AI-Driven Salient Soccer Events Recognition Framework for Next-Generation IoT-Enabled Environments

Khan Muhammad, *Member, IEEE*, Hayat Ullah, *Graduate Student Member, IEEE*, Mohammad S. Obaidat, *Life Fellow, IEEE*, Amin Ullah, *Associate Member, IEEE*, Arslan Munir, *Senior Member, IEEE*, Muhammad Sajjad, and Victor Hugo C. de Albuquerque, *Senior Member, IEEE*

***Abstract*—The salient event recognition of soccer matches in the next-generation Internet-of-Things (Nx-IoT) environment aims to analyze the performance of players/teams by the sports analytics and managerial staff. The embedded Nx-IoT devices carried by the soccer players during the match capture and transmit data to an artificial intelligence (AI)-assisted computing platform. The interconnectivity of data acquisition devices with an AI-assisted computing platform in the Nx-IoT environment will not only allow the spectators to track the formation of their favorite players during a soccer match but will also enable the managerial staff to evaluate the players' performance in the soccer match as well as in practice sessions. This Nx-IoT-enabled salient event detection feature can be provided to spectators and sports' managerial staff as a financial technology (FinTech) service. In this article, we propose an efficient deep learning-based framework for multiperson salient soccer event recognition in IoT-enabled FinTech. The proposed framework performs event recognition in three steps: first, image frames are extracted from video streams and resized in the preprocessing step to match the input of the deep network. Second, frame-level discriminative features are extracted using a pretrained convolutional neural network (CNN) architecture. Third, we employ a multilayer long short-term memory (MLSTM) network to recognize high-level events in soccer videos by exploiting the sequential relation between adjacent frames. Moreover, we introduce a new soccer video events (SVE) data set containing videos of six salient events of soccer game. To provide a strong baseline, we evaluate our newly created SVE data set using different traditional machine learning and deep learning algorithms. We also perform event recognition on untrimmed soccer videos using our proposed framework and compare the results with state-of-the-art methods. The obtained results validate the suitability of our proposed framework for salient event recognition in Nx-IoT environments.**

***Index Terms*—Computer vision, convolutional neural network (CNN), edge computing, event recognition, multilayer long short-term memory (MLSTM), next-generation Internet of Things (Nx-IoT).**

I. Introduction

THE NOTICEABLE advancement in networking technologies and efficient deep learning algorithms over edge devices has allured the sports industry to adopt next-generation Internet-of-Things (Nx-IoT)-enabled financial technology (FinTech) services for a wide range of applications. The IoT is a network, where multiple edge devices/sensors are interconnected via the Internet. The Nx-IoT devices in sports communicate and transmit data with other IoT/edge devices for edge-centric distribution and processing of sports data. Mostly, sports organizations, especially soccer officials, provide edge-based IoT environments, which can significantly improve the quality of sport analytics systems and enable the spectators to have a more enjoyable interactive experience. The video data captured by the vision sensors can be instantly processed over edge devices and then transmitted to artificial intelligence (AI)-assisted edge computing platforms for a variety of applications, such as event detection/recognition, player identification, and players formation tracking in the field of a soccer stadium equipped with Nx-IoT. Due to the exponential growth of fans following, soccer has become the world's most watchable sport with more than 4.0 billion audience worldwide [1]. According to a recent report (Google: Watch time for YouTube sports highlights jumps 80%), 90% of online viewers search for soccer videos highlights or prefers to access salient sports events (such as goal, penalties, fouls, corner-shots, etc.) rather than watching full matches [2]. Furthermore, the live spectators inside the stadium are very excited to support their favorite team/players and cheer for their best performance. The Nx-IoT-enabled edge-based event recognition service in soccer stadiums can improve the experiences of live spectators by providing an interactive entertainment environment.

Considering the complex game rules and players with different field formations, soccer is the most difficult game to analyze. Researchers around the globe are contributing to different aspects of soccer event detection and recognition. Event recognition is an essential component for high-level

Manuscript received February 25, 2021; revised June 16, 2021 and July 18, 2021; accepted August 19, 2021. This work was supported in part by the tenure of an ERCIM "Alain Benoussan" Fellowship Programme under Contract 2019–40, and in part by the Color and Visual Computing Lab, Department of Computer Science, NTNU, Gjøvik. The work of Mohammad S. Obaidat was supported in part by the PR of China Ministry of Education Distinguished Possessor under Grant MS2017BJKJ003. *(Khan Muhammad and Hayat Ullah are co-first authors.) (Corresponding authors: Khan Muhammad; Muhammad Sajjad.)*

Please see the Acknowledgement section of this paper for the author affiliations.

Digital Object Identifier 10.1109/JIOT.2021.3110341

sports video analytic tasks, such as event-aware highlights generation [3], sports videos retrieval [4], and indexing of sports videos [5]. However, soccer events are different from other sports events, where a video clip contains fascinating contents for a random time interval with semantically starting and ending boundaries of events rather than a fixed time interval. For instance, in counterattack, the brilliant assist before the goal and the celebrations after the goal are the complete soccer events. All these events are high-level semantics, which can be recognized with multiscale deep features (i.e., extract convolutional neural network (CNN) features at different layers with varying spatial dimensionality). For instance, in the soccer match, a goal is an event that involves different movements of the human body, such as running, jumping, passing ball, and shooting ball toward goalmouth.

The earlier research studies [6], [7] for soccer event detection and recognition were based on low-level or handcrafted visual features and traditional machine learning algorithms. These low-level feature-based methods rely on global features, including the texture, edge, color, shape, and motion information. Although these methods made some achievements in the last decade, they could only detect few soccer events and usually failed for complex type of events with a clutter background. To interrelate the semantic gap between the low-level semantics and high-level semantics, several traditional soccer event detection methods have been proposed. These methods utilized mid-level features to obtain the intermediate representation of soccer events, including field view classification, player tracking, scoreboard detection, and play–break segmentation. For instance, Zhao *et al.* [8] used mid-level features for video segmentation into play–break segments. They segmented the video based on the color, contour, and histogram. The works presented in [3], [9], and [10] first extracted the excitement clip from lengthy videos, and then detected the salient events using histogram and color computations. In addition, these earlier methods often required additional information such as text from score boards and audio commentary related to the game play. Despite, acquiring additional information about game, the results achieved by these methods still suffered from misclassification for complex events.

Recently, deep learning has gained tremendous attention in computer vision, and has significantly improved the performance of event detection and action recognition systems. Various CNN-assisted approaches have been proposed for soccer event detection and annotation [11], [12], which extract both the spatial and temporal features from the video frames and analyze the event's type and boundary for specific time interval. While, some researchers have adopted 3D-CNN-based soccer event detection approaches that extract spatial and temporal features from video frames [13], [14], other studies [15], [16] have presented the combined CNN and recurrent neural network (RNN) frameworks for soccer event detection and achieved state-of-the-art results. However, most of the contemporary deep learning methods are limited to certain types of events and single-person event detection and cannot be deployed in IoT-enabled environments. Considering the availability of embedded edge devices and efficient deep learning architectures, there is a need to develop an efficient edge computing-based approach for salient soccer event recognition in Nx-IoT-enabled environments. Besides the existence of smart embedded devices and energy-friendly deep learning architectures, it is very crucial to have a sufficient amount of data for the problem/task under consideration. The availability of problem-related data sets greatly helps the researchers to train and evaluate their proposed systems without devoting considerable efforts on generating new data sets. However, to the best of our knowledge, there are very few soccer videos data sets [17]–[19] and, furthermore, the existing data sets are very specific and do not cover salient events of soccer.

Therefore, in this article, we present an efficient deep learning-based framework for salient soccer event recognition over edge-centric FinTech computing platform and our newly created soccer videos events (SVE) data set. The proposed framework performs event recognition process in three steps: 1) preprocessing; 2) features extraction; and 3) sequence learning. In the preprocessing step, image frames are extracted from video streams and resized to match the input of the deep network. For feature extraction, our framework uses a pretrained CNN architecture, which extracts deep discriminative features from the video frames. While for sequence learning, a multilayer LSTM is used to analyze the video stream by capturing the temporal relation between adjacent frames. Our newly created SVE data set contains short duration clips of six different soccer events. To better understand the problem, we first evaluate the soccer event recognition with handcrafted features and a well-known machine learning classifier (HOG+SVM). Later, we investigate the soccer event recognition problem using multilayer long short-term memory (MLSTM) along with different CNNs on our SVE data set. The key contributions of our scheme can be summarized as follows.

1) To recognize the salient events in soccer matches, we investigated traditional machine learning (HOG+SVM) and deep learning-based approaches (CNN+MLSTM) for FinTech-enabled soccer event recognition service and propose an energy-efficient CNN+LSTM framework. Our proposed framework strikes a tradeoff between computational complexity and model accuracy and is a suitable solution for edge-centric FinTech computing platforms and similar domains associated with Nx-IoT environments, showing its flexibility and scalability.
2) The literature contains very few data sets for soccer event detection/recognition. However, there is no benchmark data set of key events, which defines the interest of live/offline spectators. We have created our own SVE data set, which contains salient events of soccer matches captured from multiple views. The SVE data set will be publicly available for further research to mature the event detection/recognition systems for soccer videos.
3) We have conducted comprehensive experiments on our newly created SVE data set to evaluate the performance of our framework. Furthermore, we have tested the proposed framework on relevant events from other data sets and have conducted a comparative study. The obtained results reveal that the proposed framework generalizes well and performs better than existing methods.

