FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 38, Nº 1, March 2025, pp. 39 - 52
https://doi.org/10.2298/FUEE2501039R

Original scientific paper

CLASSIFICATION OF SPORTS VIDEOS BY COMBINING RESNET50 MODEL AND FINE TUNING

Pinku Ranjan[1], Jayant Kumar Rai[1], Vaibhav Singh[1], Anand Sharma[2], Somesh Kumar[1]

[1]Department of Electrical and Electronics Engineering,
ABV- Indian Institute of Information Technology and Management, Gwalior (M.P.), India
[2]Department of Electronics and Communication,
Motilal Nehru National Institute of Technology Allahabad, Prayagraj, India

ORCID iDs:	Pinku Ranjan	https://orcid.org/0000-0002-1422-5943
	Jayant Kumar Rai	https://orcid.org/0000-0002-4183-7259
	Vaibhav Singh	N/A
	Anand Sharma	https://orcid.org/0000-0001-8566-1710
	Somesh Kumar	N/A

Abstract. *This article represents a classification of sports videos by integrating the ResNet50 model and fine-tuning methods. Broadcasting companies give significant importance to the classification of sports videos. That is a subclass of recognition of human action, which will clarify the context of videos. This work uses a deep neural network-based ResNet50 model with a fine-tuning technique for classifying popular sports types in India into their corresponding classes. This paper considers 14 main sports - badminton, basketball, boxing, cricket, football, hockey, kabaddi, swimming, shooting, table tennis, tennis, volleyball, weight lifting, and wrestling. The dataset is created to focus on sports action-based classification. Fine-tuning is nothing but networking surgery. First, a pre-trained Convolutional Neural Network model will be loaded, and then fine-tuning (network surgery) will be applied. The base model (ResNet50) will be frozen, so it will not be trained via backpropagation. After fine-tuning, the classifier will be ready to correctly classify a sports video into its category. The training accuracy of the proposed classifier is 91.73%, and testing is done on sports videos. The classifier classifies each sports video into its class correctly. A descent confusion matrix has been pertained.*

Key words: *Deep Learning, ResNet50, fine-tuning, Computer Vision, Sports Videos Classification*

Received June 4, 2024; revised August 16, 2024 and September 27, 2024; accepted November 10, 2024
Corresponding author: Pinku Ranjan
Department of Electrical and Electronics Engineering, ABV- Indian Institute of Information Technology and Management, Gwalior (M.P.), India
E-mail: pinkuranjan@iiitm.ac.in

1. INTRODUCTION

Sports are a significant section of transmitting channels like the internet, television, etc., and many sports videos flood everyday data servers. Broadcasting companies need human resources to identify each sport manually, and it's tough to index/identify solitary sports manually based on their class. Automating the task of classifying each sports video [1-4], will make it easier for broadcasting companies to manage their activity. Searching for any sports video can be done without tedious manual work. This will also help game coaches analyze a particular sports video from tons of sports video archives and help in strategy [5-9]. To classify any video, we need to analyze it sequentially; we analyze the context of a scene. Many attempts have been made to classify video information concerning the scene context [13-20]. To solve a complex task in computer vision, Convolution Neural Network (CNN) or deep learning models are very effective and are mostly used to solve computer vision problems (problems related to images and videos) [21-25]. For recognition of any sports type, only a set of actions is needed by a human being. Sometimes, only surrounding elements are sufficient to recognize a sports type. Sports videos are a sequence of images that develop a sports video classifier to classify them into multiple sports classes. The system needs to consider spatial and temporal information [26-30]. CNN is mostly used to extract spatial features [5] since it is very effective in extracting spatial features and to handle temporal features, recurrent neural networks are used because of their memory gates. These two models can be combined to analyze and recognize visual patterns, but combining CNN and RNN requires more computational resources. Now coming to some CNN architectures, such as VGG, AlexNet, GoogleNet, ResNet, etc., are multilayered neural networks designed to analyze and recognize visual patterns straightaway from pixels of images, requiring less computational resources.

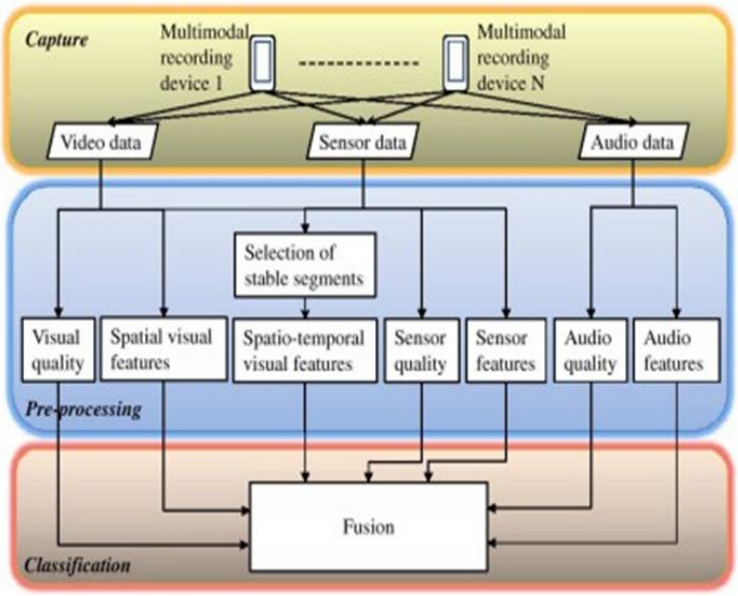

Fig. 1 Overview of the approach [3]

