

# Cricket Video Summarization Using Temporal CNNs



By

**Fahd Raza**

**Fall-2020-MS-EE-SP 00000328931 SEECS**

Supervisor

**Dr Mohsin Kamal**

**Department of Electrical Engineering**

A thesis submitted in partial fulfillment of the requirements for the degree of

In

School of Electrical Engineering & Computer Science (SEECS) ,

National University of Sciences and Technology (NUST),

Islamabad, Pakistan.

(August 2024)


## THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis entitled "Cricket Video Summarization Using Temporal CNNs" written by Fahd Raza, (Registration No 00000328931), of SEECs has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

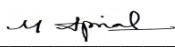
Signature:  \_\_\_\_\_

Name of Advisor: Dr. Mohsin Kamal

Date: 08-Aug-2024

HoD/Associate Dean:  \_\_\_\_\_

Date: 08-Aug-2024

Signature (Dean/Principal):  \_\_\_\_\_

Date: 08-Aug-2024

FORM TH-4


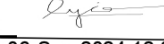
**National University of Sciences & Technology**  
**MASTER THESIS WORK**

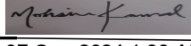
We hereby recommend that the dissertation prepared under our supervision by: (Student Name & Reg. #) Fahd Raza [00000328931]

Titled: Cricket Video Summarization Using Temporal CNNs

be accepted in partial fulfillment of the requirements for the award of Master of Science (Electrical Engineering) degree.

**Examination Committee Members**

1. Name: Salman Abdul Ghafoor Signature:   
06-Sep-2024 12:00 PM
2. Name: Wajid Mumtaz Signature:   
06-Sep-2024 12:00 PM

Supervisor's name: Mohsin Kamal Signature:   
07-Sep-2024 1:33 AM



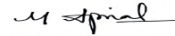
Salman Abdul Ghafoor  
HoD / Associate Dean

06-September-2024

Date

**COUNTERSIGNED**

09-September-2024  
Date



Muhammad Ajmal Khan  
Principal

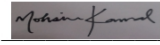
THIS FORM IS DIGITALLY SIGNED

Publish Date & Time:

## Approval

It is certified that the contents and form of the thesis entitled "Cricket Video Summarization Using Temporal CNNs" submitted by Fahd Raza have been found satisfactory for the requirement of the degree

Advisor : Dr. Mohsin Kamal

Signature:  \_\_\_\_\_

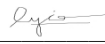
Date: 08-Aug-2024

Committee Member 1: Dr. Salman Abdul Ghafoor

Signature:  \_\_\_\_\_

09-Aug-2024

Committee Member 2: Dr. Wajid Mumtaz

Signature:  \_\_\_\_\_

Date: 09-Aug-2024

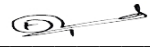
Signature: \_\_\_\_\_

Date: \_\_\_\_\_

## Certificate of Originality

I hereby declare that this submission titled "Cricket Video Summarization Using Temporal CNNs" is my own work. To the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECs or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECs or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics, which has been acknowledged. I also verified the originality of contents through plagiarism software.

Student Name:Fahd Raza

Student Signature: 

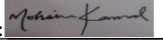
**Certificate for Plagiarism**

It is certified that PhD/M.Phil/MS Thesis Titled "Cricket Video Summarization Using Temporal CNNs" by Fahd Raza has been examined by us. We undertake the follows:

- a. Thesis has significant new work/knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.
- b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.
- c. There is no fabrication of data or results which have been compiled/analyzed.
- d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.
- e. The thesis has been checked using TURNITIN (copy of originality report attached) and found within limits as per HEC plagiarism Policy and instructions issued from time to time.

**Name & Signature of Supervisor**

Dr. Mohsin Kamal

Signature: 

# Dedication

This thesis is dedicated to all the deserving children who do not have access to quality education especially young girls.

# Acknowledgments

Glory be to Allah (S.W.A), the Creator, the Sustainer of the Universe. Who only has the power to honour whom He please, and to abase whom He please. Verily no one can do anything without His will. From the day, I came to NUST till the day of my departure, He was the only one Who blessed me and opened ways for me, and showed me the path of success. Their is nothing which can payback for His bounties throughout my research period to complete it successfully.

**Fahd Raza**



# Abstract

In this study, a new method for the summarization of very long cricket videos through the employment of an enriched deep learning approach is proposed and utilises the inherent feature of a Three-Dimensional Convolutional Neural Network (3D-CNN). Our methodology comprises several key stages as the first one is development and testing of Residual Network (ResNet) structure 3D-CNN for the identification and classification of significant events in cricket matches the second one is the annotation of the video clips divided into five classes of actions: fours, sixes, wickets, milestones and others and the third one is fine-tuning of the ResNet-based 3D-CNN with the use of the annotated

The model is expected to correctly detect important cricket events, and thus to assist in creating a highlight summary by keeps clips with fours, sixes, wickets, and milestones while discarding all unnecessary parts, to ensure we test the performance of our Summarization system, we conducted experiments by assessing the accuracy of the system after training the model on unseen cricket match videos got an average accuracy of 97%. Based on these results, it proves that our approach provides an efficient and accurate way to autonomously produce short and context-specific summaries for cricket matches using the 3D-CNN.

# Contents

<b>Acknowledgement</b>	<b>vii</b>
<b>Abstract</b>	<b>viii</b>
<b>List of Tables</b>	<b>xii</b>
<b>List of Figures</b>	<b>1</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem Statement . . . . .	2
1.1.1 Manual Editing . . . . .	2
1.1.2 Rule-Based Systems . . . . .	3
1.2 Study's Objective . . . . .	3
1.3 Advantages and Potential Applications . . . . .	4
1.4 Structure of thesis . . . . .	5
<b>2 Literature Review</b>	<b>6</b>
2.1 Machine Learning . . . . .	6
2.2 Classification . . . . .	7
2.3 Deep learning . . . . .	8
2.3.1 Deep Learning in Medical Field . . . . .	9
2.3.2 Deep Learning in Computer Vision . . . . .	9
2.4 Adversarial Attacks . . . . .	10

## CONTENTS

2.4.1	White Box Attack and White Box Attack	10
2.4.2	Targeted and Non-targeted Attacks	11
2.4.3	Adversarial Attacks Common Type	11
2.5	Related Work	11
<b>3</b>	<b>Deep Learning Models and Transfer Learning</b>	<b>15</b>
3.1	Deep Learning Models	15
3.2	Convolutional Neural Network	16
3.2.1	Alexnet	17
3.2.2	VGGNET	18
3.2.3	DenseNet	19
3.2.4	ResNet	19
3.3	Transfer Learning	20
3.4	Overview of 3D CNN	21
3.4.1	Applications	22
3.4.2	Challenges and Considerations	23
3.4.3	Conclusion	23
<b>4</b>	<b>Proposed Methodology</b>	<b>24</b>
4.1	Dataset	24
4.1.1	Dataset Collection	24
4.1.2	Dataset Preprocessing	25
4.2	Model	26
4.2.1	ResNet-34 + LSTM for Event Classification	26
4.2.2	ResNet-50 + LSTM for Sports Action Classification and Video Summarization	28
4.2.3	3DCNN Model for Sports Action Classification	30
4.2.4	ResNet-18 based 3D CNN	32