The remainder of this article is organized as follows: Section II presents the overview of the related works. The

proposed framework is presented in Section III followed by experimental evaluation of the proposed framework in Section IV. Finally, Section V concludes this article with possible future directions.

II. Related Work

In this section, we briefly describe the event recognition literature and critically discuss the soccer event recognition approaches that are reported in the recent works along with their strengths and limitations. Generally, the soccer event recognition methods can be categorized into two parts: 1) low-level features-based and 2) deep learning-based methods.

A. Low-Level Features-Based Event Recognition Approaches

Event recognition has played a very important role in different domains of sports video content analysis, including highlight generation, event-based sports video retrieval, and statistical summary generation of sports videos (e.g., soccer, hockey, etc.). A variety of methods has been proposed to automate the event recognition process in sports videos. Most of the early methods [6], [7], [20] were based on low-level features. These methods usually used handcrafted descriptors and machine learning classifiers for feature extraction and classification, respectively. For instance, Tavassolipour *et al.* [21] proposed automatic event detection in soccer videos for highlights generation. They used a hidden Markov model for the segmentation of a video into meaningful segments, named play-back patterns, followed by mid-level features extraction from each segment. Finally, they extracted discriminative features using a Bayesian network. Kolekar and Sengupta [3] proposed an automatic highlight generation system that could generate highlight from sports TV broadcasts. First, they detected the exciting clips using audio features and then segmented the clips into different scenes. Next, they assigned a concept-score to each scene within a clip using a probabilistic Bayesian belief network (PBBN) and selected the scene with a higher concept-score. Wang *et al.* [12] proposed a soccer video annotation framework based on coarse-grained time information. They annotated the soccer events by synchronizing the video clips and external text information (match reports) with coarse time constraints. Fakhar *et al.* [22] presented a learning-based soccer event detection approach based on two main concepts. First, they analyzed the frame and estimated the saliency of each frame regarding soccer events. Second, the event-oriented and discriminative dictionary was learned using their proposed K-SVD algorithm. Furthermore, Bennett *et al.* [23] proposed a technique for video-based talent identification of the youth for soccer using smart devices. Hosseini and Eftekhari-Moghadam [24] presented a fuzzy rule-based system for soccer event detection and annotation. They used statistical information quantized from audiovisual features and rule-based reasoning classifier, which constructed the semantic perception for the occurred events. However, these handcrafted or low-features-based methods are less effective and time consuming for detecting high-level soccer events. These limitations can create issues when processing lengthy videos or dealing with sports TV broadcasts. Besides these traditional handcrafted-based methods, numerous learning-based event detection/recognition methods have been proposed, which significantly improve the event detection and recognition task and overcome the limitations of traditional methods.

B. Deep Learning-Based Event Recognition Approaches

Recently, CNN-oriented methods have achieved greater success and have improved the performance of various computer vision tasks, including image classification [25], [26], object detection [27], [28], image enhancement [29], [30], speech recognition [31], [32], and activity recognition [33], [34]. For instance, Jiang *et al.* [15] proposed a deep learning-based approach for soccer video event detection. Their method utilized the combination of CNN and RNN, where they segmented the soccer video in play–break segments by determining the event boundary, and then extracted CNN features of key frames from play–break segments. Finally, RNN was deployed to recognize the salient soccer events, including the goal, goal attempt, card, and corner. Tsunoda *et al.* [35] presented a hierarchical RNN for analyzing the understanding between players of team sports activity. They integrated multiple person-centered features with LSTM cell output over temporal sequences. Furthermore, Fani *et al.* [36] introduced a parallel feature fusion (PFF) network for automatic event detection and classification in soccer broadcast videos. The PFF combined local as well as full scene features for zoom in and zoom out scene classification. A hidden observable Markov model was deployed to determine play/break status of the scenes in soccer videos. Giancola *et al.* [37] introduced a benchmark SoccerNet data set for action spotting in soccer videos. The duration of the data set was 764 h and consisted of the goal, yellow/red card, and substitution. To prevent violence incidents in football stadiums, Fenil *et al.* [38] proposed a real-time violence detection system for recognizing violence using human intelligence simulation. Their proposed method processed enormous amounts of real-time video streams from different sources, where histogram of oriented gradients (HOG) was used as a feature descriptor followed by bidirectional long short-term memory (BDLSTM). Liu *et al.* [13] proposed a soccer event detection method based on temporal action localization and play–break segmentation. First, they localized the action in soccer videos using 3-D CNN and then employed play–break rules for organizing actions into corresponding events. These deep learning-based approaches have shown remarkable performance and have effectively overcome the limitations of low-level features-based methods. On the other hand, these deep learning-based approaches require high computation power for training purposes. Different from the existing methods, our proposed framework efficiently reduces the computational complexity by adopting transfer learning and frame skip strategies.

III. Proposed Framework

Human action-oriented events typically involve sequences of specific human postures evolving in video frames, which demonstrate variations in both spatial and temporal domains.

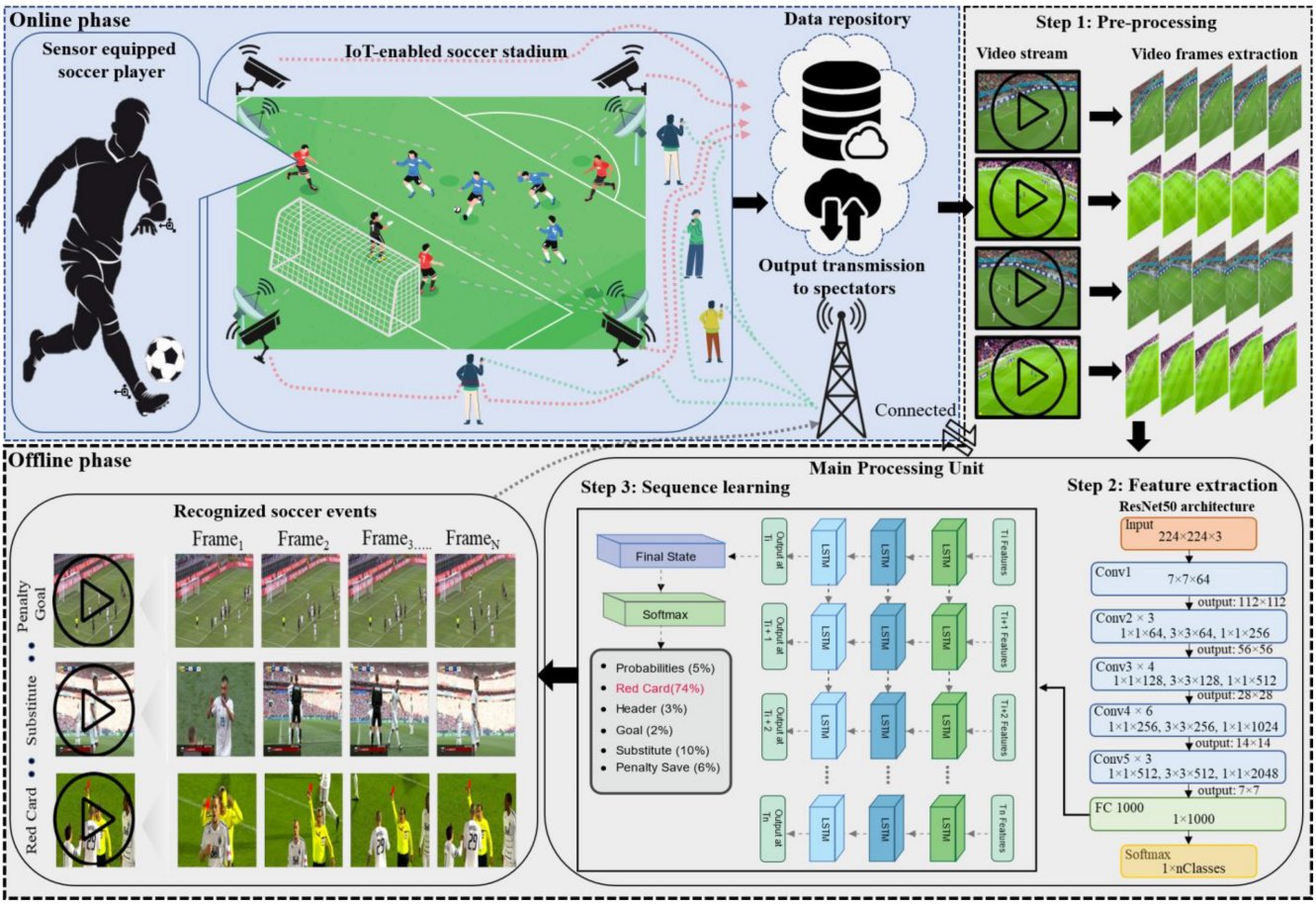


Fig. 1. Detailed overview of our proposed framework for salient soccer event recognition over edge devices in Nx-IoT-enabled environment. **Step 1**: This step involves frames extraction and frame resizing. **Step 2**: This step involves features extraction, where we employ a pretrained CNN (ResNet50) architecture to extract discriminative features from video frames. **Step 3**: This step receives CNN features and trains multilayer LSTM, which outputs a confidence score for an event detected in sequences of video frames.

