Sports Video Classification Using Convolutional Neural Network (CNN) with Normalization Flow

Nurul. A. Emran
Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka,
nurulakmar@utem.edu.my

*Nurul Izrin Md Saleh
Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka,
izrin@utem.edu.my

Muhamad Zaidi Mohd Ali
Politeknik Seberang Perai, Permatang Pauh, Pulau Pinang, Malaysia
mzaidiez@psp.edu.my

2024 5th International Conference on Artificial Intelligence and Data Sciences (AiDAS) | 979-8-3315-2855-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/AIDAS63860.2024.10730229

***Abstract*— Classifying sports videos is a crucial task in various applications, especially sports analytics, video retrieval, and content analysis. Unlike image classification, video classification is more challenging due to the dynamic nature of videos. Advancements in deep learning, particularly Convolutional Neural Network (CNN), have shown promising results in sports video classification. In this paper, we present the results of implementing CNN with normalization flow to classify several sports classes. In particular, the effect of classification accuracy in training datasets by increasing the number of classes (with similar characteristics and some noise) is analyzed. The effect of frame averaging, and the number of epochs were also observed. The results show that for the training dataset, CNN performance is slightly affected by the additional classes, but by increasing the number of epochs, the accuracy of training and validation datasets has improved. CNN can still maintain highly accurate classifications in test datasets (more than 80%) in some observations. CNN with frame averaging shows lower errors than single-frame CNN.**

Keywords*— *Sports Classification, sports video analysis, deep learning, Convolutional Neural Networks (CNN), image processing

I. INTRODUCTION

Video content has become popular digital content on the web which influences various aspects of life, including education, communication, and entertainment. This has led to a growing interest in automatic video processing, which involves identifying activities, generating text descriptions, summarizing content, and answering questions. Due to this, efficiently classifying video content is crucial for various applications including automated content tagging and performance analysis to enable real-time solutions for issues such as autonomous driving and understanding television streams. In particular, in broadcasting services, sports is a popular digital content that requires high volumes of sports videos being uploaded into the data servers [1]. Sports classification is a crucial task in supporting sports analytics, video retrieval, and content analysis. For example, sports analytics is used to design tactics or strategic decisions in soccer matches, which can affect revenue yielded in the soccer industry [2].

The capacity of traditional video classification algorithms to capture complex patterns and temporal dynamics contained in video data may be limited due to their reliance on hand-crafted features and shallow learning models. Traditional methods for sports video classification such as optical flow and Histogram of Oriented Gradients (HOG) rely heavily on manually engineered descriptors in identifying and distinguishing specific actions and dynamics within sports footage [3]–[5]. These handcrafted features, although historically important, have inherent limitations across various tasks and domains. In addition, the combination of rapid action dynamics and low-resolution video attributes are the issues that need to be dealt with [3]. For instance, the development of such features requires significant domain expertise and timely investment, which can be challenging.

Thus, in recent years, deep learning models that offer more promising feature recognition have been widely used for classifying video content based on scene context in sports videos. The results yielded from the models proved that issues on complex computer vision and signal processing tasks also have been addressed well [1]. Humans can recognize sports by considering a set of actions and surrounding elements or the scene. Thus, to build a system that can classify multiple sports classes based on sequential image data, a classification task needs to be inspired by human intuition in such a way that, it is capable of processing image information simultaneously in both the spatial and temporal domains [1], [4], [6].

Convolutional Neural Networks (CNNs) have revolutionized the field of computer vision, demonstrating exceptional performance in tasks such as image recognition, object detection, and segmentation. CNNs are particularly well-suited for video classification due to their ability to automatically learn hierarchical feature representations directly from raw pixel data. By leveraging spatial and temporal information in videos, CNNs can capture intricate details and patterns that are indicative of different sports categories. CNNs which is a type of deep learning algorithm, have emerged as a powerful tool for classifying sports videos, as CNNs have demonstrated exceptional performance in image classification tasks [3]. CNNs can be tailored to specific video datasets, enabling the identification of unique actions and dynamics across various sports.