## CONTENTS

<b>5</b>	<b>Discussion and Results</b>	<b>36</b>
5.1	Discussion of Results . . . . .	36
5.2	Training Procedure and Hyperparameters . . . . .	36
5.2.1	ResNet-34 + LSTM on Our Dataset . . . . .	37
5.2.2	X3D Model . . . . .	38
5.2.3	3D-CNN Model . . . . .	39
5.2.4	ResNet-18 with 3D-CNN Layers . . . . .	40
5.2.5	ResNet 50 + LSTM on our Dataset . . . . .	40
5.2.6	ResNet18 based 3D-CNN model on an expanded Dataset . . . . .	41
5.3	Comparison with the state of art models . . . . .	42
5.4	Conclusion . . . . .	44
<b>6</b>	<b>Conclusion and Future Work</b>	<b>45</b>
6.1	Future Work . . . . .	46
	<b>Bibliography</b>	<b>47</b>

# List of Tables

4.1	3D-ResNet34 Layer Parameters . . . . .	28
4.2	Layer Parameters for ResNet-50 . . . . .	30
4.3	Summary of 3D CNN Model Layers and Parameters . . . . .	32
4.4	Model Summary . . . . .	35
5.1	Performance of different models on our dataset . . . . .	43
5.2	Comparison of different models for video summarization . . . . .	43

# List of Figures

2.1	Adversarial Example [43]	11
3.1	General Structure of CNN	18
3.2	Architecture of Alexnet [67]	18
3.3	Architecture of VGG model [66]	19
3.4	Architecture of DenseNet [67]	19
3.5	Architecture of Residual block [68]	20
3.6	Block diagram of Transfer learning [69]	21
4.1	Frames from our own dataset	25
4.2	Block diagram of ResNet34 + LSTM	28
4.3	Block diagram ResNet-50 + LSTM	30
4.4	Block diagram of 3D-CNN	31
4.5	Block diagram of ResNet-18 based 3D CNN	33
4.6	Block diagram of Convolutional and Identity Block	34
5.1	The soccer model accuracy on our dataset of cricket.	37
5.2	X3D model accuracy on our dataset of cricket	38
5.3	3D-CNN accuracy on our dataset of cricket.	39
5.4	The soccer model accuracy on our dataset of cricket.	40
5.5	ResNet50 + LSTM performance our dataset of cricket.	41
5.6	ResNet-18 based 3DD evaluation on our dataset	42

## CHAPTER 1

# Introduction

The high modernity and technological developments experienced in the recent past have led to the generation of large amount of data, especially in forms of video and image data [1]. Much of this increase is seen across social media networks such as YouTube as well as sports streaming services where content with long duration, especially in cricket etc., is high. Sports videos are perhaps the most popular and incredibly useful videos out there, with immense economic connotations [2]. However, because of the normally long periods for most of the sporting events, it would be impossible for the audiences to view complete recordings of the events especially in the offline circumstances [3].

Cricket, one of the most popular sports in the world is based on the fact that the majority of those games last not hours, but several days. It becomes very tiresome for the fans to keep up with a very long list of events that are played all through the year [4]. Consequently, goalscorers operational videos have emerged as a crucial part of information that fans use to comprehend important and renewed aspects in the game.

One reason is that people are increasingly strapped for time because of how the flow of life has become in the modern, short form content is now more popular than ever. These changes have been to the effect of reducing the audience interest in the longer formats of the videos. Instagram Reels or TikTok are the examples of it as all these platform rely heavily on short video like reel [5]. Third, current research shows that videos of Test match have lost the attention of the audience and people like T20 more. However, the creation of video highlight brings along with it the problem of timing of summarization since the process has to be done manually and

needs a lot of professional editing.

Modern approaches to the problem of summarizing video content are carried out using very interesting methods based, for example, on the scene categorization, the OCR, the frequency analysis, sharpness and boundary scoring, the ensemble learning and the LSTM networks. There is also research under way involving reinforcement learning with intricate scoring functions, though such approaches generally entail considerable computational overhead [5].

In this work, an attempt will be made to create an automatic system capable of creating high-quality summaries from the recorded cricket match videos without prior facilitation of the same. The technique includes analyzing the input video for high energy clips which can be compiled to make a highlight video. Deep learning algorithms and natural language processing (NLP) techniques, alongside various Python libraries, are employed to develop a model capable of generating three types of summaries, Video summaries, textual summaries as well as audio summaries.

## **1.1 Problem Statement**

The case of summarizing cricket videos is felt in the area of identifying major highlights from long and often intricate games especially drawn from Test cricket that can go for days. This involves a lot of work due to the high number of videos produced as well as the very many possible significant events which include wickets, boundaries and milestones to mention but a few, which involve not only selection but analysis of the surrounding context as well. Current methods of summarization face several limitations:

### **1.1.1 Manual Editing**

The process of descending to the selection of highlights during the summary of a video is primarily done in a more or less, a manual manner and as such, is not devoid of bias as the selection is based on the observer's perception of an event as being significant or otherwise. The above subjectivity brings with it the issue of variability in the selection criteria; where important mo-



ments may be left out and less important ones picked. This is a highly manual task, which in turn is highly time-consuming especially when one has to go through large amounts of video footage in a bid to obtain a good number of segments for analysis. The ‘significance’, defined very narrowly on the level of a parish, and the tendency to filter it based on the biases and preferences of the reviewer, only add more layers of complexity to the process and might leave the final summary far from being an accurate representation of what contents matter.

### **1.1.2 Rule-Based Systems**

A rule-based system is quite rigid and unable to stand the flexibility needed to consider the aspects of different matches. In addition, it is non-adaptive in the sense that they use fixed sets of rules and apply them with no variations to the different aspects of a game. Moreover, the time stamping accuracy in such systems is usually an issue and can distort the identification of critical frames, thereby reducing the overall efficiency of generating high-spot points.

This makes it important to find more accurate systems that provide proper substitutes to more time-consuming manual methods to provide adequate and bias-free summaries of the cricket matches. These systems should be able to recognize variations in the context and therefore identify events that are worthy in relation to match context. Further, the system should be able to create summaries that are brief and accurate and written in the language understood by most people. Such systems, based on complicated algorithms and incorporating the principles of machine learning, may thereby provide enhanced accuracy and lesser human bias and inaccuracies. Lastly, these techniques would help the viewer experience a better sum-up of the cricket match information, which was easier, more inclusive and intuitive due to the specifics of the match.

## **1.2 Study’s Objective**

The key objective of this research are:

- With an aim of achieving an effective video summarization for cricket videos an efficient 3D Convolutional Neural Network based on ResNet for the automated extraction of key events from the videos is conceived

- To gather and clean a richly annotated dataset of crickets that includes various types of events to give a rich picture of what the model is trained on.
- To optimize the model to achieve high accuracy in the classification of specific cricket events, such as boundaries, wickets, and milestones, through rigorous training and validation processes.
- To generate concise highlight videos by effectively identifying and retaining significant cricket events, thereby providing an efficient method for summarizing lengthy video content.