For example, the goal event in soccer consists of more than one action where each action is the combination of different postures of human body. While analyzing the soccer video, these high-level actions can be visualized as hidden sequential patterns, which can be detected and recognized with strong representation (high-level features). In this article, we propose a deep learning-based salient event recognition framework, which analyzes the input soccer video using CNN with MLSTM. The proposed framework comprises of three steps.

Step 1 (Preprocessing): The preprocessing step extracts the image frames from the video stream and resize the frames to match the input of the deep neural network.

Step 2 (Feature Extraction): In the feature extraction step, our framework extracts deep features using a pretrained CNN network from the sequence of frames.

Step 3 (Sequence Learning): In the third step, the extracted features are fed into an MLSTM to retrieve high-level abstraction and temporal information for the event recognition task. The workflow of the proposed framework and its main components are illustrated in Fig. 1. Algorithm 1 presents the stepwise implementation of soccer event recognition process. The input and output parameters used in our proposed framework are presented in Table I.

A. Preprocessing

In the preprocessing step, frames are extracted from the video with three different frame skip strategies and then resized each frame to match the input of the deep neural network. In our proposed framework, the input is resized to $224 \times 224 \times 3$ to match the input layer dimensionality of the ResNet50 architecture.

B. Features Extraction ResNet50 Architecture

Soccer event is the combination of multiple actions (running, jumping, and passing, etc.), where each action is itself the integration of different poses. The event recognition workflow starts from low-level semantics extraction (actions) to high-level semantics (events). To represent these sequences of actions, CNN features from each frame are expressed as an individual feature vector. Similarly, the complete video can be represented as a set of feature vectors having a sequential relation between the adjacent feature vectors. On the other hand, traditional machine learning algorithms use handcrafted features for event detection and recognition tasks, which require a lot of efforts for feature engineering and scaling. Despite the extraordinary efforts for feature engineering, traditional soccer

TABLE I
DESCRIPTION OF PARAMETERS USED IN OUR PROPOSED FRAMEWORK FOR INPUT AND OUTPUT OPERATIONS

Symbols	Description
fc-1000	Fully connected layer of ResNet50 architecture.
f_s	Number of frame skip during feature extraction.
x_t	Input to LSTM at time t.
f_t	Output of forget gate.
i_t	Output of input gate.
o_t	Output of output gate.
c_t	Output of current state of LSTM cell.
c_{t-1}	Previous state of LSTM cell.
w_f	Weights of forget gate of LSTM cell.
w_i	Weights of input gate pf LSTM cell.
w_o	Weights of output gate of LSTM cell.
b_f	Biases of forget gate.
b_i	Biases of input gate.
b_o	Biases of output gate.
h_t	Final output of LSTM cell.

Algorithm 1 Event Recognition in Soccer Videos

Input: Soccer video V_{soccer}
Preparation:
1. Load pretrained ResNet50 CNN network M_f
2. Load trained multilayer LSTM network M_c

Steps:
while (V_{soccer})
1. Read frames $\leftarrow$ (f_i, V_{soccer})
2. Pass frame f_i to ResNet50 CNN
3. Extract feature $fv_i \leftarrow M_f$ (f_i)
4. Forward feature vector fv_i to trained LSTM M_c, and Predict event class $\leftarrow M_c$ (fv_i)
5. Display predicted event with confidence score

end while

Output: Display event with predicted label and confidence score

event recognition approaches are still unable to detect complex and long-duration events.

CNN is originally introduced for the image classification task [39] and has achieved state-of-the-art results. It has the ability to extract features of different scales and is equipped with a classifier at the end of architecture. CNNs are widely used for a variety of high-level computer vision tasks. The main reason behind the success and achievements of CNNs is the hierarchical nature of the architecture that contains a series of layers, including convolution, pooling, and fully connected layers. The convolutional layer generates different representations of the same image by convolving different kernels with different sizes. The pooling layer subsamples the input feature maps by selecting the high activations values, while the fully connected layer learns high-level representations and reshapes the input feature maps to a 1-D feature vector. Training a new CNN architecture from the scratch requires a huge amount of image data along with powerful machines for execution, such as graphics processing units (GPUs) or tensor processing units (TPUs). This problem can be addressed using transfer learning strategy where a pretrained model is utilized for another computer vision problem. To this end, our proposed framework uses a pretrained ResNet50 architecture [40], which is trained on the ImageNet data set containing more than 20 000 categories (door, chair, and car, etc.). The first layer of the ResNet50 architecture is input layer with dimensionality of $224 \times 224 \times 3$, the second layer is the convolution layer with a kernel size of 7×7. The rest of the architecture has four residual blocks, fully connected layers, and a Softmax layer. Each residual block contains three convolutional layers with kernel sizes of 1×1 and 3×3 followed by ReLU activation and Batch Normalization layers. We have used a fully connected layer (fc-1000) as a generic feature descriptor. Each feature vector represented a single frame of video, these features are then fed into MLSTM in the form of features block for a fixed time interval. MLSTM processes these features and learns the hidden sequential patterns from the input feature data. The detailed explanation of RNN and its variants are presented in the next section.

C. Event-Specific Sequence Learning Using Multilayer LSTM

Despite the powerful characteristics and flexibility, CNNs can only be used for tasks where inputs and outputs have fixed dimensionality and mostly fail while dealing with the data having different input and output dimensionality. Along with this limitation, CNNs are restricted to static data and cannot be used for problems dealing with time series and sequential data. Most of the problems such as speech recognition, machine translation, and activity recognition in videos are efficiently expressed with sequences having variable lengths. To solve sequential pattern learning problems or predicting time-series data, the need of such a method becomes crucial that can precisely map sequences and learn its hidden patterns from input time-series data. To meet the needs of such kind of systems, a special kind of neural network, named RNN, has been introduced which has the ability to learn from temporal features and map the temporal relation of a given time-distributed data. RNNs are specially designed for the classification of time-series and sequential data. RNNs analyze the hidden sequential patterns in both spatial and temporal dimensions by connecting the previous information with the current information and predict the future output. The suitability and efficiency of RNN for temporal data analytics has attracted the research community to investigate it for various time-series prediction and sequence classification problems and achieved incredible results. Although RNNs can decipher the hidden sequential patterns in time-series data (i.e., video, audio, or numerical data), RNNs fail to remember earlier information while interpreting long term sequences. Such a type of problem is known as a vanishing gradient or gradient exploding, which can be resolve by using a special variant of RNN known as LSTM, which has the ability to remember the earlier input information for a long-time interval.

1) Multilayer LSTM Network: The LSTM [41] is an extension of the RNN architecture, which is specially designed for interpreting long-term sequences, thereby resolving the problem of vanishing gradient and gradient exploding faced

by RNNs. The internal structure of LSTM consists of several cell units, where each cell unit contains special gates (input, output, and forget gates) that switch the flow of information and control the sequential pattern recognition process. These gates are configured in such a way that each gate receives the input from the previous stage and forwards the computed output to the next gate. All these gates are controlled by a sigmoid function. For instance, the input gate $\boldsymbol{i}_{bmt}$ decides that what portion of information should be updated, whereas the output gate o_t stores the information of the coming sequence. The forget gate f_t processes the information from the input gate and the previous cell state and removes the previous information from the memory when needed. The recurrent unit g computes the previous cell state $\boldsymbol{c}_{t-1}$ and the current input x_t using the *tanh* activation function, whereas h_t can be computed by multiplying the value of the output gate with *tanh* of the current cell state c_t. The final output can be obtained by passing the h_t to the softmax classifier. The mathematical equations of the operations performed by these gates are given in (1)–(7)

$$i_t = \sigma\left(w_i * \left[h_{t-1}, x_t\right] + b_i\right) \tag{1}$$

$$o_t = \sigma\left(w_o * \left[h_{t-1}, x_t\right] + b_o\right) \tag{2}$$

$$f_t = \sigma\left(w_f * \left[h_{t-1}, x_t\right] + b_f\right) \tag{3}$$

$$g = \tanh\left(w_g * (x_t + c_{t-1}) + b_g\right) \tag{4}$$

$$c_t = ((c_{t-1} * f_t) + (g * i_t)) \tag{5}$$

$$h_t = (\tanh(c_t) * o_t) \tag{6}$$

$$\text{output} = \text{soft}\max(h_t). \tag{7}$$

In general, the performance of any deep neural network can be improved by stacking more and more layers; similarly, the hidden sequential pattern leaning capability of an LSTM can be enhanced by increasing the number of layers in the network. Therefore, we add three layers to our LSTM network, thereby increasing the ability to analyze the given input data at different time scales and produce good results as compared to a standard LSTM. Unlike the standard LSTM, when data are fed to the MLSTM, the input data are processed in several layers in a hierarchical fashion, where each layer in the network receives the hidden state of the previous layer as an input and forwards the output to the next layer. The computational process of the memory cell of the MLSTM is the same as the standard LSTM as explained by (1)–(7). Fig. 2 depicts the building block of MLSTM, where the first hidden layer receives data from the input layer and the input of the second hidden layer is the output of the first hidden layer. Similarly, the input of the third hidden layer is the output of the second hidden layer. The final output is obtained by computing the output of the final last hidden layer using softmax.