The objective of this research paper is to develop and evaluate a novel approach for sports video classification that integrates CNNs with normalization flows. This approach aims to enhance the feature representation capabilities of CNNs by leveraging the powerful modeling abilities of normalization flows, thereby improving classification accuracy. In this paper, we report a study on the implementation of CNN for sports video classification. In particular, the study is set to analyze the effect of classification accuracy in training datasets by increasing the number of classes (with similar characteristics and some noise). The model was initially tested on three sports classes – Badminton, Football, and Swimming (2426 image frames) that are noise-free. Then, two more classes – Tennis and Weightlifting (1160 image frames) were added. These two sports classes have some noise (consisting of advertisements in the videos) that was purposely ignored. Training loss trends were observed to see the impact of the number of epochs on CNN.

This study also observes the implication of frame averaging or multi-frame video classification with normalizing flows as proposed in [7] against the classes that are not noise-free. The classification errors were measured against CNN with frame averaging and without frame averaging. The results of this study complement studies in sports video analysis that are significant for future applications in sports data analytics, content organization, and retrieval processes. This study seeks to advance the state-of-the-art in sports video classification, thereby significantly contributing to more sophisticated and reliable video processing tools in the sports domain.

The rest of the paper is organized as follows. Section II covers the related works on CNNs and their applications in the Sports domain. Section III consists of a description of the methods. Results and discussion are in Section IV and finally, Section V is the conclusion section.

II. Related Work

Videos are rich and complex data types due to their high dimensionality and size. Videos are composed of a sequence of images, known as frames. Video classification is similar to image classification but with a crucial difference: video classification involves processing consecutive frames, whereas image classification deals with static images. This dynamic nature of videos makes it challenging to develop effective models for classification tasks [8]. The simplest approach to handling video data involves extracting features from individual images and then classifying them. However, this method can lead to misclassification because a single video often contains multiple contents, such as advertisements and backgrounds. To address this, various methods can be employed, including the Long Short-Term Memory (LSTM) framework. LSTM algorithms are popular for their ability to identify long-term dependencies and patterns in data [9], [10]. This makes them a popular choice for tasks like price prediction. However, they may be overly complex for many trading applications that demand rapid and accurate predictions [11]. Thus, deep learning algorithms such as CNNs have been proposed to overcome the limitation of video classification in the earlier methods [8], [12]. One advancement in CNN is the use of normalization flow as proposed in [7], [13]. CNN with the normalization flow optimizes multi-frame video prediction data likelihood that aims to produce high-quality stochastic predictions through frame averaging [7].

Our study is closely connected to studies aimed at optimizing the use of CNNs for video classification, especially in the Sports domain (such as in [2]–[4]). Guangyu (2022) proposed sports video classification enhancement by proposing block brightness comparison coding (BICC) and block color histogram in CNN architecture [6]. In the study, the maximum mean difference (MMD) algorithm is employed to facilitate transfer learning and achieve the desired outcome.

A recent study by Gao et al. (2024) highlighted the challenge of handling sports video due to its dynamic nature and the traditional classification tasks such as optical flow and HOG are no longer adequate [3]. Thus, CNN has been optimized with a fuzzy noise reduction algorithm to enhance video quality. Li (2024) enhanced CNN by incorporating a deep learning gradient descent algorithm to classify and regress the image features of sports video output [5]. Pu et al. (2024) surveyed the use of AI in sports analytics where virtual sports pose unique requirements in sports technologies. For example, soccer match training and broadcasting require high-quality video-capturing systems, optical tracking systems, and wearable devices [2]. This indicates new challenges for classification algorithms to deal with real-time video datasets. The use of machine learning (and sports big data) in sports analysis/analytics (as reported in [14]–[16]) signals the growing interest in this domain.