### **1.3 Advantages and Potential Applications**

The research objectives outlined in this study can have a proper implication to the sports video analysis, especially to the cricket field. In one of the use cases the developed 3D Convolutional Neural Network (3DCNN) model is applied for generating cricket highlight videos. It can help save a great deal of time and reduce the load of manual video processing for broadcasters, sports analysts, and fans, offering the view on the most significant moments of a match as soon as possible. Overall by providing brief report like summary, the model is useful for fans who may not have the time to watch and follow the entire video. It enables fans to get briefed on major events in a match without being required to sit through most of the match thus making cricket more engaging to the 'layman'. Also, the rich annotated dataset obtained in the process of the research can be seen as the useful application of this work in the sphere of sports and data analytics to improve the performance of athletes, teams, and coaches, as well as to assess the match outcome and potential strategies of the teams in case of further matches.

In addition to these, the incorporation of the proposed model into broadcasting processes brings new avenue for delivery of customized content. Domains can apply this technology in the middle of a match or even after the match is over during the analysis of the event to add value to the viewer's experience. However, what is perhaps more important is the immediate capacity to classify particular cricket occurrences and to use this specificity is designing complex archival and retrieval systems within sports video databases. Such systems would allow users to easily search and identify required portions of a large set of cricket matches and thus would be useful tools for researches, historians and cricket fans. It gives the hope of extending the coverage of

cricket-related media through social media tools, where short and to-the-point summaries are highly effective.

As for the future, there are several opportunities that can be associated with the technology studied in the framework of this work. An example of the benefit of the model is if it is used for real-time event detection and summarization during the live cricket matches which is possible in the future. This advancement would help in replay of critical incidents on the spot thus making viewing during the live event much more interesting. Further, the system could also develop into broadcasting customized highlight videos for special audience interests, e.g. special player selection or type of event. The model could also be combined with the wearable technology employ by the players to include other data that would enhance the event categorization to be more precise and detail. Originally, this model was simply designed for cricket, yet it is quite possible to extend its utilization to a number of other sports, such as football, basketball, or tennis, as long as the model is retrained on the datasets of the specific sport's activities. However, the generation of highlight videos has possibilities for their monetization, for instance, the use of special paid access to the organised by sports organizations and broadcasters spectacular exclusive AI-produced highlights or detailed matches' analysis

### **1.4 Structure of thesis**

The remaining sections of this dissertation are as follows: The literature overview is in Chapter 2, followed by deep learning models and transfer learning in Chapter 3, and the methodology of the proposed work is discussed in Chapter 4. In chapters 5 and 6, the results as well as future work are discussed.

# Literature Review

## 2.1 Machine Learning

With the increasing amount of data, a tool to analyse it and generate information from it is needed, and Machine Learning can help [6]. Machine learning is a branch of artificial intelligence that assists in the development of autonomous systems in which computers learn about the task at hand without being explicitly coded [7] [8]. Machine Learning may be used in a variety of vocations [9] [10] .

For machine learning algorithms to work, handcrafted features are necessary [11]. Feature selection is a critical stage in Machine Learning, when important characteristics are chosen to allow the model to train and converge smoothly [12] . In medical diagnostics, a number of machine learning applications have been used [13]. In the pharmaceutical industry, many Machine learning applications have recently been developed, with the system being able to identify patients who are more likely to benefit from the therapies [14] [15].

Machine learning applications can be broadly categorized into three fundamental areas: It follows that some subcategories of Machine Learning include Supervised Learning, Unsupervised Learning, and Reinforcement Learning [16]. In Supervised Learning the data is already tagged or labeled and the model is trained to find out the dependency of the output variable with the input variable. This makes the model a capable one in a sense that can predict on new unseen data. Supervised learning can be further divided into two key tasks: discrete value such as logistic regression forms a classifier which predicts discrete categories while a regressive model which predicts a continuous scalar value. For instance, in a classification context the model assigns samples in two categories such as cat images and dog images in contrast to regression where the

model estimates the house price given some input attributes. Most of the supervised learning models rely heavily on labeled data, therefore data labeling which is a part of data preparation step is an important task.

There being no labeled data, the model is trained through Unsupervised Learning. However, the aim of the model is to find some invariance of the data, identify some structure or some property of the data, e.g., grouping of similar data points [17]. This kind of learning is particularly useful in exploratory data analysis where it is an objective to find previously unknown relations or clusters. Of all the categories of machine learning methods, unsupervised learning is most useful in areas where it is difficult or costly to get labeled datasets, for instance in language processing or genomics. While the motivation the latter is clear, a problem, which arises consequently, is the fact that evaluation of unsupervised models is a very difficult process, especially when no labeled data is used, which means that more subtle metrics and validation procedures have to be employed.

Last is Reinforcement Learning which is a subset of machine learning where the model works with and learning from its environment using feedback in the form of rewards or penalties regarding its actions [18]. Such approach is most useful when the model is required to take sequence of decisions at different times, for example in game playing or robotics. Through experience the model learns how to select its actions in such a way that its total rewards increases from one time to another thereby enhancing its decision making. Reinforcement learning is effectively applied in large systems, where the action's outcome is hardly determined immediately – as in the case of self-driving cars or stock exchange operations. But it was computationally intensive and the model was highly sensitive to a number of hyperparameters to achieve good performance.

## **2.2 Classification**

When data is divided into more than two categories or classes the goal of the machine learning model is to understand the mapping between input features and the respective output in order to be able to categorize or classify new unseen data into the right categories or classes [19]. This is done by the model whereby it looks at the training data and sees if there are any patterns it can use to set boundaries between the different classes. These decision boundaries are mathematical functions which define the region of space within which objects of the different classes exist and then provides the model with ability to separate between them [20].

While learning, the model changes its weights to reduce the error for each input example – the input’s assigned class – a process with the help of an optimization procedure like a gradient descent [21]. In other words, given a new input data it is classified to the category that corresponds to a certain region in the higher dimensional space having the decision boundaries specified by the coefficients received during training. On the other hand the performance of the model in providing unmistakable discrimination of the various classes depends on the intricacy of the data set, suitability of features adopted and how the model generalizes the learned data to unseen samples. Feature selection and effective training are thus vital so that the model does not either over-complicate or under complicate the classes and their distribution [22].

### **2.3 Deep learning**

There are several types of Machine Learning, some of which are very specific and others more general; deep learning is part of the later and has steadily gained a lot of attention and adoption over the last couple of years for the fact that it can learn the representations of the data or the features on its own without the interference of the user [23]. As opposed to more conventional Machine Learning techniques where the ‘good’ features are required in order to obtain high performance, the Deep Learning methodologies, through a process of hierarchical representation learning, are capable of learning these features from scratch from raw data. This ability to learn at multilayer abstractions is what makes deep learning models feasible for highly complex task such as image and speech recognition; natural language processing, and even autonomous driving [24].

The attraction to Deep Learning is, therefore, for its capability and versatility; it is also easy to use as has been made easy by additional levels of frameworks such as TensorFlow and PyTorch [25]. It enables practitioners to design and train complex neural networks rather conveniently as compared to other traditional approaches, which demands great amount of domain knowledge in the process of feature extraction. Also, deep learning models are more accurate than conventional Machine Learning methods particularly in problems with big data and unstructured data [26]. This is because deep neural network has the capability of discovering features inherent in data in a more complex manner than the conventional shallow neural networks thus enhances the prediction and classification. Subsequently, Deep Learning is now the favored process in many advanced imposing applications, which has enhanced modern developments in computer vision and natural language processing, and bioinformatics.