IV. Experimental Results and Discussion

This section presents comprehensive experimental evaluation of our proposed framework and details about the SVE data set used in experiments. We have first performed the salient event recognition using SVM with HOG descriptor, and then assessed the performance of different state-of-the-art architectures with MLSTM for salient event recognition. Furthermore, the proposed framework is implemented

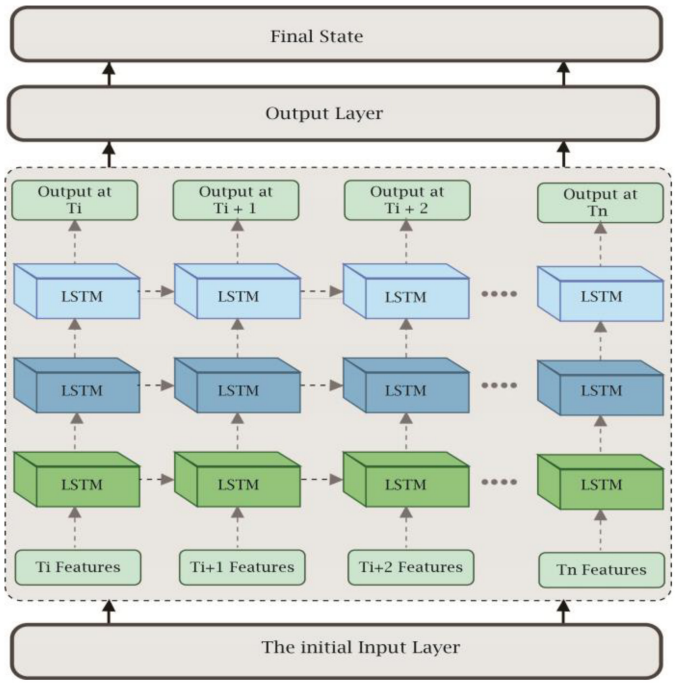


Fig. 2. External structure of a multilayer LSTM.

using MATLAB 2018 with Matconvnet on PC equipped with 3.60-GHz Intel Core i7 processor and NVIDIA GTX 1080 with 4-GB GPU. We have initialized the training with a random weight initializer for 60 epochs with a batch size of 32. For weight optimization, we have used the Adagrad optimizer with learning rate 0.0001. For performance evaluation, we have used five evaluation metrics, including the Precision, Recall, true-positive rate (TPR), false-positive rate (FPR), and $F1$-score.

A. Details of the Data Set

For any advanced computer vision problem, the data acquisition phase is very crucial because without appropriate and sufficient amount of data, one cannot achieve desirable results. Furthermore, the collected data must be labeled properly according to the type of data and nature of the problem. Since this article is focusing on recognition of salient events in soccer videos, this research problem requires a sufficient amount of labeled soccer video data.

To the best of our knowledge, very few soccer videos data sets are presented so far for specific types of tasks, including ball tracking, player position, and movement tracking. However, these data sets do not consist of the generic type of events, such as, the Goal, Substitute, and Red Card, etc. Therefore, in this article, we present a newly created balance SVE data set of soccer videos, which comprises of short video clips of six different events, including the goal, penalty save, penalty goal, card, head goal, and substitute. Also, our newly created SVE data set contains event videos captured from different views with both far and close field of views that offer great variety in the data. The SVE data set is created in three distinct phases: 1) we collect soccer videos

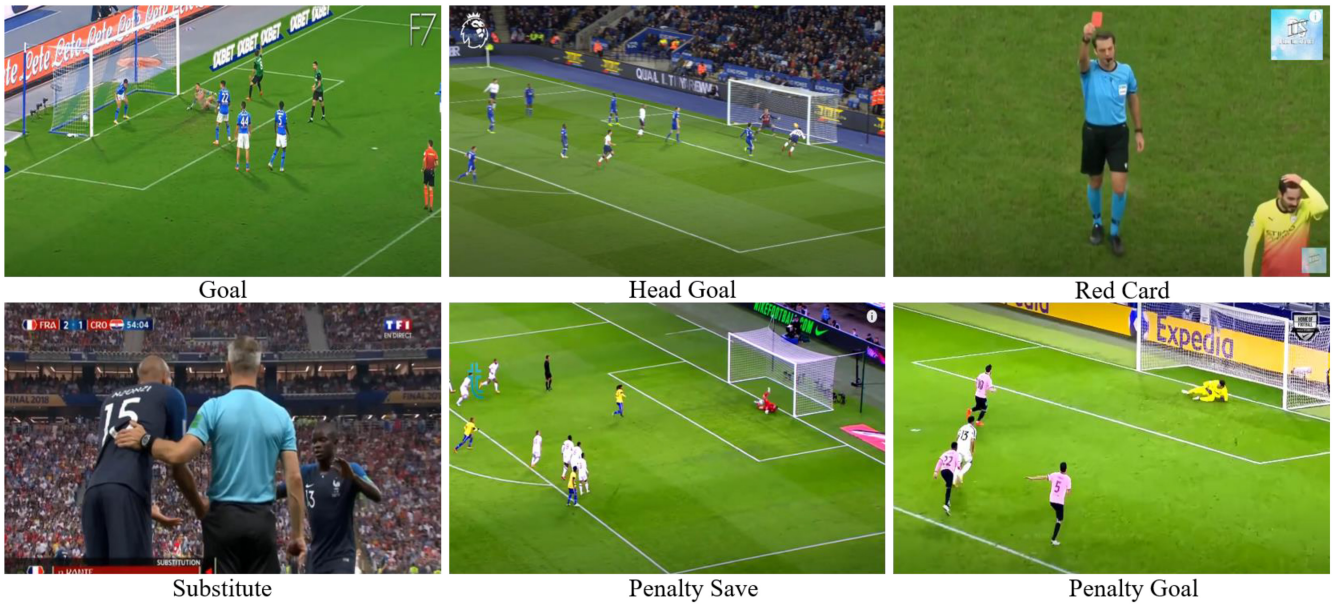


Fig. 3. Sample event classes from our newly created soccer videos data set.

TABLE II
STATISTICS OF TRAINING, VALIDATION, AND TESTING SETS OF THE SVD DATA SET

	Number of videos clips	Video clip type	Duration in seconds	Frame rate (fps)
Training data	360	MP4	3s – 6s	29
Validation data	120	MP4	3s – 6s	29
Testing data	120	MP4	3s – 6s	29

from multiple sources(such as UEFA Champions League, English Premier League, FIFA World Cup 2018, Bundesliga and Primera Division); 2) extract event-specific short video clips from the downloaded soccer videos; and 3) annotate the event-specific video clips with start and end boundary of event. The SVE data set contains a total of 600 short video clips, which are divided into three subsets, including train, validation, and test set with a split ratio of 60%, 20%, and 20%, respectively. The detailed information of data set is presented in Table II, where the representative images for each event of our SVD data set are depicted in Fig. 3.

B. Experimental Analysis

1) Experiment 1 (SVM Classifier With Histogram of Oriented Gradient Features): We have evaluated the SVE data set with a conventional machine learning technique, where we have used a HOG as a feature descriptor and SVM as a classifier to detect the salient events in soccer videos. The HOG descriptor represents gradient orientation and magnitude of objects in a particular region of an image and captures shape-relevant information of detected objects in a video frame. After feature extraction process, we have trained the SVM classifier on extracted features and have evaluated the trained classifier on the test data set. The results obtained from the test data set are shown in Table III, where the diagonal values represent the true positive produce by the SVM classifier. The

TABLE III
CONFUSION MATRIX OF SVM(HOG)

Actual Class	**Predicted Class**					
	Goal	Head Goal	Penalty Save	Penalty Goal	Red Card	Substitute
Goal	18	7	2	2	0	0
Head Goal	2	13	2	1	0	0
Penalty Save	0	0	11	3	0	0
Penalty Goal	0	0	4	14	3	2
Red Card	0	0	0	0	15	1
Substitute	0	0	1	0	5	20
Recall (%)	**90.0**	**65.0**	**55.0**	**70.0**	**75.0**	**100.0**
Precision (%)	**62.0**	**72.0**	**78.0**	**77.0**	**100.0**	**76.0**

precision and recall scores from Table III reveal that there is still a considerable room for the improvement of event recognition rate, especially for Head Goal, Penalty Goal, and Penalty save. These scores can be significantly improved using deep learning techniques such as CNN and RNN.