III. Methods

The operational framework used in this study is based on the classification of videos using CNN with normalization flow architecture (as proposed by [13]). Based on the architecture, video classification is the extension phase for the image frame detection produced by the robust CNN with an additional variable of k predictions [13]. The overall process begins with the acquisition of sequence frame images of selected video where those images are then extracted with k list prediction. The last k predictions were taken and averaged before the label with the largest probability was selected. The label is chosen to classify the current frame. The assumption is that subsequent frames in a video will have similar semantic contents [17]. If the assumption holds, we can leverage this temporal relationship to improve video classification. By averaging the predictions across multiple frames, the system can smooth out fluctuations and enhance the overall accuracy of the video classifier. An experiment was conducted within the operational framework. As proposed by [13], the key normalization steps for CNN features in video classification are:

1. Frame Sampling and Calibration
2. CNN Feature Extraction
3. Spatial and Temporal Pooling
4. Feature Normalization
5. Classification

This pipeline was designed to fully exploit image-trained CNN architectures for video classification. The authors emphasize that the choice of CNN layer, normalization technique, and classifier type are the most critical factors affecting performance.

A. Experiment setup

An experimental approach similar to studies in [1] and [11] is used to test the hypothesis that: 1) CNN can maintain high accuracy with additional class (with similar characteristics and noise), and 2) using normalization flow of frames (frames averaging) can reduce classification error even though noise is present in the datasets. The experiment was configured on the CPU of Intel ® Core™ i7-6500M @ 2.50GHz and 8 GB memory on the workstation of a Linux-based platform with OpenCV and Keras Framework installed. Fig. 1 illustrates the flow of video classification using CNN with normalization flow which is used in this study.

At the first stage of the process, the video will be split into several frames to occupy the CNN framework. Then, for each frame of the video, once it passes through the CNN framework, the predicted class (label) will be written. The prediction result for each frame will be kept under a list called k-prediction. K-prediction stores the prediction value of each frame for the whole video. Then, the average of the prediction from each frame and the right label will be assigned based on the largest corresponding probability in the k-prediction. As a result, if the largest list is football, then the label with "football" will be written in the current frame video. If the largest list has changed, the label will be changed to match the

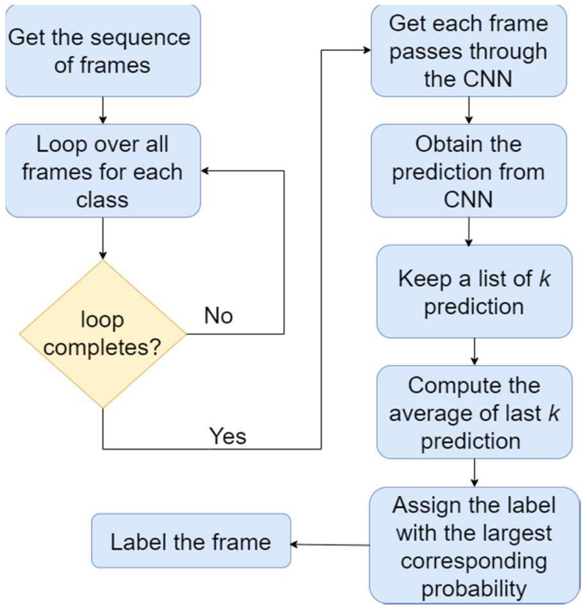


Fig. 1: Flow of video classification using CNN with normalization flow.

current average prediction. Subsequently, all frames will undergo the same processes.

The experiment was implemented in two phases, for two groups of datasets, as described next. CNN uses fine-tuning that updates the model architecture itself by eliminating the existing fully connected layer heads and creating new, randomly initialized ones. These new layers are then trained to predict the classes of the target dataset, while the weights of the lower convolutional layers are either frozen or fine-tuned with a lower learning rate.

B. *Datasets*

To classify sports, we need to focus on specific sports actions that combine spatial and motion features. When classifying individual images, spatial features alone are sufficient. However, for this task, we require a video dataset. Moreover, randomly selected sports videos are not suitable because they often contain non-sports content, such as interviews or audience reactions. Therefore, it is essential to use a dataset that consists of videos primarily featuring sports-action sequences that are free from 'noise'.