Fine-tuning is applied to the CNN (ResNet50) model because ResNet50 is a pre-trained model on the Imagenet dataset. To adjust a pre-trained model—which has learned generic characteristics from a huge dataset (like ImageNet)—to a particular job using new data, CNNs like ResNet50 must be fine-tuned. Fine-tuning reduces training time and increases performance compared to starting from scratch, which takes a lot of effort and enormous datasets. We allow the model to acquire new, task-specific characteristics while utilizing its pre-existing knowledge of fundamental picture properties by freezing the early layers and updating the later ones. This method makes training faster, more accurate, and requires less data.

The novelty of the proposed work is as follows:

(i) Integrating ResNet50 with Fine-Tuning: To categorize sports images, especially those related to Indian sports, this work interestingly combines the ResNet50 model with fine-tuning approaches.

(ii) Sports Action Classification Dataset: A new dataset on classifying 14 prominent sports in India according to sporting actions has been generated.

(iii) High Accuracy: The proposed classifier achieves an impressive training accuracy of 91.73%, highlighting its effectiveness in classifying sports videos.

(iv) Focussing on Recognising Sports Action: The work emphasizes sports video categorization, which is a subclass of human action recognition.

(v) Freeze and Fine-Tune Approach: The model is specialised for sports video classification, providing a new technique in this area. It is achieved by freezing the underlying ResNet50 model and fine-tuning the top layers.

Related works are described in section II, the collection of datasets is discussed in section III, and section IV describes methodology, results, and conclusion in sections V and VI, respectively.

2. Related Works

Various methodologies have been used to classify sports and TV shows. These methodologies differ based on input data types like image, audio, video, etc [6]. Video classification is quite like image classification. Some deep learning enthusiasts are quick and treat video classification as simple, incorrect image classification. Since a video contains spatial and temporal features. One can understand a video as a series of images. Therefore, most deep learning enthusiasts consider a video classification task as an image classification task by considering N number of frames in a video. In [2], the authors have developed three different methods to classify sports videos - two types of neural networks and texture code cues. They used their combination and performed best on neural net cues. They have only considered five different types of sports - Tennis, Cycling, Track events, Swimming and Yachting. In [4], Support Vector Machine (SVM), and Random Forest (RF) techniques have been used to classify broadcasting shows. They have classed broadcasting shows into general categories such as talk shows, sports, news, movies, cartoons, politics, animation, and, more clearly, summer and winter sports. They took 1250000 frames of general categories and 3250000 frames of summer and winter sports and got 0.80 and 0.77 F1 scores for general categories and summer and winter sports, respectively, using SVM. The F1 scores are quite good, but as the dataset is large enough, there might be a chance to improve the F1 scores. A random forest (RF) with 50 trees was used. The number of trees can be increased in the RF, which makes the RF more

stable and robust. In [3], the authors said that extracting semantic information from mobile videos is difficult because of their unconstrained nature. The domain knowledge of sports videos recorded by multiple users is extracted, and after that, the classification of sports videos into soccer, basketball, football, tennis, ice hockey, or volleyball.

In this paper, the approach used by authors was multi-user and multimodal, as shown in Fig. 1. Since in this paper, the authors have considered only six different sports types. In [7], the authors combined audio and video features to classify sports videos. First, MFCC (Mel Frequency Cepstral Coefficients) is extracted from audio, and then PCA is used to reduce the features' dimensions. K- nearest neighbor (KNN) classifier is used to classify sports types. K-fold cross-validation with K=10, a correct classification rate of 96.11%, is obtained with multimodal features. They got a satisfactory result, though only three types of sports (Soccer, Basketball, volleyball) were considered. The authors used the Hidden Markov Model (HMM), a statistical Markov Model [8]. The overview of the proposed framework is given below in Fig. 2. In this paper, the authors have proposed a framework to recognize sports events. Training of neural networks becomes more difficult as it becomes deeper. So, in this paper [9], a residual learning framework has been presented to make training neural networks easier. In [5], a deep learning approach is used to classify the sports videos. The authors combined a convolutional neural network (CNN) and a recurrent neural network (RNN). CNN was used for spatial feature extraction and RNN for temporal feature extraction. They got a decent accuracy of 96.66 % but considered only five different sports classes - football, cricket, tennis, basketball, and ice hockey. The combination of CNN and RNN deep learning model is a resource-hungry deep neural network model that requires RAM of 32 GB, a GPU of 12 GB, and a very good processor. Classification of Soccer videos and events are two different tasks. In [10], the authors combined these two tasks and solved them simultaneously using a pre-trained CNN model and transfer learning, achieving an accuracy of 89%. Here, the researchers only consider soccer sports. In [11], seven different sports are considered - ski, football, futsal, basketball, box, swimming, and tennis. The authors proposed an ensemble classifier to classify sports videos. The ensemble classifier was built by combining 4 classifiers - linear discriminant analysis, nearest neighbor, probabilistic neural network, and decision tree. Literature has already been made to classify sports videos into their corresponding class. Some of the research [3], [4], [7], [8] have been done using traditional machine learning models, and in this research, a smaller number of different sports types are considered. Some consider only general categories of TV shows; some consider three different sports and some 5 or 7 different sports. After that, some researchers developed a deep learning approach [5], [10] to classify sports videos. They have considered only several different sports types, some of which have considered event classification in a particular sport [10].

