preprocessing section, raw audio transformed into visually insightful spectrogram images which co-ordinates the extraction of audio segments corresponding to predefined start and end times, thereby the extraction of audio segments slicing audio into meaningful fragments. These fragments are transformed into MFCC as shown in Figures 3 and 4 which are stored based on their classification category with appropriate filenames in predefined directories classified as "event" and "no event".

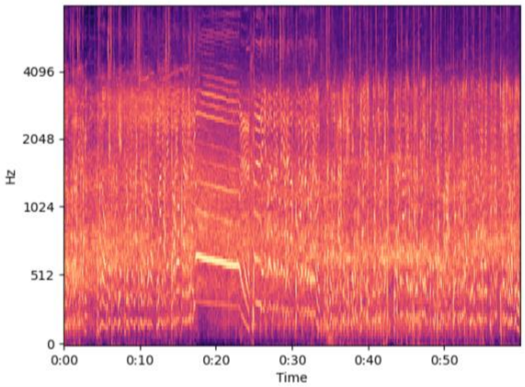

Figure 3. MFCC image for "event"

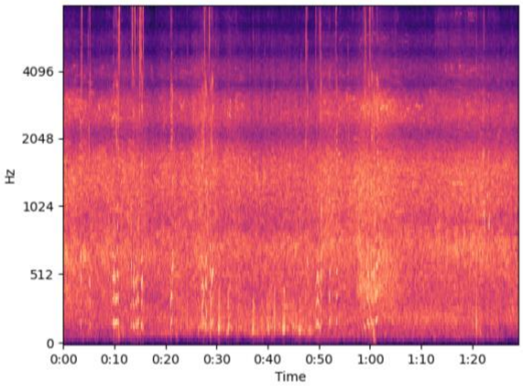

Figure 4. MFCC image for "no event"

2.2.3. Transfer learning with ResNet-50

Transfer learning methods, applied across various domains utilize knowledge acquired from one source to address classification, regression, and clustering challenges in a different destination [25]. This section focused on the applying transfer learning on ResNet-50 model as shown in Figure 5. First, it reads images from a specified directory and assigns inferred labels based on the subdirectory structure. The categorical label mode is chosen, and images are resized to 256×256 as implemented in [19]. The extended ResNet-50 model serves as the foundational backbone for the classification architecture, as depicted in Figure 5. Initially, all layers of the ResNet-50 are designated as non-trainable. Subsequent augmentation involves the addition of extra layers, including global average pooling, dense layers with dropout for regularization, and a final dense layer with softmax activation for binary classification. The trainModel is intricately designed to compile and train the model for a predetermined number of epochs. The binary cross-entropy loss function is employed, and accuracy is monitored in real-time during training. Additionally, training history is systematically logged for subsequent analytical purposes. The trained model is permanently stored at a specified location for future deployment.

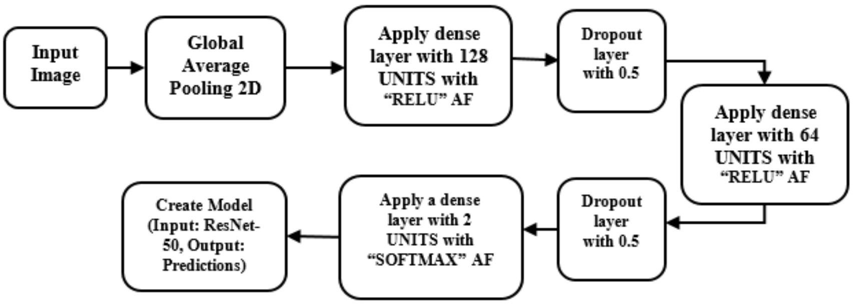
Input Image
Global Average Pooling 2D
Apply dense layer with 128 UNITS with "RELU" AF
Dropout layer with 0.5
Apply dense layer with 64 UNITS with "RELU" AF
Dropout layer with 0.5
Apply a dense layer with 2 UNITS with "SOFTMAX" AF
Create Model (Input: ResNet-50, Output: Predictions)

Figure 5. Flowchart of extended ResNet-50

2.2.4. Event prediction from audio images

This section introduces two crucial processes: "preprocess_image" and "predict_file_events". "preprocess_image" handles image files, processes them, and readies them for prediction. "predict_file_events" is responsible for the entire process of image preprocessing, event prediction, and result recording. The preprocessing step involves loading an audio image from the specified path, converting it into an array, and normalizing pixel values. Subsequently, the audio is divided into 60-second intervals. For each segment, a Mel spectrogram image is created as shown in Figure 4 and saved with an appropriate filename. After the preprocessing stage, the binary classification model trained with our extended ResNet-50 architecture. Image files from the specified location are loaded, extended predictions are made for each image, and the model's output determines the predicted class label. This information is then stored with the corresponding start and end times in the predictions list as per Table 3. After completing this process, we compare the observed event with predicted event. If they match, we classify the prediction outcome as a "match"; otherwise, it is classified as a "no match." Based on this comparison, we calculate the classification metrics.