The dataset used for this study consists of sports videos obtained from publicly available sources. For data collection, samples of sports videos were obtained from Google and Bing search engines. Each video of the classes under study was annotated manually into several series of image frames corresponding to the sport types. In the first phase of the experiment, three classes of sports namely Badminton, Football, and Swimming were prepared. The dataset consists of 2426 image frames of three sports classes. Data preprocessing was conducted for better representation and quality of the data by extracting individual frames from the sports videos to create a uniform dataset. Videos were loaded using OpenCV library and decomposed into individual frames at 50 to 120 frames from each video clip to ensure uniform temporal solution. Later, the individual frames of images were converted into greyscale images, and noises which are the non-sports content (e.g. interviews, advertisements, and audience reactions were also removed).

In the second phase, two more classes namely Tennis and Weightlifting were added. The dataset consists of 1162 image frames of sports classes. The dataset consists of some noise as it consists of sports video frames with some advertisements. Tennis was chosen as it has a high similarity to Badminton. Weightlifting is a class that is less similar to other classes. This is purposely set to test the first hypothesis. The details of the datasets are shown in Table I. For normalization, this study implements feature normalization where L2 normalization is applied for hidden layer features and root normalization for output layer features. The L2 norm, also known as the Euclidean norm, calculates the square root of the sum of the squared values of the vector elements. Mathematically, the L2 norm of a vector x with n elements can be defined as: $||x||_2 = \sqrt{(x_1^2 + x_2^2 + \ldots + x_n^2)}$

For model training, the number of epochs was initially set to 50. An epoch is the number of iterations a training dataset takes to complete an algorithm. The number of epochs was increased to 75 in the second phase of the experiment when the new classes were added. Training loss trends were observed to see the impact of the number of epochs on the training datasets. The accuracy of classification is measured against the test datasets.

TABLE I. SPORTS DATASETS

Class	Training (75%)	Validation (25%)	Total
Badminton	703	235	938
Football	517	172	689
Swimming	599	200	799
Tennis	600	200	800
Weightlifting	270	92	362
Total	**2689**	**899**	**3588**

To test the second hypothesis, we measured the impact of frame averaging by implementing CNN with normalization flow and without the normalization flow. We use the default size frame for averaging as set in the OpenCV library which is 128. For this purpose, two datasets were used namely Tennis and Weightlifting, which both consist of some noise.

C. *Evaluation Criteria*

In the context of sports video classification, it is crucial to evaluate the performance of the model to guarantee its effectiveness and reliability. In this study, common metrics are used to assess the performance of CNN which are F1-score, precision, recall, and accuracy.

F-measure or F1-score is the measure of a model's performance that takes consideration of weighted harmonic means of the precision and recall of the test. F1-score is usually used for imbalanced data such as video datasets. In this study, the F1-score was used to measure the classification accuracy of the training set. Eq. (1) is the formula for F1-score, followed by Eq. (2) and Eq. (3) for precision and recall respectively:

$$F1\ score = \frac{precision.recall}{precision+recall} \tag{1}$$

$$precision = \frac{True\ Positive}{True\ Positive+False\ Positive} \tag{2}$$

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

Loss measure is used to measure the average loss across each image batch within an epoch during training and validation. Based on the loss measure, the accuracy measure within an epoch during training and validation can be calculated. The loss measure is critical in determining the detriment of the predictor model when the input data points are classified in a dataset. The less the failure, the better the classifier models the relationship between the data input and the results output.

Accuracy is a measure of how often a classifier correctly predicts. Accuracy is defined as the ratio of the number of correct predictions and the total number of predictions, which can be calculated from the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as in Eq. (4):

$$accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4)$$

IV. Results and Discussion

Table II and Table III consist of the results of training datasets evaluation for phase 1 and 2 respectively. Based on the tables we can see that most readings for F1-score, Precision, and Recall of CNN have dropped in the second phase of the experiment (Table II) for Badminton, Football, and Swimming. This is expected as the additional classes which are Tennis and Weightlifting consist of some noise. Nevertheless, there is no abrupt drop as the amount of decrease is at most 0.12 (see the Badminton class). F1-score for Tennis and Weightlifting is no less than 0.50 which is still within the acceptable range. Therefore, this result indicates that the CNN model achieves a good balance between precision and recall, signifying effective performance in the sports video classification.