2) Experiment 2 (Integration of MLSTM With AlexNet Architecture): We have analyzed the soccer event recognition using MLSTM with the AlexNet architecture. First, we have extracted discriminative CNN features using a pretrained AlexNet CNN architecture, and then classified the event types by inputting the extracted features to an MLSTM. For feature extraction, we have used the fully connected layer fc-7 of the pretrained AlexNet model as a generic feature descriptor, which converts a video frame into a 1×4096 feature vector. After the feature extraction process, MLSTM is trained on extracted features. Finally, we have evaluated our trained model on the test data set, where 20 video clips per class are given to the trained model for the event recognition task. The obtained results using this approach are presented in Table IV. From Table IV, we can observe that the recognition rate for the head goal, penalty goal, penalty save, and red card is improved as compared to the results obtained by SVM(HOG) in Experiment 1.

TABLE IV
CONFUSION MATRIX OF OUR SOCCER DATA SET FOR MLSTM + ALEXNET

Actual Class	Predicted Class					
	Goal	Head Goal	Penalty Save	Penalty Goal	Red Card	Substitute
Goal	16	1	0	0	0	0
Head Goal	4	19	1	1	0	0
Penalty Save	0	0	19	0	0	0
Penalty Goal	0	0	0	19	3	2
Red Card	0	0	0	0	17	2
Substitute	0	0	0	0	0	16
Recall (%)	**80.0**	**95.0**	**95.0**	**95.0**	**85.0**	**80.0**
Precision (%)	**94.1**	**76.0**	**100.0**	**79.2**	**89.5**	**100.0**

TABLE V
CONFUSION MATRIX OF OUR SOCCER DATA SET FOR MLSTM + GOOGLENET

Actual Class	Predicted Class					
	Goal	Head Goal	Penalty Save	Penalty Goal	Red Card	Substitute
Goal	18	3	2	0	0	0
Head Goal	2	17	1	0	0	0
Penalty Save	0	0	16	0	0	0
Penalty Goal	0	0	1	20	3	2
Red Card	0	0	0	0	17	1
Substitute	0	0	0	0	0	17
Recall (%)	**90.0**	**85.0**	**80.0**	**100.0**	**85.0**	**85.0**
Precision (%)	**78.3**	**85.0**	**100.0**	**76.9**	**91.8**	**100.0**

3) Experiment 3 (Integration of MLSTM With GoogleNet Architecture): In this set of experiments, we have replaced the AlexNet with a deeper CNN architecture named GoogleNet. It is a deeper architecture with 22 convolution layers and extracts more useful features, which significantly improves the performance of MLSTM for event recognition task. To extract features, we have used the loss3-classifier as a feature descriptor and have obtained the feature vector having a length of 1×1000. Furthermore, MLSTM is trained on extracted features of length 1×1000, and then the performance of trained model is evaluated on the test data. We test 20 video clips per event using trained model. The obtained results are presented in Table V. It can be inferred from Table V that the GoogleNet with MLSTM achieves more or less similar results in terms of precision and recall as obtained in our second set of experiments (i.e., Experiment 2) but improved the event recognition rate with a minor increment of 0.24%.

4) Experiment 4 (Integration of MLSTM With ResNet50 Architecture): Finally, we have evaluated the performance of our ultimate framework, which combined ResNet50 + MLSTM for the recognition of salient events in soccer videos. Our proposed event recognition framework first performs a feature extraction process, where we extract CNN features from video frames using the fully connected layer fc-1000 of ResNet50. After feature extraction, we have trained MLSTM on extracted features. Furthermore, we have performed our model testing, where 20 videos per event have been tested on the trained model. The obtained results are presented in Table VI. It can be observed from Table VI that MLSTM with the ResNet50 architecture not only achieves the best results in terms of precision and recall but also improves the overall accuracy on our SVE data set instances, which are correctly classified. This is indicated by diagonal values, whereas the precision and recall scores are listed at the bottom of Table VI.

TABLE VI
CONFUSION MATRIX OF OUR SOCCER DATA SET FOR MLSTM + RESNET50

Actual Class	Predicted Class					
	Goal	Head Goal	Penalty Save	Penalty Goal	Red Card	Substitute
Goal	18	1	1	0	0	0
Head Goal	2	18	1	0	0	0
Penalty Save	0	1	18	0	0	0
Penalty Goal	0	0	1	20	2	0
Red Card	0	0	0	0	17	1
Substitute	0	0	0	0	1	19
Recall (%)	**90.0**	**90.0**	**90.0**	**100.0**	**85.0**	**95.0**
Precision (%)	**90.0**	**85.0**	**94.7**	**90.9**	**94.4**	**95.0**

C. Overall Performance Comparison

In this section, we have compared the performance of investigated approaches for salient event recognition in experiments 1–4 on test data. Fig. 4 depicts the categorywise accuracy of our four different experiments. The results reveal that ResNet50 + MLSTM, as adopted in our framework, outperforms other approaches for salient soccer event recognition in terms of accuracy. We further compare the presented approaches in terms of TPR, FPR, Precision, Recall and $F1$-score. The results obtained by our proposed framework are presented in Table VII. The evaluation metrics TPR and FPR represent the predicted TPR and FPR of each investigated method in our soccer data set. Furthermore, we calculated the precision and recall measure of each method. Finally, the $F1$-Score is computed using precision and recall. The accuracy measures presented in Table VII validate that our proposed solution (LSTM + ResNet50) dominates all pervious investigated methods in terms of FPR, TPR, Precision, Recall, and $F1$-score. Fig. 5 depicts the predictions of our proposed framework for event recognition in soccer videos (please see the section on "Visual Results" for more details on the predictions of our proposed framework). Furthermore, the training and validation performance of our proposed framework and other investigated techniques are given in Fig. 6. Moreover, we have investigated the performance of the proposed framework using three different fame skip schemes for event recognition in soccer videos. Table VIII presents the statistics of the experiments conducted based on different frame skip strategies. It can be noted in Table VIII that our proposed four-frame-skip strategy shows overwhelming performance improvement over other frame skip strategies (i.e., 8-frame-skip strategy and 6-frame-skip strategy). Therefore, we adopt four-frame-skip strategy in our approach which enables us to achieve reasonable accuracy with acceptable time complexity.

D. Visual Results

We have further evaluated our proposed framework on random videos with predefined events. During evaluation, the