3. DATASET COLLECTION

The dataset is crucial to any machine learning/deep learning method. The performance/ accuracy of any deep learning/machine learning method depends on the dataset on which the model will be trained. In this work, 14 different types of sports are taken into consideration. So, we have collected the dataset of 14 different types of sports. Google Image Search 2 dataset (gi2ds) is used to download the datasets of different sports classes. gi2ds queries the image from Google Image and is used to build an image dataset. When we

query a particular image through gi2ds, it shows all images related to the query. We need to scroll down the current window to capture the URLs of the images. URLs of the images are saved in a .txt file, and this .txt file is used by Python code to download the images. The following Fig. 3 (grid of images) shows the sports that have been considered.

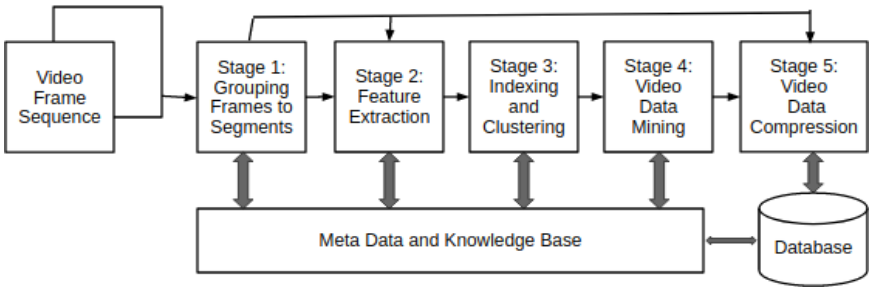

Fig. 2 Overview of the proposed framework [8]