TABLE II: Results for the first phase of the training dataset

Class Accuracy: 0.89	F1-score	Precision	Recall
Badminton	0.87	0.84	0.90
Football	0.95	0.82	0.86
Swimming	0.91	0.92	0.93

TABLE III: Results for the second phase of the training dataset

Class Accuracy: 0.78	F1-score	Precision	Recall
Badminton	0.75	0.70	0.81
Football	0.85	0.86	0.84
Swimming	0.86	0.78	0.95
Tennis	0.65	0.73	0.58
Weightlifting	0.76	0.85	0.68

Next, Fig. 2 shows the graph of training loss and validation loss of CNN for the first phase of the experiment that involves three classes namely Badminton, Football, and Swimming. Training accuracy and validation accuracy are also plotted in the graph. The graph depicts that, as the number of epochs increases, the training and validation loss are both decreasing. Training loss was higher than validation loss at the beginning, but the loss values started to overlap from the 25th epoch. As for the training accuracy, we can see there is a sharp increase at the beginning, until 10 epochs. After 10 epochs, training accuracy and validation accuracy tend to overlap and show a slight increase until epoch 50. The result shows that the model is well-fitted after the 10^{th} epoch as there is no big gap between the training loss and the validation loss.

Fig. 3 shows the graph of training loss and validation loss of CNN for the second phase of the experiment that involves five classes, where Tennis and Weightlifting are the newly added classes. A decreasing trend for the training loss (red line) can be observed until the 75^{th} epoch. However, unlike in the first phase, validation loss was higher than training loss at the beginning. There is some inconsistency in the loss trend for the validation dataset towards the 75^{th} epoch, with some spikes, especially at the beginning and after the 60^{th} epoch. This indicates that the image prepared in the validation datasets consists of some noise or high similarity with other sports classes. This is expected because we know that the two additional datasets consist of some noise. In addition, tennis and badminton are quite similar. A few similar images might be detected during the validation.

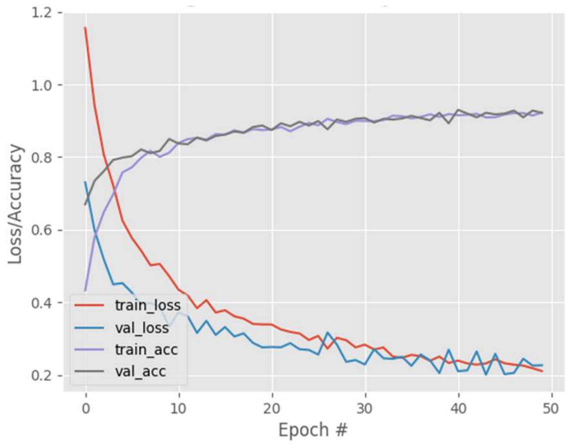


Fig. 2: Training and validation loss (and accuracy) trend for phase 1.

Next, Fig. 4 illustrates the average scores of CNN classification accuracy across validation data sets of five sports classes. Based on the figure, Badminton, Swimming, and Weightlifting exhibit outstanding average accuracy scores, which are between 83% to 98.9%. These classes also have high scores for training data set accuracy (75% and above). Nevertheless, football shows low accuracy (50%) which can be explained by false classification as shown in Table V. For example, based on the sample images, the third football image has been falsely classified as Badminton (with -41.99% error).

For simplicity, we present three sample images (with scores) for each sports class that was tested in this study as shown in Table V. Tennis has the lowest accuracy (27.5%) which could be caused by its high similarity with Badminton. Thus, Tennis tends to be mistakenly classified as Badminton. As shown in Table V, the third image of Tennis was falsely classified as Badminton (with -99.14% error). With this result, the first hypothesis is only partially supported.