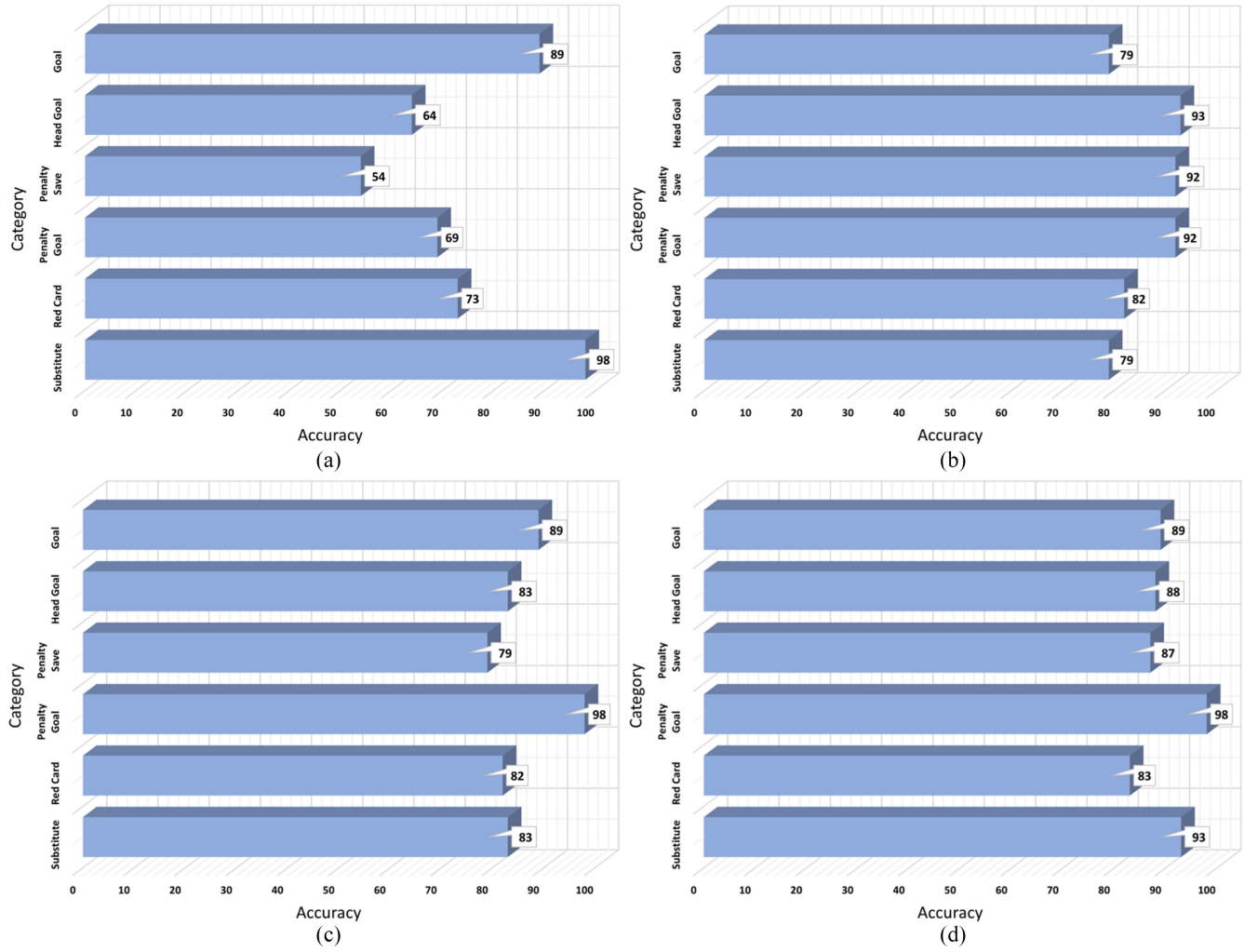


Fig. 4. Categorywise accuracy of four different experimental evaluations of our data set. (a) HOG + SVM, (b) AlexNet + MLSTM, (c) GoogleNet + MLSTM and (d) ResNet50 + MLSTM.

TABLE VII
OVERALL COMPARISON OF THE INVESTIGATED METHODS

Method	FPR (%)	TPR (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM (HOG)	0.13	0.81	75.83	77.5	76.65
MLSTM+ AlexNet	0.08	0.92	88.33	89.80	88.05
MLSTM+ GoogleNet	0.07	0.93	87.50	89.11	88.29
MLSTM + ResNet50	**0.05**	**0.96**	**91.66**	**91.78**	**91.74**

TABLE VIII
ACCURACY OF THE PROPOSED FRAMEWORK ON DIFFERENT FRAME SKIP STRATEGIES

Experiments	Frame Skip	Average Time Complexity (Sec)	Accuracy (%)
8 frame skip strategy	8	0.97	89.31
6 frame skip strategy	6	1.19	90.06
Proposed (4 frame skip strategy)	4	1.43	**91.74**

proposed framework makes prediction for each video that could be correct or incorrect. While testing the input video, our method extracted CNN features with the four-frame-skip strategy. The extracted features are then fed to Multilayer LSTM for analyzing the video sequences and predict the type of event present in the video. In Fig. 5, each row represents specific events, row 3 is misclassified, where "Penalty Save" is classified as "Penalty Goal." This misclassification is due to the visual similarity of contents, motion of the player (i.e., running), and a similar background.

E. Comparison With Exiting Soccer Event Recognition Methods

This section presents the comparative study of our proposed framework with existing soccer event recognition approaches [15], [21], [42]. The results of our proposed framework are evaluated on test videos of our SVE data set. To validate the effectiveness of our proposed framework, we have compared the proposed framework with five existing soccer event recognition methods. The obtained results are shown in Table IX. For a comparison with state-of-the-art methods, we have used accuracy as the evaluation metric. From Table IX, it can be observed that the performance of each method varies

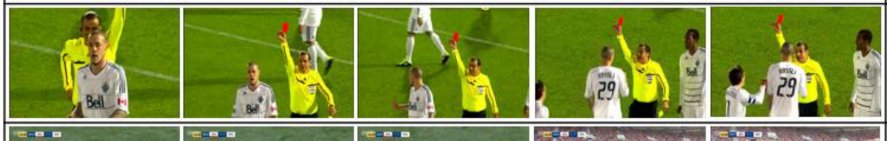
Event specific video frames	Ground Truth	Predictions	Confidence Score
	Goal	Goal	0.78
	Head Goal	Head Goal	0.63
	Penalty Save	Penalty Goal	0.29
	Penalty Goal	Penalty Goal	0.59
	Red Card	Red Card	0.81
	Substitute	Substitute	0.87

Fig. 5. Predictions of our proposed framework for event recognition in soccer videos, where classifications with high confidence values are indicated in green color, classifications with moderate confidence values are in blue color, and misclassified predictions are in red color.

from one event to another. For instance, the method [21] has the best accuracy for "Red card" event. Whereas, our proposed framework dominates the existing soccer event recognition methods, in particular, for detecting "Goal" and "Penalty Save or Penalty attempt" events. Our proposed framework increases the recognition accuracy for "Goal" and "Penalty save or penalty attempt" events by 1.13% and 3.57%, respectively, on average, as compared to the existing methods.

V. Conclusion and Future Work

The advent of smart cameras, Nx-IoT, and efficient learning algorithms for sports video analytics will enhance the performance of players as well as facilitate the live spectators inside the stadium. The smart cameras in the Nx-IoT soccer environment are interconnected through wireless networks, which capture and transmit the data to an AI-assisted computing platform. Majority of the spectator are very enthusiastic to watch and celebrate the better performance of their favorite teams. The IoT-enabled soccer environment will provide the spectators with live information (visual and textual) related to the important events of the match and will allow them to share and discuss the match situation in real time, which can be provided to the spectators as a FinTech service. Therefore, in this article, we have proposed an efficient deep learning-based framework for multiperson salient soccer event recognition in Nx-IoT-enabled environments. The proposed framework recognizes the salient events in the soccer video, including the goal, substitute, penalty save, penalty goal, red card, and head goal. Our proposed framework examines different CNN architectures with multilayer LSTM and proposes an efficient CNN+LSTM approach for soccer event learning and recognition in Nx-IoT-enabled environments. Furthermore, we have developed a new soccer data set SVE, containing six most salient soccer events (i.e., Goal, Red card, Penalty save, Penalty goal, Substitute, Head goal). The results obtained from the experimental evaluation validate the suitability and accuracy of our proposed framework for soccer event recognition in FinTech-enabled Nx-IoT environments.

This article mainly focuses on the recognition of salient soccer events in IoT environment using combined CNN+MLSTM deep learning framework. Although, the current approach uses an efficient CNN architecture in terms of feature enrichment, the series of residual blocks employed in this approach increase the overall computation complexity. Also, the proposed system has no suitable mechanism for ball tracking and player position tracking in the ground field. Moreover, the current system sometime misclassifies Penalty Save event as a Penalty Goal. Considering these limitations of our current method, in future we are aiming to use a light-weight CNN architecture having lower computation complexity. Furthermore, we have intentions to extend our proposed framework by introducing more robust and discriminative features for efficient event recognition task, such