Neural networks are used in deep learning. It employs an artificial neural network that attempts to replicate the human brain. Different layers are combined to form a neural network [27]. Neural network is made up of three layers: an input layer, a hidden layer, and an output layer. These levels are made up of 'nodes.' The data is sent into the input layer, the hidden layer does some calculations, and the output layer outputs the required results[28]. The weights and biases of these hidden layers are changed in order to minimize the loss function and allow the model to converge. The depth of a neural network is increased by adding numerous hidden layers, which is why the term "deep" is employed [29].

Deep learning models work well when there is a lot of data. Depending on the activities they're utilized for. Neural networks come in a number of shapes and sizes. When working with photographs, Convolutional Neural Networks (CNN) are used. You'll want to use a Recurrent Neural Network if you're working on things like Natural Language Processing (RNN)[30]. Robotics, video synthesis, facial recognition, and diagnosis of disease are all examples of deep learning applications[31].

### **2.3.1 Deep Learning in Medical Field**

Because of deep learning, many applications for medical diagnostics have been developed. Deep learning is assisting health professionals in identifying more precise and efficient ways to treat patients. These models aid in medication discovery by examining the patient's medical history, allowing for improved therapy. It can also be used to forecast whether or not a patient's medical insurance claim will be fraudulent. The diseased photos were digitized with full slide scanners and then employed in the deep learning model. This has aided pathologists in analyzing complex tasks such as cancer detection. These models also make heavy use of radiological scans such as CT and MRI images, which assists radiologists[32] [33] [34][35] [36].

### **2.3.2 Deep Learning in Computer Vision**

Computer vision is one of the areas that has been boosted through deep learning, that has resulted to achievement of various vision related tasks. Perhaps the most popular of them is the image classification where Convolutional Neural Networks (CNN) outperforms other similar methods in categorizing images into prescribed classes [37]. In addition to image classification, deep learning finds application in object detection, for example; YOLO and Faster R-CNN do

not only recognize objects in images but locate them while being useful in applications such as self-driving cars and security cameras. One is image segmentation in which models like the U-net splits an image into segments so that, for instance, the details of certain structures of the body can be distinguished during diagnosis and planning of treatment depending on the situation in medical imaging [38].

Moreover, it has been applied widely in facial recognition and the development of this technology has been made prominent over the recent years; it is now possible to identify faces with a great degree of accuracy in difficult orientations, lighting, and even expression [39]. In the proposed family of applications, deep learning models are employed for image sequences in order to detect the movement of objects in between frames in action recognition, video summary and surveillance systems [40]. Also, there are generative models like Generative Adversarial Networks used in image generation where they can generate images from the noise, which contributes to fields such as entertainment, design, and art. [41]. Such a variety of applications shows how deep learning dramatically changed the field of computer vision and contributed to the development of new technologies in different spheres for the purpose of automation of the process of visual recognition.

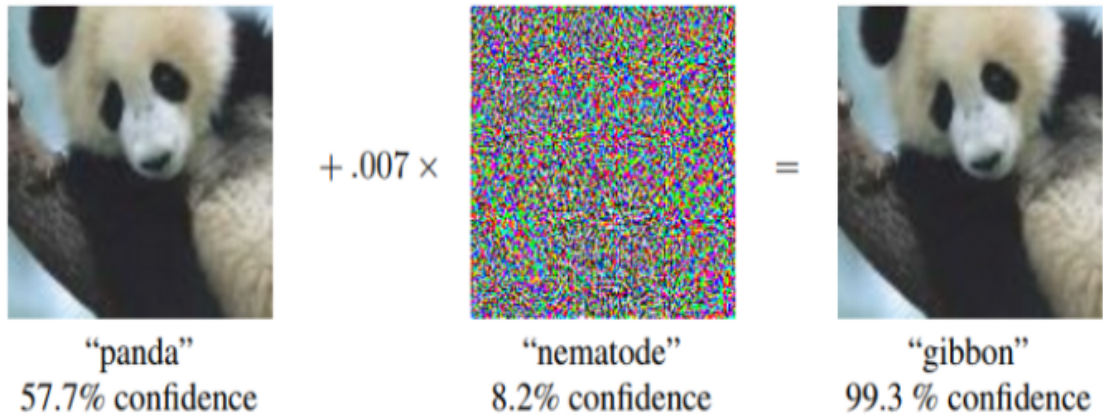
## 2.4 Adversarial Attacks

Adversarial attack entails skillfully altering an original image in a way that the modifications remain imperceptible to human vision. The resulting modified image, called an adversarial image, is incorrectly classified by a classifier, while the original image is classified correctly. These attacks can have serious real-world implications; for example, altering a traffic sign could confuse autonomous vehicles, potentially leading to accidents. Another concern is the possibility of illicit content being subtly altered to avoid detection by content moderation algorithms on major websites or by law enforcement web crawlers. The extent of alteration is typically measured using the  $l$  norm, which quantifies the maximum absolute shift in a single pixel [42].

### 2.4.1 White Box Attack and White Box Attack

In white-box attacks, the attacker has access to the model's parameters, allowing them to generate adversarial images with precise knowledge of how the model functions. Conversely, in black-box attacks, the attacker lacks access to these parameters. Instead, they create adversarial





**Figure 2.1:** Adversarial Example [43]

images using either a different model or no model at all, relying on the expectation that these images will transfer to the target model successfully[43].

#### 2.4.2 Targeted and Non-targeted Attacks

Non-targeted attacks aim to induce misclassification of the adversarial image by the model, while targeted attacks aim to manipulate the model into classifying the image as a specific target class different from its true class [44].

#### 2.4.3 Adversarial Attacks Common Type

Gradient-based methods are frequently employed in adversarial attacks. In these methods, attackers adjust the image according to the gradient of the loss function with respect to the input image [45]. There are two main approaches to conducting these attacks: one-shot attacks involve a single adjustment in the gradient direction, while iterative attacks involve multiple successive adjustments instead of a single step[46]

### 2.5 Related Work

There are different methodologies that have been suggested for the categorisation and summarisation of scenes in sports videos; cricket in particular. A good example of the formulation of the

techniques for specific application is the method for the scene classification introduced in the work on sports video summarization with the focus on cricket videos. The research entailed the classification of some of the vital events such as batting, bowling, and boundaries. The authors created a five classes dataset including batting, bowling, crowd, boundary and close-up to train a classification model [47]. They used transfer learning in which the pre-trained AlexNet CNN was used in the study and extended by three new fully connected layers for refining scene classification. To reduce overfitting, dropout was performed in the first two layers and to classify the final output the SoftMax activation was used [48]. As for sources of data, cases of augmented data were used to increase the accuracy of the model, which reached 99.26% but the model took longer to converge, in this small amount of data set. It was shown that the given model was much more accurate than other strategies, however, the basic issue that the approach had was connected with classification of frames which did not suit the best for one of the types. This approach demonstrates that Transfer Learning can be beneficial in sports video summarization but at the same time, it shows the problem of when the dataset is limited and specific to a certain area.

Another attempt to the automated Cricket video summarization focused on creating the summary by fragmenting the innings into Video shots in which the set of frames with respect to a single camera is considered. The match scores were detected through optical character recognition and the features that were used included scoreboards and audio cues [49]. The generated highlights were compared to official match highlights which showed the ability of the model to capture most of the important incidences like wickets, fours, and sixes with an accuracy of 89.45%. However, model failed to capture some boundary events and one of the models had issues with the processing speed because of the huge number of video streams that needed to be processed. This brings out the issues of real-time processing and the fact that there is need for proper data handling methods for large scale sports video summarization.