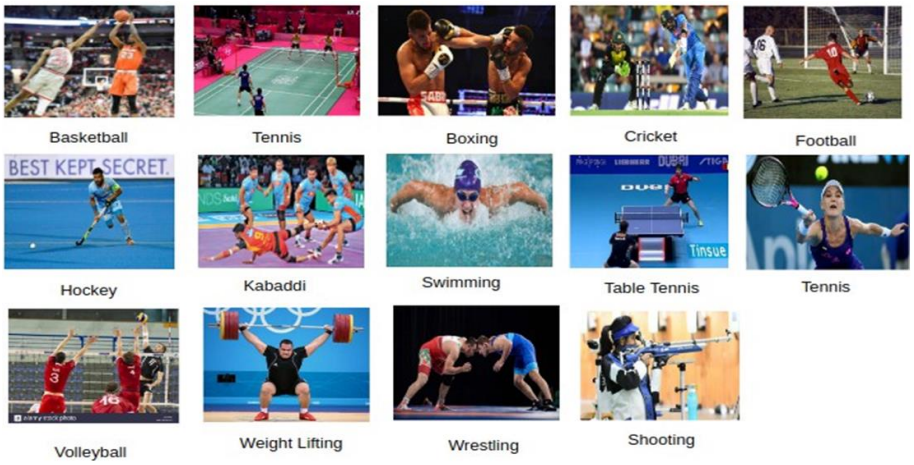

Fig. 3 Dataset of different sports

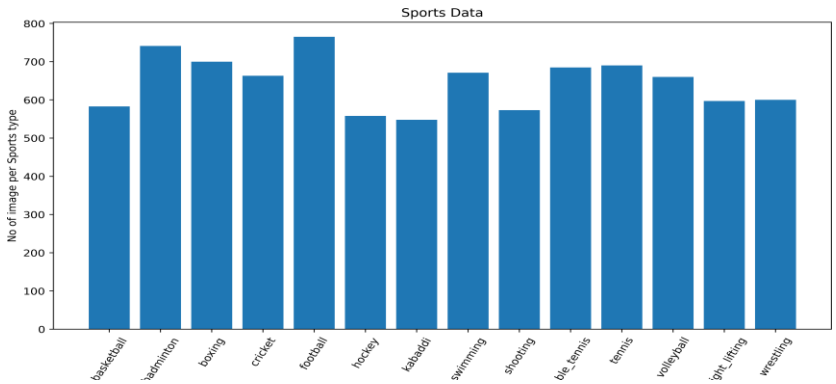

Fig. 4 Distribution of data for different sports

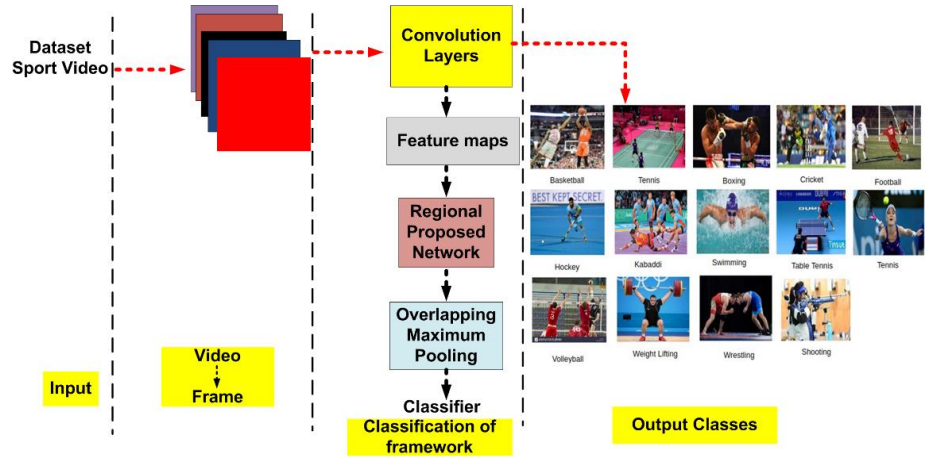

Fig. 5 The proposed flowchart for different classes

Fig.6 The flowchart for the prediction of accuracy

3.1. ResNet50 model and its optimization process

ResNet50 is a deep convolutional neural network (CNN) that enhances image identification tasks and is designed with 50 layers. Using "residual connections," or shortcuts that let data bypass some levels of processing, distinguish ResNet50 from other models. This contributes to the resolution of the "vanishing gradients" issue, which causes networks' performance to deteriorate with depth. Even with deep designs, ResNet50 retains high performance by omitting some layers.

Optimization Process:

The optimization process of the ResNet50 model is as follows: -

Pre-training:

Large datasets like ImageNet are typically used to pre-train ResNet50, teaching it to recognize common characteristics like shapes, edges, and textures.

Fine-Tuning:

The model is fine-tuned for a particular goal, such as sports video categorization. How to do it is as follows: The generic features-capturing early layers are frozen, meaning they are not modified throughout training. To learn features particular to the new dataset, the subsequent layers, which oversee task-specific characteristics, use backpropagation to adapt and unfreeze.

Backpropagation:

The weights of the unfrozen layers are modified throughout this optimization procedure. It computes the error that results from comparing the predicted label with the actual model label, propagates this error backward through the network, and adjusts the weights to minimize the error.

Gradient Descent:

Gradient descent is the optimization process used in backpropagation, which iteratively modifies the model weights to lower the loss function and enhance the model's task-specific performance.

ResNet50 gets optimized for the particular problem by freezing the early layers and fine-tuning the later ones while still utilizing its prior knowledge, which expedites and improves the optimization process.

3.2. Preprocessing of dataset

The collected dataset cannot be fed to the model directly. Cleaning and preprocessing need to be done before feeding it to the model. The downloaded dataset has some irrelevant images, so these images need to be deleted. Irrelevant images are deleted manually. ResNet50 is a pre-trained model trained on an ImageNet dataset of size (224×224). Therefore, the images are converted into (224×224) size. A large amount of data leads to better performance of a deep neural network. So, we need to increase the number of images for each sports type. A tool known as data augmentation is used to increase the size of the dataset. Using a data augmentation tool, the number of images for each sports type increased. Shifting, zooming, and flipping have been used for data augmentation rotation. This study can further train the model, as the data is sufficient. The Fig. 4 represents the distribution of data for each sports class. The graph shows that each class is evenly distributed. The proposed flowchart for different classes is shown in Fig. 5. The flowchart for the accuracy prediction is shown in Fig. 6.

4. METHODOLOGY

This section shows a proposed novel approach to classify sports videos in Fig. 4. Each pixel's width, height, and depth make up the three dimensions of the CNN's input layer. Meanwhile, depth represents the RGB color channel, and width and height indicate the horizontal and vertical pixels. We have converted the raw sports video collection into frames to minimize computational complexity. Taking part in network training. Video classification is quite like image classification, considering the number of subsequent frames in a video. However, video classification is different from simple image classification. In video classification, we make some assumptions, such as, based on the semantic contents, subsequent frames are correlated in a video. The advantage of the temporal nature of videos can be considered to improve the accuracy of the actual video classifier. The proposed solution will use the ResNet50 model to classify sports videos [12]. Resnet50 neural network contains 50 layers. ResNet50 is one of the categories of (CNN) trained on more than a million images taken from the ImageNet database [9]. As an outcome, the ResNet50 network has learned rich feature depictions for a broad range of images. ResNet50 is deeper than Visual Geometry Group (VGG) networks, but ResNet50 has lower complexity than VGG networks 2.