Table 3. Event prediction evaluation

Observed event	Predicted event	Start time (sec.)	End time (sec.)	Prediction outcome	Class label
Event	No event	0	59	No match	FN
No event	Event	60	119	No match	FP
No event	No event	120	179	Match	TN
Event	Event	180	239	Match	TP
…	…	…	…	…	…
Event	Event	5,400	5,459	Match	TP

3. RESULTS AND DISCUSSION

The proposed methodology was applied on two distinct soccer test audio inputs each of 90 minutes soccer game downloaded from YouTube with four different epochs like 25, 30, 35, and 40. Both test audio inputs were classified into "event" and "no event" at 60 second intervals by two different football fans. The football fans precisely recorded each event. After that the algorithm's predicted events were compared with the observed events noted by the football fans, and the results were subsequently generated and analyzed for further evaluation as per the Table 3. A confusion matrix is created in classification to evaluate the performance of a model. These metrics collectively provide assessment of a classification models by calculating precision, accuracy, recall, and F1-score considering both correct and incorrect predictions as proposed in [26]. The results were quantitatively evaluated for accuracy and compared with those obtained from pre-trained models like VGG19, DesNet121, and EfficientNetB7. Table 4 shows accuracy comparison of our proposed approach with other pre-trained models, and Table 5 displays precision, recall, and F1-score values for different methods at epoch 40. Accuracy is measured as the overall correctness of the model by calculating the ratio of correctly predicted events to the total events [20]. Test audio 1 contains a total of 101 events, while test audio 2 comprises 105 events. Our experiments were conducted in the Google Colab environment. In our observations, the proposed model achieves an accuracy close to 80% after 40 epochs. Figures 6 and 7 show the accuracy measurements of both test audio files over different epochs. Increasing the number of epochs can potentially improve accuracy. Similarly, other pre-trained models also showed enhanced performance with more epochs but encountered memory limitations, often resulting in crashes. However, this is not the case with the extended ResNet-50. Increasing the number of epochs with the extended ResNet-50 leads to higher accuracy, precision, recall, and F1-score with reasonable processing time.

Table 4. Accuracy comparison of proposed model vs. pre-trained models of test audio

Epoch=40		Accuracy (%)	Precision	Recall	F1-score
Test audio-1	EfficientNetB7	58.42	0.36	0.22	0.27
	VGG19	48.51	0.22	0.25	0.23
	Desnet121	48.51	0.22	0.25	0.23
	Proposed model	79.21	0.79	0.45	0.58
Test audio-2	EfficientNetB7	65.35	0.31	0.31	0.31
	VGG19	69.52	0.36	0.35	0.35
	DesNet121	40.59	0.24	0.62	0.34
	Proposed model	79.05	0.54	0.77	0.63

Table 5. Performance metrics at epoch 40 for test audio-1 and test audio-2

Epoch=40		Accuracy	Precision	Recall	F1-score
Test Audio-1	EfficientNetB7	0.36	0.22	0.27	0.27
	VGG19	0.22	0.25	0.23	0.23
	Desnet121	0.22	0.25	0.23	0.23
	Proposed model	0.79	0.45	0.58	0.58
Test Audio-2	EfficientNetB7	0.31	0.31	0.31	0.31
	VGG19	0.36	0.35	0.35	0.35
	DesNet121	0.24	0.62	0.34	0.34
	Proposed model	0.54	0.77	0.63	0.63