In another piece of work, similar to this, the authors had come up with a text-based, method of video summarization using semantic information. The approach combined deep visual-captioning with an extractive technique that was designed to obtain textonic summaries of video clips [50]. The authors captured the keyframes by employing an open-source vision computer library, OpenCV, and employed the scoring criteria that came with cinematographic rules to identify segments of interest. The frames were then reconstructed via a ResNet CNN model and then to the LSTM model to produce captions. These captions were also used in order to generate a comprehensible summary with the help of Python's Natural Language Toolkit (NLTK). How-

ever, this method proved quite useful in generating concise summaries of videos while at the same time, being resource-intensive because of the repeated frame analysis. This study shows that by incorporating both deep learning and natural language processing in their approach, there is a possibility of other methods of summarization if optimized will offer a better result.

From the latter, more advance in cricket video summarization has been made using the Deep Cricket Summarization Network (DCSN) with deep and reinforcement learning. The DCSN model used LSTM and CNN structures and reinforcement learning as the decoder with a keyframe selection [51]. Two bi-directional LSTMs employed enhanced the robustness of the model while 4/ 5 MOS showed that the users were satisfied with the system. But this approach incurred more space complexity and considerable performance overhead. Overall, the future work on video summarization with reinforcement learning is rather encouraging especially for dynamic and temporal nature of sports videos but the research is still open to several computational issues.

Another approach involved in generating likely cricket clips was molded on the event-based and excitement-based stimulus. This approach used CNNs and Support Vector Machine (SVM) frameworks where it was identifying essential incidents like wickets, boundaries, and sixes [52]. The concept was highly useful in segmenting games into specific scenes by employing OCR to read scores from the scoreboard. While this was helpful in often finding little occurrences, the model at times failed to capture less noteworthy events, which showed the highlights could be severely skewed by simple things like weather interferences. Such an approach points out the need to consider context, in case of event detection in video summarization, especially when the event is a sports event since the environment has a bearing on the game.

A different approach to video summarization is using the audio context and creating highlight based on the loudness of the content [53]. This method simply involved breaking the audio tracks into segments to get at sections where commentators were speaking and sections where the crowds were cheering. It split sections that it designated as 'exciting' and those that it considered 'not exciting' in regard to short-term energy computations. Although this approach offered a new way of generating summaries, it was less efficient compared with the video-based methods especially were the video was short and summarized basically based on the audio. The vulnerabilities of ASRF put forward the necessity of developing multimodal approaches that will include both, vision and hearing to improve the preciseness and the relevance of the generated preview.

Other than cricket, in the topical area of sports video summarisation, a work described deep

learning approaches for summarising soccer video. The researchers used jointly a 3D CNN and an LSTM for key-frame selection and ball tracking and they obtained an MOS of 4 or 5 for the video summaries [54]. As expected, the given model was simplistic and functional, but several issues involved computational implementation; most critically, ResNet based feature extraction. It applied more successfully to ball games such as soccer and basketball where the movement of the ball is a key feature of the story of the game. This approach shows the effectiveness of the deep learning methods on all the three sports while at the same time showing the compromise that has to be made between the model size and real-time processing.

Finally, one of the methods reviewed on the subject of advanced video compression was related to video summarization with the efficient selection of the keyframe using the FCSNs model [55]. This approach made use of common semantic segmentation networks for video summarisation and the method provided optimal accuracy as documented by experts in benchmark databases. LSTMs can be replaced by FCSNs in order to increase the parallelism in the model as such solution scales better on videos. Incorporating FCSNs can be seen as a major improvement to the kind of video summarization technologies, especially when it comes to timed problems and need for highly effective results [56].

# Deep Learning Models and Transfer Learning

This section provides a detailed explanation of the architecture of deep learning models, including their key components and structure. Additionally, it delves into the concept of transfer learning, discussing how pre-trained models on large datasets can be adapted to specific tasks, improving performance and efficiency by leveraging previously learned features and knowledge.

## 3.1 Deep Learning Models

Deep learning is an emerging field which consists of following basic models.

- Supervised models
- Unsupervised models

In supervised learning models, the training is done by feeding the model with data that it is expected to learn from and which has labels assiduously attached to each input and the outputs they are expected to produce. This makes it possible for the model to find the relation between the inputs and the outputs thus making it possible for the model to advance its knowledge with the new information it has not learnt from [57]. Most of the available algorithms stem from supervised learning; well-known models of this approach are the Multilayer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs). MLPs are feedforward artificial neural networks that comprise of more than one layer of nodes, to which each node is a neuron that takes in inputs and processes its data using a weighted sum and activation function [58]. CNNs, in turn,

are rather effective when it comes to structured grid data, such as images, because convolutional layers allow for automatic learning of the spatial hierarchies of features, for instance, [59]. The MLP and CNNs have widely used in different fields such as image, language, diagnosis where the data sets are well labeled.

However, the unsupervised learning models do not involve the set of labelled data during the training process. Rather, the input data are given to the model only and the model needs to find the patterns, structures, or relations in the input data on its own initiative [60]. The model tries to find out clusters, relationships, or dimensions that might be hidden in the input data as primary characteristics of it. Some of the typical examples of unsupervised learning methods are k means clustering, PCA and auto encodings. These models are especially used in exploratory data analysis, dimensionality reduction and anomaly detection, processes where one is typically more interested in mining patterns in the data rather than trying to predict the values of those patterns [61]. Another key advantage of unsupervised learning is that it is useful in situations where there is limited labeled data, or where getting a hold of such data can be very costly, which makes it very useful knowledge discovery toolkit in large data sets.

It is for this reason that we can see that the classification of learning methods into supervised or unsupervised based on the availability of labeled data is only half the story; the kinds of tasks they are suitable for are the other half. Unlike supervised learning which thrives in predictive tasks and where limited goals are set in terms of inputs to be mapped to outputs, unsupervised learning focuses more on discovering the latent patterns of the data. Supervised and unsupervised learning involves the considerate selection of the type of data available in the database, the nature of the problem under analysis and the expected output of the analysis [62]. They both are basic to the general subdiscipline of machine learning, and provide different methods for solving a vast array of practical problems.

## 3.2 Convolutional Neural Network

When dealing with pictures or Computer Vision tasks, a convolutional neural network (CNN) model is used. Numerous CNN applications have been created to assist with object detection, classification, and segmentation for applications such as human face detection, vehicle identification, and so on. CNN is often utilised in the medical field to aid in disease diagnosis. The general structure of CNN is shown in the Fig 3.1