4.1. ResNet Model

The network depth is important to represent and generalize the feature [13]. Adding the convolutional layers to extend the depth of a model will not help achieve better training and generalization performance [14]. To build a deeper convolutional neural network, Kaming et al. Introduced Residual Networks [9]. A residual block of the ResNet model is shown in Fig. 7, which can be assumed as mapping function H(x). Here, the input to the first layer is represented by x, and a residual aligning function is represented by F(x), which can be represented mathematically as

$$F(x) = H(x) - x \tag{1}$$

$$H(x) = F(x) + x \tag{2}$$

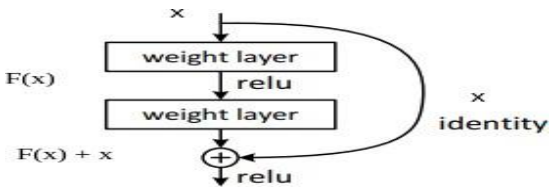

Fig. 7 Residual block

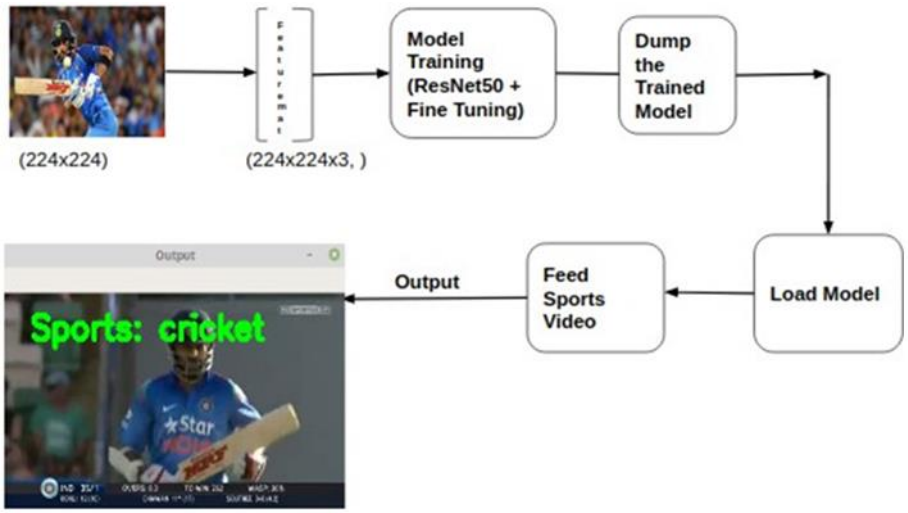

Fig. 8 Block diagram of the proposed methodology

The shortcut connections are known as identity functions (mapping). When the models get deeper, the problem of model degradation arises. Deep residual networks can effectively solve the problem of model degradation. When fine-tuning the ResNet, we set trainable as equal to false, which results in F(x) = 0. By doing this, we can use the pre-trained weight of the ResNet model.

4.2. Fine-tuning

Fine-tuning the pre-trained ResNet50 model to this goal of sports activity recognition from video frames entails adjusting the model for 14 sports videos. Using the fine-tuning method is explained in detail below:

(i)Pre-trained ResNet50 as a Feature Extractor

A large dataset such as ImageNet was used to pre-train the ResNet50 model, teaching it to extract common visual characteristics like edges, textures, and patterns. Because the lowest layers of ResNet50 (which identify fundamental characteristics) are probably beneficial across a broad range of visual tasks, including recognizing human motions in sports, we employ this pre-trained model for sports video classification.

(ii) Freezing the Base Layers of ResNet50

Basic picture properties, such as edges and textures, are captured by the lowest layers of ResNet50 and are useful for various tasks. The weights of these layers do not change throughout training since they stay frozen. This guarantees the preservation of the model's general-purpose feature extraction capabilities. When working with smaller datasets, freezing layers lowers the computational burden and the chance of overfitting.

(iii) Backpropagation:

Only the weights of these recently inserted layers are adjusted via backpropagation during training; the frozen layers remain unaltered. Here, it's important to make sure the model picks up on mapping sports-specific attributes—like motions or behaviors common to several sports—to the appropriate sport.

(iv) Parameter Optimization

The learning rate of the recently inserted layers is often configured to be higher than that of the frozen layers. This keeps the pre-trained weights in the frozen layers intact while enabling the model to modify the new layers' parameters efficiently. Certain deeper levels of ResNet50 can be gradually unfrozen once the new layers have been trained, providing additional model optimization. This makes the model more flexible and able to adjust to the unique subtleties of categorizing sports videos.