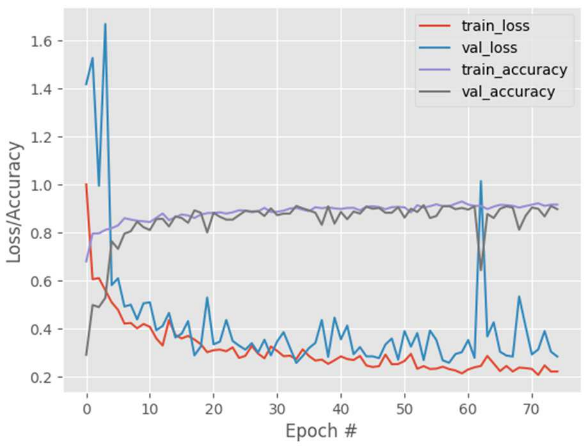


Fig. 3: Training and validation loss (and accuracy) trend for phase 2.

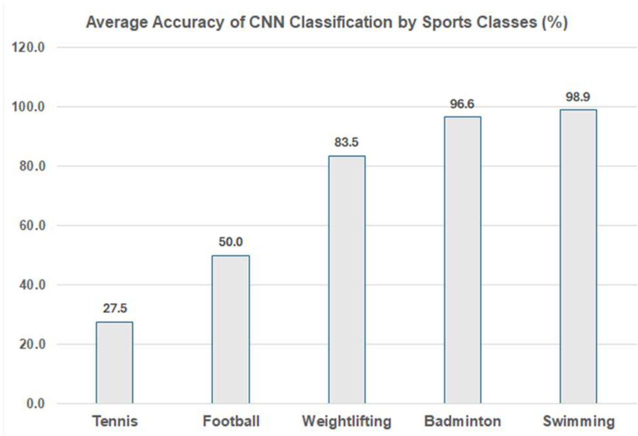


Fig. 4: The results of the average accuracy of CNN for five sports classes.

Finally, Table IV shows the results of classification for CNN that is configured with normalization flow/frame averaging (multiple frames) and CNN without it (single frame). Observations were made against validation datasets which are Tennis and Weightlifting. For both datasets, we can see that CNN with frame averaging shows fewer errors as compared to CNN which does not implement normalization flow. Thus, the results support the second hypothesis of this study.

TABLE IV

Datasets	Classification Results	Single Frame	Averaged Frame
Tennis (200 frames)	False	31 (15.5%)	22 (11.0%)
	True	169 (84.5%)	178(89.0%)
Weightlifting (92 frames)	False	54 (58.69%)	21 (22.82%)
	True	38 (41.30%)	71 (77.17%)

V. CONCLUSIONS

This study presents the results of implementing a Convolutional Neural Network (CNN) to classify several sports classes with normalization flow. In conclusion, the results show that classification accuracy in the training dataset is slightly affected by the additional classes, but the accuracy of the training and validation datasets improves with an increase in the number of epochs. The use of frame averaging shows lower errors than the single-frame CNN, where observation has been made against data sets with noise where errors are expected. The CNN can maintain highly accurate classifications in test datasets (with more than 80% on average) in some observations. These results demonstrate the effectiveness of the proposed CNN model in classifying sports videos despite the presence of non-noise-free classes. Nevertheless, there is still room for improvement, especially in improving the sensitivity of classification against classes with similar characteristics (e.g., Badminton and Tennis). Future improvement on CNN might be in terms of combining multiple and diverse architectures that can help improve the robustness and generalization of the model to various image categories through Ensemble learning.

ACKNOWLEDGMENT

This study is funded by Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM). The authors would like to thank the Centre for Research and Innovation Management (CRIM), UTeM, and members of the IDEAL research group for all their support. The authors would like to thank all individuals and the research team for their support.

REFERENCES

[1] M. A. Russo, L. Kurnianggoro, and K. H. Jo, "Classification of sports videos with combination of deep learning models and transfer learning," *2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019*, pp. 1–5, 2019.