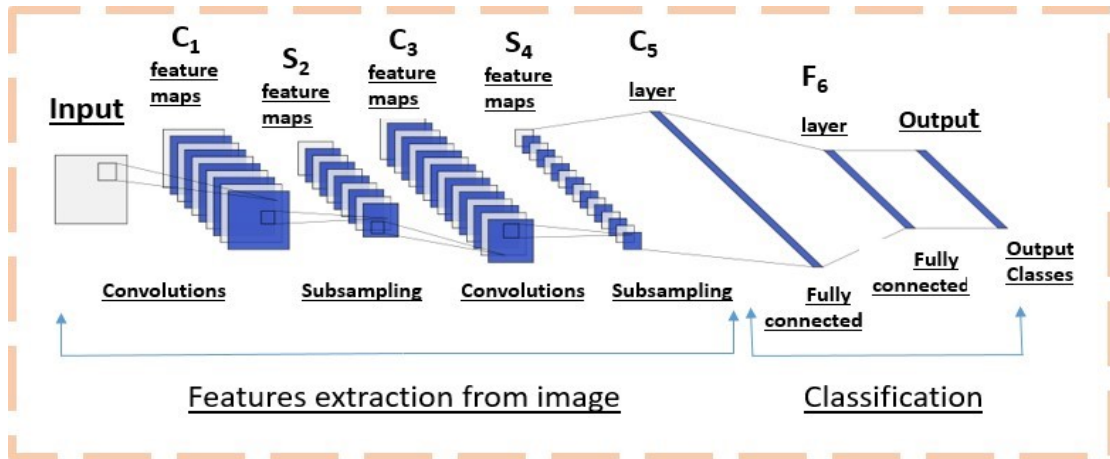
There are mainly three building blocks in the architecture of CNN: Convolution Layer: This layer is the most important layer of CNN and also known as 'Kernel' which is basically used to learn the features of an input image. It convolves a set of filters (or kernels for short) with the input image, where each filter is shifted over the image and every position of the filter performs dot product with a small patch of the input image to produce Activation Map or Feature Map. These feature maps are representations of some aspects of the image for instance edges, texture or patterns and due to the architecture of convolutional layers that add more layers the model is able to learn features at different levels of abstraction [63].

Max-Pooling layer: The Max-Pooling Layer as is known as subsampling or down-sampling layer is involved to reduce the size of the feature maps generated from the convolutional layers for height and width. It does so by employing a window (for instance 2x2) and extracts the maximum value of the feature map within the window and thereby eliminating and retaining lower order features. This process makes the computational load less by reducing the number of parameters and also in turn helps in improving the learnt feature towards obtaining a translation invariance to slight movements in the input image [64].

Fully Connected Layer: Fully connected layer is generally the last layer of a CNN and is responsible for the aggregation of the learned features that leads to the predictions. In this layer, each neuron is connected to every neuron in previous layer; thus it is advantageous for the network to oversee all the features that has been extracted in the convolutional and pooling layers. These features are converted to one-dimensional vector which shows the probability of different classes (in case of classification problems) through FC layer and activation function like SoftMax [65].

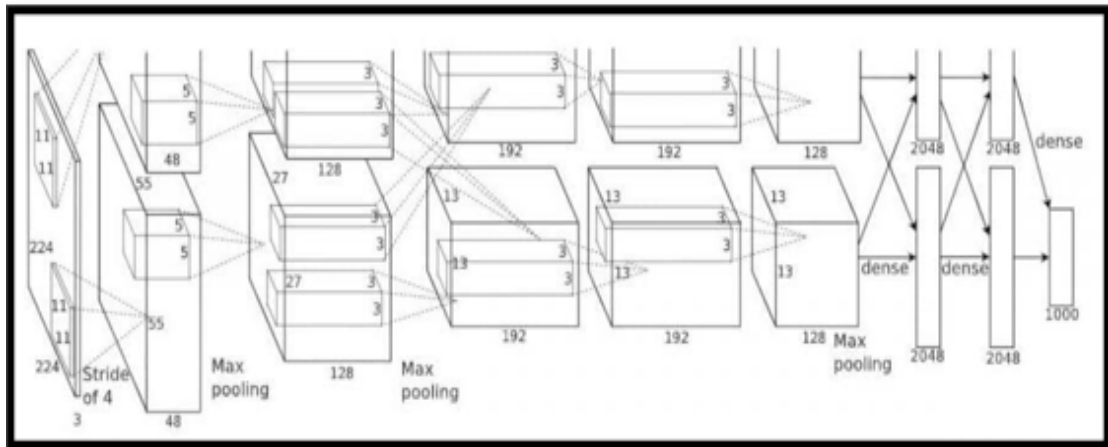
### 3.2.1 Alexnet

AlexNet was suggested in 2012 and was the winner of the 2012 ImageNet classification contest. It was the first time that a deep neural network was used to classify images in an imagenet. AlexNet has eight layers in total, excluding the pooling levels. Maxpooling is the pooling method employed. Five convolutional layers and three fully linked layers comprise this image. It employs a wide receptive field (11 x 11) and a small receptive field (5 x 5) in the initial layers and a smaller receptive field (3x3) in the subsequent layers. Each convolutional layer is



**Figure 3.1:** General Structure of CNN

followed by a ReLU activation function, and the final layer classifies 1,000 classes using the softmax function. The architecture of AlexNet is shown in the Fig 3.2



**Figure 3.2:** Architecture of Alexnet [67]

### 3.2.2 VGGNET

The Oxford Visual Geometry Group developed the VGG architecture, which is the state-of-the-art model in 2014. VGG was an evolution of the AlexNet architecture. VGG has fewer parameters than AlexNet due to the usage of a set of filters with 12 tiny receptive fields of size (3 x 3). On imagenet data, it achieved a top 5 test accuracy of 97.2 percent for the Classification challenge. Vgg has two different variants. One is Vgg-16 which is 16 layer dense model while the other one is Vgg-19 which is 19 layers dense network[66].



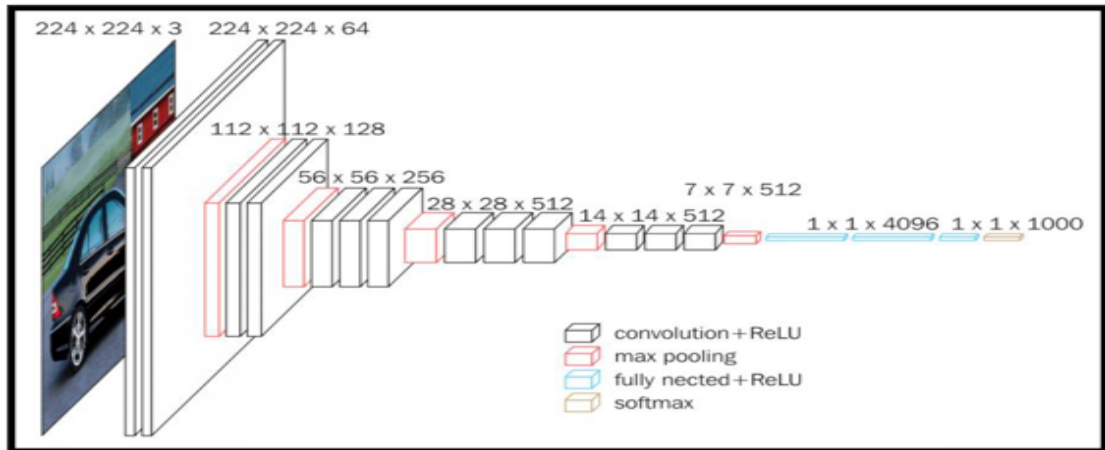


Figure 3.3: Architecture of VGG model [66]

### 3.2.3 DenseNet

Dense Convolutional Networks (DenseNet) require fewer parameters than typical CNNs since they never learn redundant mapping features. DenseNet's layers are relatively thin, consisting of 12 filters, which result in a smaller collection of new feature maps. DenseNet is available in four flavours: DenseNet121, DenseNet169, DenseNet201, and DenseNet264. The computational cost is minimised since each dense block is directly connected to the input image and loss function gradient [67].