(v) Training the Model

The collection comprises video frames, or frame sequences, from fourteen different sports, including cricket, badminton, basketball, boxing, and more. The ResNet50-based model is fed these preprocessed frames to be classified. Every video has a label indicating the sport it is associated with. Using the sports dataset, the refined model is trained. Throughout the training process, the model gains the ability to link specific visual patterns, motions, and situations to the corresponding sports categories. A validation set is employed to keep an eye on the model's performance and adjust hyperparameters (such as batch size and learning rate) to avoid overfitting.

(vi) Performance Evaluation

The model is assessed using an independent test set of sports videos after training. The accuracy of the classifier is measured; in this instance, it is 91.73%, meaning that the model mostly properly identifies sports footage.

4.3. Block diagram of methodology

The proposed methodology is shown in Fig. 8. This block diagram shows how the proposed solution will be implemented for sports video classification. First, the images are resized into (224 × 244) shapes, and the resize () function from computer vision (cv)

is used for resizing the images. These images are converted into a feature matrix. The size of the feature matrix is (224×224×3). Now, the pre-trained ResNet50 model has been loaded as a base model, and then the head of the model has been constructed and will be put on top of the model. The head consists of the following layers: average pooling, flattening, dropout, and dense. The base model has been frozen, so it will not be trained during backpropagation. Now, the feature matrix is fed to the model. The model will be trained by learning from the feature. After training the model, the trained model is dumped into the drive. While training the model, the ReduceLR On Plateau ( ) function is used to handle when the model gets stuck on local minima. SGD (Stochastic gradient descent) function is used as an optimization function with various learning rates and decay rates to get a decent performance.

5. Results and Discussion

In this section, the model will be trained. For training the model, Google Colab, a free cloud service that provides 12 GB RAM, 108 GB storage, and 12 GB NVIDIA Tesla K80 GPU that can be used continuously for up to 12 hours, has been used. Confusion matrix, precision-recall, and f1 score are used as performance metrics. During the model's training, many experiments were conducted with the different values of hyperparameters. The best result is obtained from all the experiments using the hyperparameters listed in Table 1. The dataset is split into an 80:20 ratio; 80% of the data is used as a training set, and 20% as a validation set. The model has been trained for 100 epochs. The history of how the training accuracy and validation accuracy vary concerning epoch number and how the loss varies are shown in Fig. 9.

5.1. Evaluation Parameter

Precision, recall, and f1-score are good metrics for measuring the performance of a classification model and are listed in Table 2. The classifier is trained for 14 popular sports types in India. A confusion matrix is to describe the prediction summary of a classification model. The obtained confusion matrix is shown in Fig. 10. Precision, recall, and f1-score are calculated [30-33] through equations (3) to (5):

(i) Precision (P): It is the ratio of accurately identified classes to all classes.

$$P = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{3}$$

(ii) Recall (R): The percentage of actual classes detected in the video compared to the overall number of classes.

$$R = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{4}$$

(iii) F1 Score: The F1 score represents the harmonic mean of recall and precision.

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{5}$$

Now, the prediction is done on sports videos. The dumped model has loaded and a particular sports video is fed to the model. Prediction is done for each frame which leads to label flickering. So, to handle the flickering problem, the prediction results are stored in a list, and the final output is the class that is in the majority. Here, the prediction for the cricket video is shown in Fig. 11. Table 3 compares the obtained results with some other existing results.

Table 1 Hyperparameters

Hyperparameter	Value
Learning rate	0.001
Decay rate	0.001/epochs
Patience level	5
Factor to reduce the learning rate	0.3
Batch size	32

Table 2 Precision, Recall, and f1-score

Sports class	Precision	Recall	f1-score	Support
Badminton	0.88	0.81	0.84	148
Basketball	0.80	0.92	0.86	117
Boxing	0.95	0.94	0.95	140
Cricket	0.92	0.84	0.88	133
Football	0.82	0.91	0.86	153
Hockey	0.86	0.76	0.81	112
Kabaddi	0.80	0.94	0.86	109
Shooting	0.94	0.90	0.92	115
Swimming	0.96	0.96	0.96	134
Table Tennis	0.94	0.89	0.91	137
Tennis	0.84	0.87	0.85	138
Volleyball	0.89	0.88	0.89	132
Weightlifting	0.93	0.89	0.91	119
Wrestling	0.87	0.85	0.86	120