[2] Z. Pu *et al.*, "Orientation and Decision-Making for Soccer Based on Sports Analytics and AI: A Systematic Review," *IEEE/CAA J. Autom. Sin.*, vol. 11, no. 1, pp. 37–57, 2024.

[3] T. Gao *et al.*, "Sports Video Classification Method Based on Improved Deep Learning," *Appl. Sci.*, vol. 14, no. 2, 2024.

[4] 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, pp. 1–15, 2022.

[5] W. Li, "Applied Mathematics and Nonlinear Sciences," *Appl. Math. Nonlinear Sci.*, vol. 9, no. 1, pp. 1–15, 2024.

[6] H. Guangyu, "Analysis of Sports Video Intelligent Classification Technology Based on Neural Network Algorithm and Transfer Learning," *Comput. Intell. Neurosci.*, vol. 2022, 2022.

[7] M. Kumar *et al.*, "VideoFlow : A Flow-Based Generative Model for Video VideoFlow : A Flow-Based Generative Model for Video," in *8th International Conference on Learning Representations, ICLR 2020 (2020)*, 2020.

[8] Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang, "Deep learning for video classification and captioning," in *Frontiers of Multimedia Research*, Association for Computing Machinery and Morgan & Claypool, 2017, pp. 3–29.

[9] M. Ghislieri, G. L. Cerone, M. Knaflitz, and V. Agostini, "Long short-term memory (LSTM) recurrent neural network for muscle activity detection," *J. Neuroeng. Rehabil.*, vol. 18, no. 1, pp. 1–15, 2021.

[10] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, "A survey on long short-term memory networks for time series prediction," *Procedia CIRP*, vol. 99, no. July 2020, pp. 650–655, 2021.

[11] M. A. Russo, A. Filonenko, and K. H. Jo, "Sports Classification in Sequential Frames Using CNN and RNN," *2018 Int. Conf. Inf. Commun. Technol. Robot. ICT-ROBOT 2018*, pp. 1–3, 2018.

[12] C. Heipke and F. Rottensteiner, "Deep learning for geometric and semantic tasks in photogrammetry and remote sensing," *Geo-Spatial*

Inf. Sci., vol. 23, no. 1, pp. 10–19, 2020.

[13] S. Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov, "Exploiting Image-trained CNN Architectures for Unconstrained Video Classification," in *British Machine Vision Conference*, 2015.

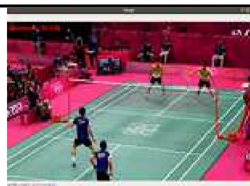
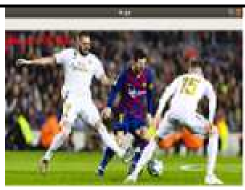
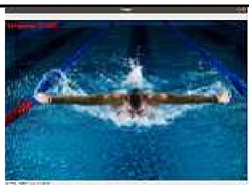
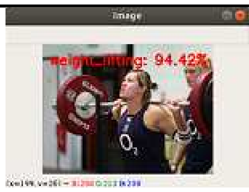
[14] M. Gifford and T. Bayrak, "A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression," *Decis. Anal. J.*, vol. 8, no. July, p. 100296, 2023.

[15] Z. Bai and X. Bai, "Sports Big Data: Management, Analysis, Applications, and Challenges," *Complexity*, vol. 2021, no. 1, 2021.

[16] Y. F. Roumani, "Sports analytics in the NFL: classifying the winner of the superbowl," *Ann. Oper. Res.*, vol. 325, no. 1, pp. 715–730, 2023.

[17] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 3, no. 1, pp. 1–37, 2007.

TABLE V: Sample results of CNN classification accuracy for five sports classes

Class	**Sample images**		
Badminton	 True: 99.48	 True: 90.48	 True: 99.83
Football	 True: 99.86	 True: 96.12	 False: (-41.99)
Swimming	 True: 97.36	 True: 99.47	 True: 99.99
Tennis	 True: 85.03	 True: 96.52	 False: (-99.14)
Weightlifting	 True: 98.55	 True: 94.42	 True: 57.56