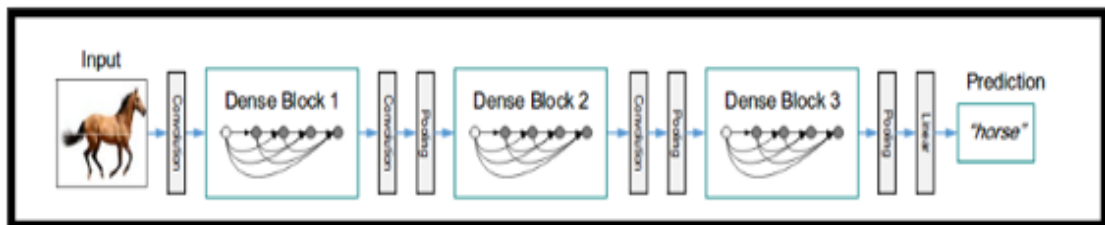


Figure 3.4: Architecture of DenseNet [67]

### 3.2.4 ResNet

The number of layers in deep learning models enhances the network's depth. However, when the network's depth increases, it runs into problems, culminating in a network that is extremely bad and inaccurate. The following are the most common challenges that a deep network encounter:

- Exploding/Vanishing Gradient: As the number of layers in the network rises and the weights are changed, gradients become unstable. As a result of continuous multiplication,

the gradient value may climb to an infinitely large value or shrink to an eternally small value, and the weights are not updated.

- Network Degradation: Another issue is that as you get deeper into the model, its performance begins to deteriorate. The network's accuracy suffers as a result of low performance.

Microsoft created a network called Residual Neural Network (ResNet) to overcome the challenges outlined[68]. The ResNet architecture's basic concept is to use skip connections after numerous levels. After a few more levels, the output from the preceding layers can be applied as is. As a result of the increase in depth, the exploding /vanishing gradient problem is avoided, as well as network performance decrease. These leftover blocks are simply put together to increase the depth of the networks. The input and output dimensions are the same in an identity skip connection. Resnet has been released in many versions, such as resnet50, resnet101, and resnet151. The only distinction between them is the number of layers.

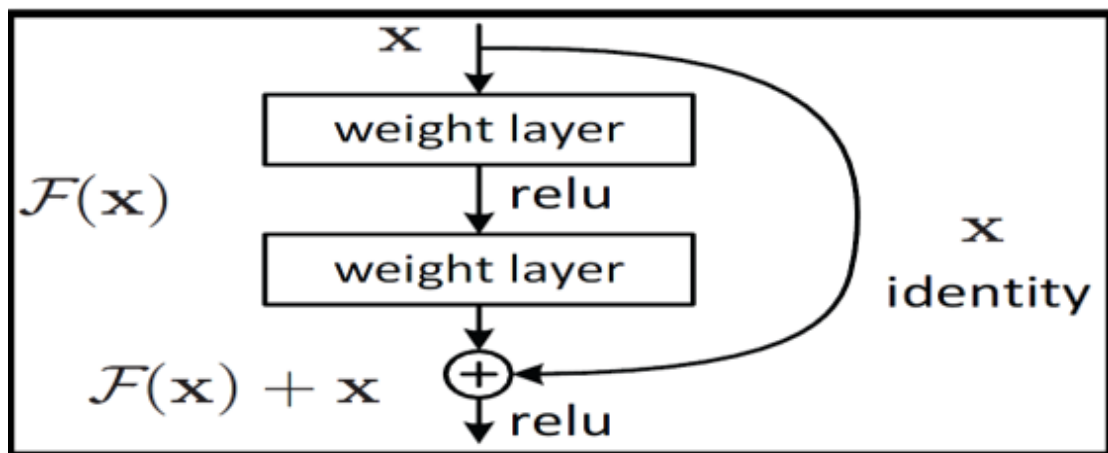


Figure 3.5: Architecture of Residual block [68]

### 3.3 Transfer Learning

When faced with a huge dataset, it is well established that deep transfer learning CNNs outperform smaller networks. As a result, transfer learning is frequently used when smaller datasets are available. The following figure can assist in comprehending the concept of transfer learning. Without sacrificing efficiency, a model trained from a bigger dataset (ImageNet) can be put to use for smaller datasets. Transfer learning has recently been used to for many image recognition tasks as well as for medical images for disease diagnosis[69].

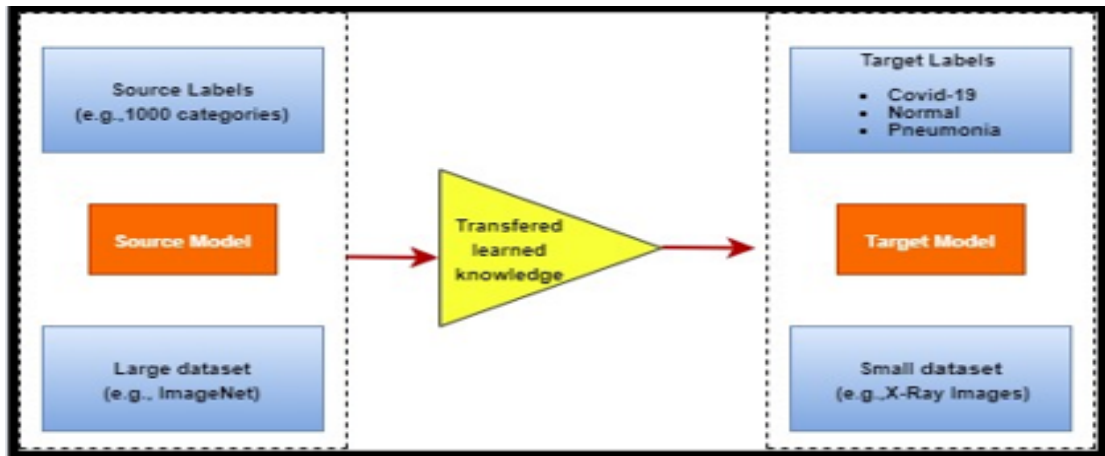


Figure 3.6: Block diagram of Transfer learning [69]

### 3.4 Overview of 3D CNN

3D Convolutional Neural Networks (3D CNNs) consist of; the 3D convolutional layer whereby a three-dimensional kernel is moved over the input volume. While common 2D CNNs are only capable of identifying patterns in terms of height and width of the networks inputs 3D CNNs are well equipped to recognize patterns of motion over the networks inputs width, height and depth. This capability is important especially where volumetric data is involved for instance in medical imaging, when some sections of MRI or CT scans have to be reconstructed in 3D, or in video processing where temporal motion between frames is as important as spatial [70]. Depending on the functions, 3D CNNs can process data in three dimensions and provide detailed patterns and structures which always play a big role in high-dimensional data in some fields.

Besides the 3D convolutional layers, 3D CNNs include 3D max pooling layers or strides as a way for down sampling in space. This process performs a great role in minimizing computational cost especially while processing big data while at the same time preserving important details that are key ones and which dictate accurate processing and analysis. In three dimensions Max pooling enhances the extraction of features while omitting unnecessary subordinate details which is very advantageous when receiving high resolution signals, for instance medical scans or video signals since it is very computationally expensive [71].