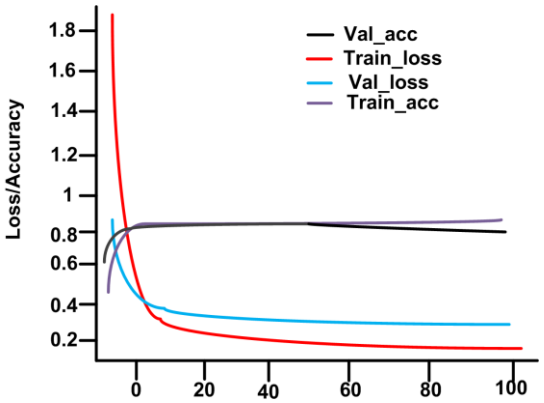

Fig. 9 Training accuracy and loss of the dataset

predictions	0	1	2	3	4	5	6	7	8	9	10	11	12	13	All
actual															
0	120	0	1	0	1	3	3	2	0	3	10	4	0	1	148
1	0	108	0	0	0	0	2	0	0	0	0	1	1	5	117
2	0	5	132	0	0	0	0	0	0	0	0	0	1	2	140
3	1	0	0	112	13	4	1	0	0	0	2	0	0	0	133
4	1	1	0	5	139	3	0	0	1	0	2	1	0	0	153
5	3	1	2	4	9	85	4	0	0	1	2	0	0	1	112
6	0	2	0	1	1	0	102	0	0	1	0	0	1	1	109
7	1	2	0	0	0	1	2	103	0	1	1	2	2	0	115
8	0	0	1	0	0	0	0	1	129	0	0	2	0	1	134
9	2	0	0	0	2	0	1	1	0	122	5	1	2	1	137
10	7	0	1	0	4	1	1	0	3	0	120	1	0	0	138
11	2	5	0	0	1	1	5	0	0	1	0	116	0	1	132
12	0	1	1	0	0	0	3	3	1	1	0	1	106	2	119
13	0	10	1	0	0	1	3	0	0	0	1	1	1	102	120
All	137	135	139	122	170	99	127	110	134	130	143	130	114	117	1807

Fig. 10 Confusion matrix

Fig. 11 The prediction for cricket video

Table 3 Comparison of the proposed work with other existing work

Ref.	Dataset	Model	Optimization Techniques	Class	Accuracy	Output
[23]	Open Video Project and YUV Video Sequences	Not Given	Graph clustering	Not Given	Not Given	Video summarization
[24]	Not Given	SVM, and CNN	K-mean, K-medoid	Two classes	90%	An audio talk show
[25]	Soccer Dataset for Shot, Event, and Tracking	DevNet, VGG, LSTM GoogLeNet	Not Given	Multiclass	Not Given	Shot segmentation, event detection, player tracking
[26]	YouTube, Cricinfo	Deep learning	K-means clustering	Multiclass	Not Given	Frames extracted
[27]	Olympic Games event image	AlexNet, VGG-16, ResNet-50	Transfer learning	Multiclass	90%	Olympic event identification
[28]	Sports	CustomizedCNN	SGDM	Multiclass	89.75	Keyframes
This Work	14 different Sport event	Deep learning with fine-tuning	Transfer learning	Multiclass	91.73	14 sports type (popular in India)

6. CONCLUSION

In this paper, a sports video classifier has been developed using a pre-trained deep-learning model. Fine-tuning is used, which helps in using the pre-trained weight of the model instead of training it from scratch. The proposed sports videos classifier correctly classified the sports videos (badminton, basketball, boxing, cricket, football, hockey, kabaddi, swimming, shooting, table tennis, tennis, volleyball, weight lifting, and wrestling). Image data have been used for training while the prediction is done on the sports videos, and it found that the classifier can classify the sports correctly. This classifier helps broadcasting companies handle their sports data in a well-organized manner.

REFERENCES

[1] D. Brezeale and D. J. Cook, "Automatic video classification: A survey of the literature", *IEEE Trans. Syst. Man. Cybern. Part C (Applications and Reviews)*, vol. 38, no. 3, pp. 416-430, 2008.

[2] K. Messer, W. Christmas and J. Kittler, "Automatic sports classification", In Proceedings of the IEEE International Conference on Pattern Recognition, Quebec City, QC, Canada, 2002, vol. 2, pp. 1005-1008.

[3] F. Cricri, M. J. Roininen, J. Leppanen, S. Mate, I. D. Curcio, S. Uhlmann and M. Gabbouj, "Sport type classification of mobile videos", *IEEE Trans. Multimedia*, vol. 16, no. 4, pp. 917-932, 2014.

[4] P. Campr, M. Herbig, J. Vanek and J. Psutka, "Sports video classification in continuous TV broadcasts", In Proceedings of the 12th IEEE International Conference on Signal Processing (ICSP), 2014, pp. 648-652.

[5] M. A. Russo, A. Filonenko and K.-H. Jo, "Sports classification in sequential frames using CNN and RNN", In Proceedings of the IEEE International Conference on Information and Communication Technology Robotics (ICT-ROBOT), 2018, pp. 1-3.

[6] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho and J. Huopaniemi, "Audio-based context recognition", *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 1, pp. 321-329, 2005.

[7] R. Gade, M. Abou-Zleikha, M. Græsbøll Christensen and T. B. Moeslund, "Audio-visual classification of sports types", In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 51-56.