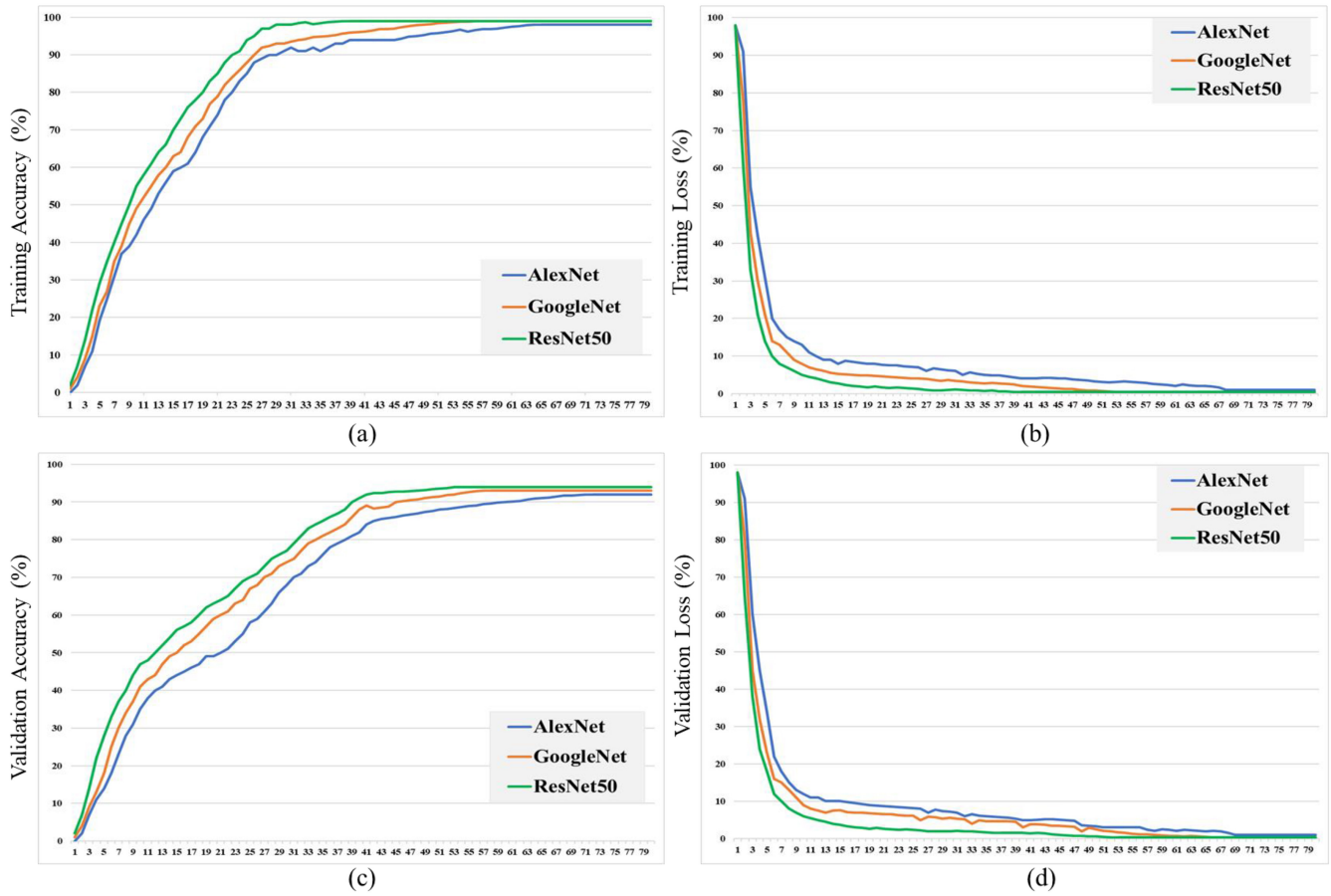


Fig. 6. Training and validation evaluation of our proposed framework along with different investigated techniques for recognition of salient events in soccer videos. (a) Training accuracy, (b) Training loss, (c) Validation accuracy, and (d) Validation loss.

TABLE IX
ACCURACY COMPARISON WITH STATE-OF-THE-ART SOCCER EVENT RECOGNITION METHODS

Method	Goal (%)	Head goal (%)	Penalty goal (%)	Penalty save/attempt (%)	Red card (%)	Substitute (%)
[45]	88.03	-	-	-	90.35	-
[21]	90.29	-	-	86.48	**96.42**	-
[15]	89.68			91.66	86.66	
Our Method	**90.81**	87.42	92.29	**95.23**	89.45	95.0

as optical flow, motion saliency, and C3D features (C3D features are utilized by 3-D ConvNets) along with more robust sequence learning algorithm such as gated recurrent unit (GRU). Furthermore, we plan to integrate other modules, such as player position tracking, player identification, and soccer ball tracking in our current framework for efficient soccer event recognition and streaming in Nx-IoT-enabled environments.

ACKNOWLEDGMENT

Khan Muhammad and Hayat Ullah are with the Department of Software, Sejong University, Seoul 143-747, Republic of Korea (e-mail: khan.muhammad@ieee.org; hayatullah@ieee.org).

Mohammad S. Obaidat is with the College of Computing and Informatics, University of Sharjah, Sharjah, UAE, also with the King Abdullah II School of Information Technology, University of Jordan, Amman 11942, Jordan, and also with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China (e-mail: m.s.obaidat@ieee.org).

Amin Ullah is with the CORIS Institute, Oregon State University, Corvallis, OR 97331 USA (e-mail: ullaham@oregonstate.edu).

Arslan Munir is with the Department of Computer Science, Kansas State University, Manhattan, KS 66506 USA (e-mail: amunir@ksu.edu).

Muhammad Sajjad is with the Digital Image Processing Laboratory, Islamia College University Peshawar, Peshawar 25000, Pakistan, and also with the Norwegian Colour and Visual Computing Laboratory, Department of Computer Science, Norwegian University of Science and Technology, 2815 Gjøvik, Norway (e-mail: muhammad.sajjad@icp.edu.pk; muhammad.sajjad@ntnu.no).

Victor Hugo C. de Albuquerque is with the Department of Teleinformatics Engineering, Federal University of Ceará, Fortaleza 60811-905, Brazil (e-mail: victor.albuquerque@ieee.org).

REFERENCES

[1] Worldatlas. *The Most Popular Sports in the World*. Accessed: Feb. 13, 2020. [Online]. Available: https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html

[2] M. Dive. *Google: Watch Time for YouTube Sports Highlights Jumps 80%*. Accessed: Feb. 14, 2020. [Online]. Available: https://www.marketingdive.com/news/google-watch-time-for-youtube-sports-highlights-jumps-80/516281/

[3] M. H. Kolekar and S. Sengupta, "Bayesian network-based customized highlight generation for broadcast soccer videos," *IEEE Trans. Broadcast.*, vol. 61, no. 2, pp. 195–209, Jun. 2015.
[4] M. G. I. Rathod and M. D. A. Nikam, "Review on event retrieval in soccer video," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 5601–5605, 2014.
[5] A. Ghosh and C. Jawahar, "SmartTennisTV: Automatic indexing of tennis videos," in *Proc. Nat. Conf. Comput. Vis. Pattern Recognit. Image Process. Graph.*, 2017, pp. 24–33.
[6] M. Xu, N. C. Maddage, C. Xu, M. S. Kankanhalli, and Q. Tian, "Creating audio keywords for event detection in soccer video," in *Proc. Int. Conf. Multimedia Expo.*, vol. 2, 2003, pp. 281–284
[7] Q. Ye, Q. Huang, W. Gao, and S. Jiang, "Exciting event detection in broadcast soccer video with mid-level description and incremental learning," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 455–458.
[8] W. Zhao, Y. Lu, H. Jiang, and W. Huang, "Event detection in soccer videos using shot focus identification," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, 2015, pp. 341–345.
[9] A. Raventos, R. Quijada, L. Torres, and F. Tarrés, "Automatic summarization of soccer highlights using audio-visual descriptors," *SpringerPlus*, vol. 4, no. 1, pp. 1–19, 2015.
[10] M.-H. Sigari, H. Soltanian-Zadeh, and H.-R. Pourreza, "A framework for dynamic restructuring of semantic video analysis systems based on learning attention control," *Image Vis. Comput.*, vol. 53, pp. 20–34, Sep. 2016.
[11] B. Fakhar, H. R. Kanan, and A. Behrad, "Event detection in soccer videos using unsupervised learning of Spatio-temporal features based on pooled spatial pyramid model," *Multimedia Tools Appl.*, vol. 78, no. 12, pp. 16995–17025, 2019.
[12] Z. Wang, J. Yu, and Y. He, "Soccer video event annotation by synchronization of attack–defense clips and match reports with coarse-grained time information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 5, pp. 1104–1117, May 2017.
[13] W. Huang and Z. Wang, "Soccer video event detection using 3D convolutional networks and shot boundary detection via deep feature distance," in *Proc. 24th Int. Conf. Neural Inf. Process.*, 2017, pp. 440–449.
[14] M. Z. Khan, S. Saleem, M. A. Hassan, and M. U. G. Khan, "Learning deep C3D features for soccer video event detection," in *Proc. 14th Int. Conf. Emerg. Technol. (ICET)*, 2018, pp. 1–6.
[15] H. Jiang, Y. Lu, and J. Xue, "Automatic soccer video event detection based on a deep neural network combined CNN and RNN," in *Proc. IEEE 28th Int. Conf. Tools Artif. Intell. (ICTAI)*, 2016, pp. 490–494.
[16] G. Yaparla, S. Allaparthi, S. K. Munnangi, G. Ramamurthy, and G. Canaria, "A Novel framework for fine grained action recognition in soccer," in *Proc. Int. Work-Conf. Artif. Nerual Netw.* 2019, pp. 137–150.
[17] S. A. Pettersen *et al.*, "Soccer video and player position dataset," in *Proc. 5th ACM Multimedia Syst. Conf.*, 2014, pp. 18–23.
[18] J. Yu, A. Lei, Z. Song, T. Wang, H. Cai, and N. Feng, "Comprehensive dataset of broadcast soccer videos," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval (MIPR)*, 2018, pp. 418–423.
[19] N. Feng *et al.*, "SSET: A dataset for shot segmentation, event detection, player tracking in soccer videos," *Multimedia Tools Appl.*, vol. 79, no. 39, pp. 28971–28992, Aug. 2020.
[20] H. Ullah and M. Sajjad, "Salient event detection in soccer videos using histogram of oriented gradient," in *Proc. 4th Int. Conf. Next Gener. Comput. (ICNGC)*, 2018, pp. 231–233.
[21] M. Tavassolipour, M. Karimian, and S. Kasaei, "Event detection and summarization in soccer videos using Bayesian network and Copula," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 2, pp. 291–304, Feb. 2014.
[22] B. Fakhar, H. R. Kanan, and A. Behrad, "Learning an event-oriented and discriminative dictionary based on an adaptive label-consistent K-SVD method for event detection in soccer videos," *J. Vis. Commun. Image Represent.*, vol. 55, pp. 489–503, Aug. 2018.
[23] K. J. Bennett, A. R. Novak, M. A. Pluss, A. J. Coutts, and J. Fransen, "Assessing the validity of a video-based decision-making assessment for talent identification in youth soccer," *J. Sci. Med. Sport*, vol. 22, no. 6, pp. 729–734, 2019.
[24] M.-S. Hosseini and A.-M. Eftekhari-Moghadam, "Fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video," *Appl. Soft Comput.*, vol. 13, no. 2, pp. 846–866, 2013.
[25] S. Khan, K. Muhammad, S. Mumtaz, S. W. Baik, and V. H. C. de Albuquerque, "Energy-efficient deep CNN for smoke detection in foggy IoT environment," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9237–9245, Dec. 2019.
[26] K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik, "Efficient deep CNN-based fire detection and localization in video surveillance applications," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 7, pp. 1419–1434, Jul. 2019.
[27] J. Jia *et al.*, "EMBDN: An efficient multiclass barcode detection network for complicated environments," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9919–9933, pp. 1483–1498, Dec. 2019.
[28] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021.
[29] T. Liu, H. Liu, Y.-F. Li, Z. Chen, Z. Zhang, and S. Liu, "Flexible FTIR spectral imaging enhancement for industrial robot infrared vision sensing," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 544–554, Jan. 2020.
[30] W. Ren *et al.*, "Low-light image enhancement via a deep hybrid network," *IEEE Trans. Image Process.*, vol. 28, pp. 4364–4375, 2019.
[31] L. Liu, G. Feng, D. Beautemps, and X.-P. Zhang, "Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition," *IEEE Trans. Multimedia*, vol. 23, pp. 292–305, Feb. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9016100
[32] A. Chowdhury and A. Ross, "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 1616–1629, 2019.
[33] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Activity recognition using temporal optical flow convolutional features and multilayer LSTM," *IEEE Trans. Ind. Electron.*, vol. 66, no. 12, pp. 9692–9702, Dec. 2019.
[34] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments," *Future Gener. Comput. Syst.*, vol. 96, pp. 386–397, Jul. 2019.
[35] T. Tsunoda, Y. Komori, M. Matsugu, and T. Harada, "Football action recognition using hierarchical LSTM," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 99–107.
[36] M. Fani, M. Yazdi, D. A. Clausi, and A. Wong, "Soccer video structure analysis by parallel feature fusion network and hidden-to-observable transferring Markov model," *IEEE Access*, vol. 5, pp. 27322–27336, 2017.
[37] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem, "SoccerNet: A scalable dataset for action spotting in soccer videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 1711–1721.
[38] E. Fenil *et al.*, "Real time violence detection framework for football stadium comprising of big data analysis and deep learning through bidirectional LSTM," *Comput. Netw.*, vol. 151, pp. 191–200, Mar. 2019.
[39] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[42] C.-L. Huang, H.-C. Shih, and C.-Y. Chao, "Semantic analysis of soccer video using dynamic Bayesian network," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 749–760, Aug. 2006.