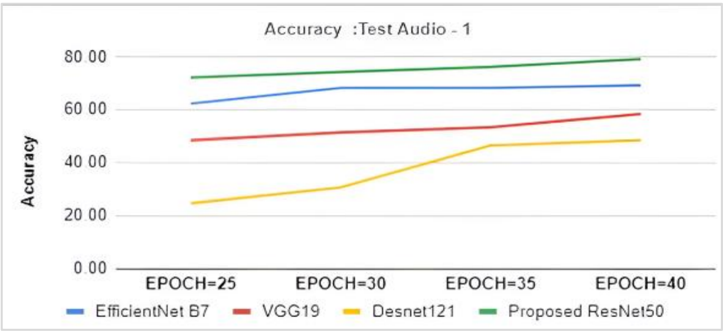

Figure 6. Accuracy measure of test audio-1 across various epochs

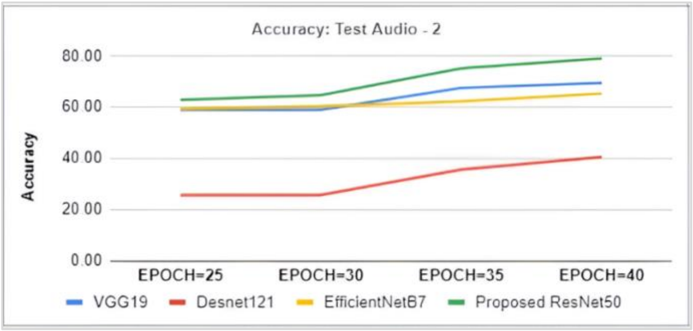

Figure 7. Accuracy measure of test audio-2 across various epochs

It is also noticeable from Figures 8 and 9 that our proposed model achieves high precision. Figures 10 and 11 illustrates that while maintaining significantly high precision, extended ResNet-50 manages to achieve a reasonable level of recall at epoch 40 for both test audio. This suggests that the model effectively identifies a substantial portion of actual event and indicates its ability to minimize false positives and enhance the relevance of detected event. Overall, the general and concluding observation is that as the training epochs increase, there is a noticeable improvement in the performance metrics for all models. Among them, extended ResNet-50 consistently stands out, securing the highest accuracy and maintaining a well-balanced precision, recall, and F1-score. VGG19 and EfficientNetB7 demonstrate slow improvement in performance in different aspects of precision and recall. On the other hand, DesNet121 falls behind the other models concerning overall accuracy and precision. Despite the promising results, our study is limited by the reliance on a manually annotated dataset and the constraints of computational resources available during testing. While our model effectively distinguishes between "event" and "no event," the diversity of soccer match scenarios and varying audio qualities could affect the generalizability of our results. Further testing on more diverse and larger datasets is needed to validate the broader applicability of our method.

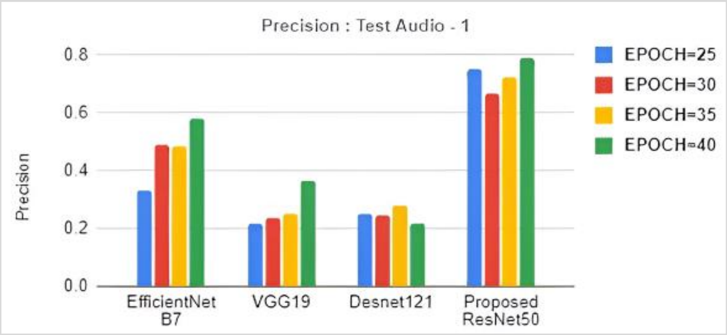

Figure 8. Proposed model vs. other pre-trained models: test audio-1 precision across different epochs

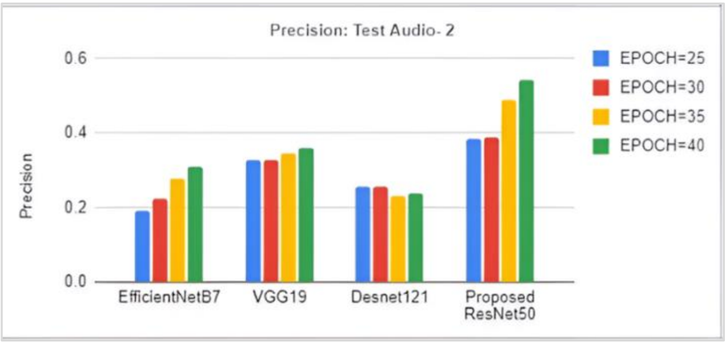

Figure 9. Proposed model vs. other pre-trained models: test audio-2 precision across different epochs