[8] V. Ellappan and R. Rajasekaran, "Event recognition and classification in sports video",” In Proceedings of the 2017 Second IEEE International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), 2017, pp. 182-187.
[9] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.
[10] Y. Hong, C. Ling and Z. Ye, "End-to-end soccer video scene and event classification with deep transfer learning", In Proceedings of the IEEE International Conference on Intelligent Systems and Computer Vision (ISCV), 2018, pp. 1-4.
[11] M. H. Sigari, S. A. Sureshjani and H. Soltanian-Zadeh, "Sports video classification using an ensemble classifier", In Proceedings of the 7th Iranian IEEE Conference on Machine Vision and Image Processing, 2011, pp. 1-4.
[12] Y. Xing, C. Lv, H. Wang, D. Cao, E. Velenis and F.-Y. Wang, "Driver activity recognition for intelligent vehicles: A deep learning approach", *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5379-5390, 2019.
[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", *arXiv preprint*, arXiv:1409.1556, 2014.
[14] R. K. Srivastava, K. Greff and J. Schmidhuber, "Highway networks", *arXiv preprint*, arXiv:1505.00387, 2015.
[15] A. Ekin, A. M. Tekalp and R. Mehrotra, "Automatic soccer video analysis and summarization", *IEEE Trans. Image Process.*, vol. 12, no. 7, pp. 796-807, 2003.
[16] H. Jiang, Y. Lu and J. Xue, "Automatic soccer video event detection based on a deep neural network combined CNN and RNN",” In Proceedings 28th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2016, pp. 490-494.
[17] S. J. Pan and Q. Yang, "A survey on transfer learning", *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345- 1359, 2009.
[18] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in *Advances in neural information processing systems*, 2012, pp. 1097-1105.
[19] C. K. Mohan and B. Yegnanarayana, "Classification of sport videos using edge-based features and autoassociative neural network models", *Signal Image Video Process.*, vol. 4, no. 1, pp. 61-73, 2010.
[20] C. Xu, J. Cheng, Y. Zhang, Y. Zhang and H. Lu, "Sports video analysis: Semantics extraction, editorial content creation and adaptation", *J. Multimedia*, vol. 4, no. 2, pp. 69-79, 2009.
[21] J. Wang, C. Xu and E. Chng, "Automatic sports video genre classification using pseudo-2D-HMM", Jinjun Wang, Changsheng Xu and E. Chng, "Automatic Sports Video Genre Classification using Pseudo-2D-HMM," In Proceedings of the IEEE 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 2006, pp. 778-781.
[22] L. Li, N. Zhang, L.-Y. Duan, Q. Huang, J. Du and L. Guan, "Automatic sports genre categorization and view-type classification over large-scale dataset",” in Proceedings of the 17th ACM International Conference on Multimedia, 2009, pp. 653-656.
[23] V. Chaudhary, Rashmi and V. Uniyal, "An effective video noise removal algorithm", *Int. Res. J. Eng. Technol. (IRJET)*, vol. 3, no. 8, pp. 2031-2034, 2016.
[24] X. Yunjun, "A sports training video classification model based on deep learning",” *Scientific Programming*, vol. 2021, p. 7252896, 2021.
[25] S. M. Daudpota, A. Muhammad and J. Baber. "Video genre identification using clustering-based shot detection algorithm", *Signal, Image and Video Process.*, vol. 13, pp. 1413-1420, 2019.
[26] S. Zhang, "Detection of aerobics action based on convolutional neural network", *Comput. Intell. Neurosci.*, vol. 2022, p. 1857406, 2022.
[27] Y. I. Mohamad, S. S. Baraheem and T. V. Nguyen, "Olympic games event recognition via transfer learning with photobombing guided data augmentation", *J. Imaging*, vol. 7, no. 2, p. 12, 2021.
[28] M. Ramesh and K. Mahesh, "Sports Video Classification Framework Using Enhanced Threshold Based Keyframe Selection Algorithm and Customized CNN on UCF101 and Sports1-M Dataset", *Comput. Intell. Neurosci.*, vol. 2022, p. 218431, 2022.
[29] N. Feng et al., "SSET: a dataset for shot segmentation, event detection, player tracking in soccer videos." *Multim. Tools Appl.*, vol. 79, pp. 28971-28992, 2020.
[30] M. Tabish, ZuR. Tanooli and M. Shaheen, "Activity recognition framework in sports videos." *Multim. Tools Appl.*, vol. 83, pp. 15101-15123, 2021.
[31] M. Rafiq et al, "Scene classification for sports video summarization using transfer learning", *Sensors*, vol. 20, no. 6, p.1702, 2020.
[32] F. Wu, Q. Wang, J. Bian, N. Ding, F. Lu, J. Cheng, D. Dou and H. Xiong, "A Survey on Video Action Recognition in Sports: Datasets, Methods and Applications", *IEEE Trans. Multim.*, vol. 25, pp. 7943-7966, 2022.
[33] D. Xiao, F. Zhu, J. Jiang and X. Niu, "Leveraging Natural Cognitive Systems in Conjunction with ResNet50-BiGRU Model and Attention Mechanism for Enhanced Medical Image Analysis and Sports Injury Prediction", *Front. Neurosci.*, vol. 17, p. 1273931, 2023.