Khan Muhammad (Member, IEEE) received the Ph.D. degree in digital contents from Sejong University, Seoul, Republic of Korea, in February 2019.

Since March 2019, he has been working as an Assistant Professor with the Department of Software, Sejong University, where he is currently the Director of the Visual Analytics for Knowledge Laboratory. He has registered eight patents in South Korea (seven)/Australia (one) and has contributed more than 150 articles in peer-reviewed journals and conference proceedings in his areas of research. He was recently selected among top 100 000 scientists around the globe by Stanford Researchers List. His research interests include intelligent video surveillance, medical image analysis, information security, video summarization, multimedia data analysis, computer vision, IoT/IoMT, and smart cities.

Hayat Ullah (Graduate Student Member, IEEE) received the B.S. degree in computer science from Islamia College University Peshawar, Peshawar, Pakistan, in 2018, and the M.S. degree in computer science from Sejong University, Seoul, Republic of Korea, in 2021.

He secured a scholarship-based Ph.D. Fellowship with the Intelligent Systems, Computer Architecture, Analytics, and Security Laboratory, Department of Computer Science, Kansas State University, Manhattan, KS, USA, and will officially start his Ph.D. soon. His research interests include image processing, sports video analytics, deep learning, computer vision, image enhancement, and image/video quality assessment.

Mohammad S. Obaidat (Life Fellow, IEEE) received the Ph.D. degree in computer engineering with a minor in computer science from Ohio State University, Columbus, OH, USA, in 1986.

He is an internationally known academic/researcher/scientist/scholar. He has received extensive research funding and published to date (2019) about 1200 refereed technical articles. About half of them are journal articles, over 95 books, and about 70 book chapters. He is currently the Founding Dean and a Professor with the College of Computing and Informatics, University of Sharjah, Sharjah, UAE.

Dr. Obaidat received many best paper awards for his papers. He also received the Best Paper awards from IEEE SYSTEMS JOURNAL in 2018 and in 2019 (2 Best Paper Awards). In 2020, he received four best paper awards from IEEE SYSTEMS JOURNAL. In 2021, he also received the IEEE SYSTEMS JOURNAL Best Paper award. In 2021, he was ranked by Guide2Research as Number 1 Computer Scientist in UE in terms of Number of Publications. He has chaired numerous (over 175) international conferences and has given numerous (over 175) keynote speeches worldwide. He is an SCS Fellow.

Amin Ullah (Associate Member, IEEE) received the Ph.D. degree in digital contents from Sejong University, Seoul, South Korea, in 2021.

He is currently working as a Postdoctoral Researcher with the CoRIS Institute, Oregon State University, Corvallis, OR, USA. He has published several papers in reputed peer-reviewed international journals and conferences, including IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE INTERNET OF THINGS JOURNAL, IEEE ACCESS, *Future Generation Computer Systems* (Elsevier), *Applied Soft Computing* (Elsevier), *International Journal of Intelligent Systems*, *Multimedia Tools and Applications* (Springer), *Mobile Networks and Applications* (Springer), and IEEE Joint Conference on Neural Networks. His major research focus is on human action and activity recognition, sequence learning, image and video analytics, content-based indexing and retrieval, 3-D point clouds, IoT and smart cities, and deep learning for multimedia understanding.

Arslan Munir (Senior Member, IEEE) received the M.A.Sc. degree in electrical and computer engineering from the University of British Columbia, Vancouver, BC, Canada, in 2007, and the Ph.D. degree in electrical and computer engineering from the University of Florida, Gainesville, FL, USA, in 2012.

He is currently an Associate Professor with the Department of Computer Science, Kansas State University, Manhattan, KS, USA. He was a Postdoctoral Research Associate with the Electrical and Computer Engineering Department, Rice University, Houston, TX, USA, from May 2012 to June 2014. His current research interests include embedded and cyber–physical systems, secure and trustworthy systems, parallel computing, artificial intelligence, and computer vision.

Dr. Munir received many academic awards including the doctoral fellowship from Natural Sciences and Engineering Research Council of Canada. He earned gold medals for best performance in electrical engineering, and gold medals and academic roll of honor for securing rank one in pre-engineering provincial examinations (out of approximately 300 000 candidates).

Muhammad Sajjad received the master's degree from the Department of Computer Science, College of Signals, National University of Sciences and Technology, Rawalpindi, Pakistan, in 2012, and the Ph.D. degree in digital contents from Sejong University, Seoul, South Korea, in 2015.

He is currently working as an ERCIM Research Fellow with Norwegian University of Science and Technology, Trondheim, Norway. He is an Associate Professor with the Department of Computer Science, Islamia College University Peshawar, Pakistan. He is also the Head of the Digital Image Processing Laboratory, Islamia College University Peshawar, where many students are involved in different research projects under his supervision, such as big data analytics, medical image analysis, multimodal data mining and summarization, image/video prioritization and ranking, fog computing, Internet of Things, autonomous navigation, and video analytics. He has published more than 65 papers in peer-reviewed international journals and conferences. His primary research interests include computer vision, image understanding, pattern recognition, robotic vision, and multimedia applications, with current emphasis on economical hardware and deep learning, video scene understanding, activity analysis, fog computing, Internet of Things, and real-time tracking.

Dr. Sajjad is serving as a professional reviewer for various well-reputed journals and conferences. He is currently an Associate Editor for IEEE ACCESS and acting as a Guest Editor for IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

Victor Hugo C. de Albuquerque (Senior Member, IEEE) received the graduated degree in mechatronics engineering from the Federal Center of Technological Education of Ceará (CEFETCE), Fortaleza, Brazil, in 2006, the M.Sc. degree in teleinformatics engineering from the Federal University of Ceará, Fortaleza, in 2007, and the Ph.D. degree in mechanical engineering from the Federal University of Paraíba, Joao Pessoa, Brazil, in 2010.

He is a Professor and a Senior Researcher with the Department of Teleinformatics Engineering/Graduate Program on Teleinformatics Engineering, Federal University of Ceará. He is a specialist, mainly, in image data science, IoT, machine/deep learning, pattern recognition, automation and control, and robotics.