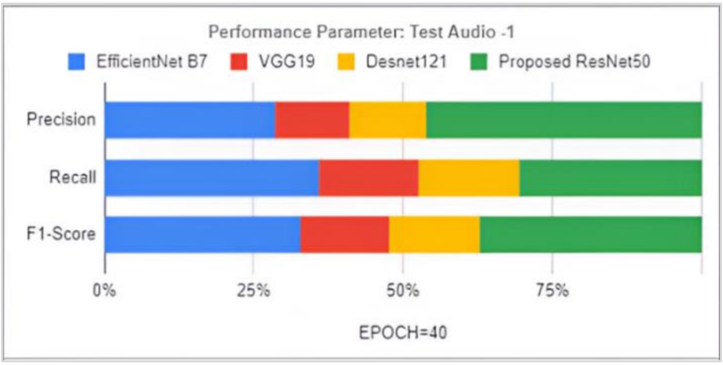

Figure 10. Measures of performance parameter over epoch 40 of test audio-1

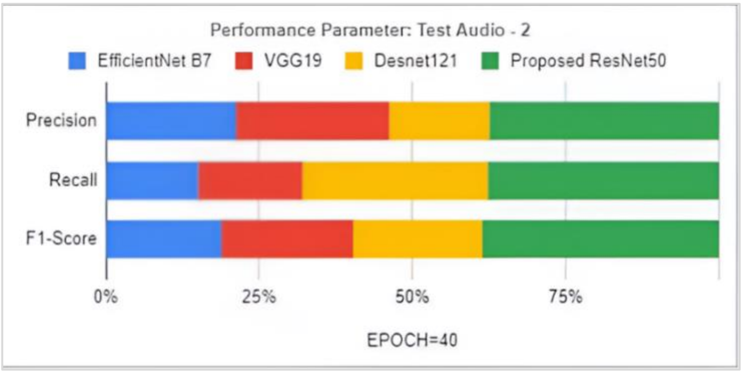

Figure 11. Measures of performance parameter over epoch 40 of test audio-2

4. CONCLUSION

This paper presents a novel approach to soccer audio classification using an extended ResNet-50 based deep learning model. The proposed methodology, validated with the precisely compiled SADC, demonstrated superior performance in accurately classifying significant in-game events. A comparative analysis was conducted between the proposed model and pre-trained models such as VGG19, DesNet121, and EfficientNetB7. Among these, the proposed model emerged as the most effective in extracting relevant events from soccer audio while filtering out irrelevant ones. The results, evaluated across different epochs, highlight the model's stability and accuracy in distinguishing important from unimportant events within the given soccer audio input. In the broader context of sports analytics, the proposed model stands out as a promising solution for content creators, analysts, and fans seeking concise and informative soccer highlights. Looking ahead, this approach could be applied to other field games like cricket or hockey and enhanced by incorporating visuals to further improve accuracy.

REFERENCES

[1] A. G. Money and H. Agius, “Video summarisation: a conceptual framework and survey of the state of the art,” *Journal of Visual Communication and Image Representation*, vol. 19, no. 2, pp. 121–143, 2008, doi: 10.1016/j.jvcir.2007.04.002.

[2] V. K. Vivekraj, S. E. N. Debashis, and B. Raman, “Video skimming: taxonomy and comprehensive survey,” *ACM Computing Surveys*, vol. 52, no. 5, 2019, doi: 10.1145/3347712.

[3] B. U. Gadhia and S. S. Modasiya, “An evaluation-based analysis of video summarising methods for diverse domains,” *Journal of Innovative Image Processing*, vol. 5, no. 2, pp. 127–139, 2023, doi: 10.36548/jiip.2023.2.005.

[4] M. Basavarajaiah and P. Sharma, “GVSUM: generic video summarization using deep visual features,” *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 14459–14476, 2021, doi: 10.1007/s11042-020-10460-0.

[5] E. Mendi, H. B. Clemente, and C. Bayrak, “Sports video summarization based on motion analysis,” *Computers and Electrical Engineering*, vol. 39, no. 3, pp. 790–796, 2013, doi: 10.1016/j.compeleceng.2012.11.020.

[6] Y. Takahashi, N. Nitta, and N. Babaguchi, “Video summarization for large sports video archives,” in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 1170–1173, doi: 10.1109/ICME.2005.1521635.

[7] Y. S. Khan and S. Pawar, “Video summarization: survey on event detection and summarization in soccer videos,” *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 11, 2015, doi: 10.14569/IJACSA.2015.061133.

[8] S. Jadon and M. Jasim, “Unsupervised video summarization framework using keyframe extraction and video skimming,” in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, 2020, pp. 140–145, doi: 10.1109/ICCCA49541.2020.9250764.

[9] O. A. N. Rongved *et al.*, “Real-time detection of events in soccer videos using 3D convolutional neural networks,” in *2020 IEEE International Symposium on Multimedia (ISM)*, 2020, pp. 135–144, doi: 10.1109/ISM.2020.00030.

[10] A. T. D. -Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, “Summarization of user-generated sports video by using deep action recognition features,” *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2000–2011, 2018, doi: 10.1109/TMM.2018.2794265.

[11] S. H. Emon, A. H. M. Annur, A. H. Xian, K. M. Sultana, and S. M. Shahriar, “Automatic video summarization from cricket videos using deep learning,” in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, 2020, pp. 1–6, doi: 10.1109/ICCIT51783.2020.9392707.

[12] M. Sanabria, F. Precioso, and T. Menguy, “Hierarchical multimodal attention for deep video summarization,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 7977–7984, doi: 10.1109/ICPR48806.2021.9413097.

[13] G. Evangelopoulos *et al.*, “Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention,” *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013, doi: 10.1109/TMM.2013.2267205.

[14] B. Vanderplaetse and S. Dupont, “Improved soccer action spotting using both audio and video streams,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 3921–3931, doi: 10.1109/CVPRW50498.2020.00456.

[15] M. Ilse, J. M. Tomczak, and M. Welling, “Attention-based deep multiple instance learning,” in *35th International Conference on Machine Learning, ICML 2018*, 2018, vol. 5, pp. 3376–3391.

[16] M. Sanabria, Sherly, F. Precioso, and T. Menguy, “A deep architecture for multimodal summarization of soccer games,” in *Proceedings Proceedings of the 2nd International Workshop on Multimedia Content Analysis in Sports*, Oct. 2019, pp. 16–24, doi: 10.1145/3347318.3355524.

[17] R. Agyeman, R. Muhammad, and G. S. Choi, “Soccer video summarization using deep learning,” in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp. 270–273, doi: 10.1109/MIPR.2019.00055.

[18] Z. Ji, F. Jiao, Y. Pang, and L. Shao, “Deep attentive and semantic preserving video summarization,” *Neurocomputing*, vol. 405, pp. 200–207, 2020, doi: 10.1016/j.neucom.2020.04.132.

[19] M. Rafiq, G. Rafiq, R. Agyeman, G. S. Choi, and S.-I. Jin, “Scene classification for sports video summarization using transfer learning,” *Sensors*, vol. 20, no. 6, 2020, doi: 10.3390/s20061702.

[20] A. Deliege *et al.*, “SoccerNet-v2: a dataset and benchmarks for holistic understanding of broadcast soccer videos,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 4503–4514, doi: 10.1109/CVPRW53098.2021.00508.

[21] C. Liu, Q. Huang, S. Jiang, L. Xing, Q. Ye, and W. Gao, “A framework for flexible summarization of racquet sports video using multiple modalities,” *Computer Vision and Image Understanding*, vol. 113, no. 3, pp. 415–424, 2009, doi: 10.1016/j.cviu.2008.08.002.

[22] A. Raventós, R. Quijada, L. Torres, and F. Tarrés, “Automatic summarization of soccer highlights using audio-visual descriptors,” *SpringerPlus*, vol. 4, no. 1, 2015, doi: 10.1186/s40064-015-1065-9.

[23] H.-C. Shih, “A survey of content-aware video analysis for sports,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1212–1231, 2018, doi: 10.1109/TCSVT.2017.2655624.

[24] E. Tsalera, A. Papadakis, and M. Samarakou, “Comparison of pre-trained CNNs for audio classification using transfer learning,” *Journal of Sensor and Actuator Networks*, vol. 10, no. 4, 2021, doi: 10.3390/jsan10040072.

[25] N. Zakaria, F. Mohamed, R. Abdelghani, and K. Sundaraj, “VGG16, ResNet-50, and GoogLeNet deep learning architecture for breathing sound classification: a comparative study,” in *2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP)*, 2021, pp. 1–6, doi: 10.1109/AI-CSP52968.2021.9671124.

[26] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010, doi: 10.1109/TKDE.2009.191.

BIOGRAPHIES OF AUTHORS

Bijal Utsav Gadhia is pursuing Ph.D. in computer engineering from Gujarat Technological University (State University), Gujarat, India. Currently, she is a faculty member at Government Engineering College, Gandhinagar (Government Employee), Gujarat, India and has served several governmental activities around the university and outside. Her research interests are the application of deep learning, machine learning, image processing, and data science. She has published various research papers in the field of image processing and deep learning. She can be contacted at email: bij.1988@gmail.com.

Dr. Shahid S. Modasiya is an Assistant Professor at the Department of Electronics and Communication Engineering at Government Engineering College, Gandhinagar under the affiliation of Gujarat Technological University. His research interest areas are image processing, artificial intelligence, RF and microwave and antenna design. He has also published two patents and various papers in the field of his research interest. He can be contacted at email: shahid@gecg28.ac.in.