In addition, the highly developed 3D CNN architectures have been adopted with skip connections, popular in the ResNet structures, allowing to combine the accurate detail in the different level of a network. Skip connections reduce vanishing gradient by allowing the gradient to pass through layers of the network, during backpropagation, and hence, to make deep architectures

more learnable. At the final layers, fully connected layers could commonly be used to perform the designated actions of the model for instance, classification, regression, or segmentation. These fully connected layers accept the representations of learnt features from the convolution and pooling layers and transforms them to the desired output making the 3D CNNs applicable in a vast of field including object recognition in 3D worlds, actions prediction in video sequences among others [72].

### 3.4.1 Applications

**Video Summarization:** 3D CNNs seem to be especially appropriate for tasks such as video summarization which require capturing spatial relations as well as temporal ones. Thanks to the fact that such networks process the video sequences as 3D data, it can be effectively used for identifying the major events, action and transitions which facilitate the process of generating the summaries of the video material by default.

**Action Recognition:** In the action recognition problems, 3D CNNs take the role of extracting the information from the video inputs where the human actions are involved. It makes 3D CNNs suitable for this purpose given that they are equally capable of capturing temporal properties of a video depending on the variations in the motion and interactions over time.

**Medical Image Analysis:** It is noteworthy that COVID-19 3D CNNs are most significant in analysis of the volumetric data that is obtained from MCT, MRI and ultrasound, and other imaging methods. In this case, these networks may compartmentalize organs, inform and diagnose the presence of disease based on the structure's three-dimensional shape, which renders crucial visual information for health care provision.

**Other Computer Vision Tasks:** Apart from video summarization and medical imaging 3D CNNs have found applications in other computer vision problems such as detection of objects in 3D space, depth map estimation and even self-driving cars where perception of 3D is vital.

### 3.4.2 Challenges and Considerations

**Computational Demand:** One of the largest problems concerning 3D CNNs is the fact that they are computationally expensive. 3D data consumes a lot of memory and time during training of the model compared to 2D data since it is more complex.

**Data Requirements:** Due to the realities of 3D, the datasets are commonly significantly bigger, so data management is both different and requires substantial storage. Similarly, getting and labeling 3D data, especially in logic, can be expensive, and especially in medical imaging.

**Model Complexity:** In this paper, the various shortcomings of 3D CNNs are described, which includes over fitting especially when the network is trained under a limited data set. The model should not over-fit, so appropriate model design such as the use of regularizations techniques as well as data augmentation needs to be implemented.

### 3.4.3 Conclusion

3D CNNs are particularly useful because they can be applied in various domains as a potent tool for processing 3D data. To begin with, in tasks like summarizing videos, recognizing actions, as well as diagnose the medical conditions, 3D CNNs are excellent in representing spatial temporal relationship that are key to these tasks. However, as the champions for the most difficult and intricate computer vision tasks, their value in modern deep learning perspectives can hardly be overestimated.

# Proposed Methodology

In this section, we explain the procedure used to collect the dataset as well as pre-processing of the data and the models used in classifying events in cricket videos. It is discussing several deep learning architectures that would be powerful for the act of analyzing video. Among the models that had been considered include ResNet-34+LSTM, 3D Convolutional Neural Network (3DCNN), ResNet-50 + LSTM, X3D, ResNet-18 based 3DCNN. These models are selected to make the best use of convolutional networks for feature learning and the temporal analysis which comes with the LSTM or 3D convolution for classifying events in the cricket videos

## 4.1 Dataset

### 4.1.1 Dataset Collection

Cricket matches, which are longer in duration as compared to football matches, offer a wealth of video data that can be accessed through different platforms including YouTube, Hotstar, and Dailymotion. For this study, we initially downloaded long-format cricket match videos from YouTube, encompassing all three major formats: Country profiles also contain information about the test, one day internationals (ODI), and T20 cricket matches. Furthermore, to avoid data bias or lack of sources, some selected Pakistan Super League (PSL) matches were also incorporated as the study sample.

Host video clips were manually labeled using open source OpenShot Video Editor for classifying various instances of cricket events. The dataset was organized into five distinct classes: Here we have four classes of Fours, Sixes, Wickets, and Milestones, and an additional class called “Nothing”. Annotation process includes deciding on the segments of text that belong to

the particular class. For instance, in the “Four” class, the clip would begin with the run-up to the bowler, and would end with the ball being hit to the boundary by the batsman. As many various shots and directions as possible were collected, and each class was compiled from 200 clips.

The same criteria were applied to the “Six” class: 200 clips from different matches with variation of shot type and trajectory were chosen. The “Wicket” class was more diverse, for it consisted of 200 clips of various forms of dismissal like bowled, caught, run out, and stumped sourced from all forms of the game. The “Milestones” class recorded memorable events, such as batsmen completing fifty or hundred, bowlers taking three wickets in three balls or any other landmark, with 200 clips. The “Nothing” class consisted of miscellaneous events like advertisements, singles, doubles, drinks breaks, and other instances which are not suitable for match descriptions.

In the given dataset, each video clip is about five to six seconds long, and the frame rate is 25 frames per second. To sum up, the dataset is 1000 video clips and belongs to five classes where each class has 200 video clips. Basically, the proposed dataset comprises of all the raw videos in clips with test cricket, ODI cricket, one T20 match from PSL, and all the events related to cricket including batting, bowling, fielding, wicket taking, etc. The following Fig 4.1 shows some of the frames of each class in the dataset.



**Figure 4.1:** Frames from our own dataset

#### 4.1.2 Dataset Preprocessing

After the collection of dataset for the five classes: Fours, Sixes, Wickets, Milestones and Nothing consists of 1000 videos (200 videos each class), the first phase of data preprocessing was done. Firstly, we obtained the arithmetical mean of the length and the mean number of frames in all the videos. The study shown that average duration of the videos was almost six seconds

making the videos to have an average of 150 frames per video.

To meet this clean process, and make all the video clips of the same format, we underwent through frame adjustment. To videos which had frames less than 150, we repeated the last frame in an attempt to make the length of the video equal 150 frames. On the other hand, for the videos with frames more than 150, we only considered frames up to 150 frames only. This preprocessing step was very useful in order to sort out the scores and ensure that there was less possibility of biases within the data set while the model was being trained or while the summary was being generated.

After standardizing the frame count, we converted the videos into arrays, creating one array per class: an array for fours, an array for sixes, an array for wickets, an array for milestones and an array for nothing. These arrays were down-sampled to match the input dimensions of these models, next to the input requirements. In particular, each of them was resized to a 112 by 112 pixels resolution. Therefore, the input data was reshaped to tensor of  $l \times h \times w \times c := 150 \times 112 \times 112 \times 3$  for each video. Therefore, the last dimension of input shape for each video clip was  $150 * 112 * 112 * 3$ .

The decision of the execution of this preprocessing strategy was to bring the video clips into equal lengths, format, and resolution used by the classification models and for proper event detection in the cricket videos.

## 4.2 Model

### 4.2.1 ResNet-34 + LSTM for Event Classification

Today deep learning models are widely used everywhere, and they are employed in many areas such as, object identification, face recognition, human motion recognition, pose estimation, human computer interaction, motion capture, augmented reality, video processing. In this work, we lay our emphasis on classifying sports action from cricket videos using a ResNet-34 deep learning model with LSTM layers [73]. Cricket is a complex sport with a wide variety of events, but this study narrows down the focus to five specific events: As the century partner, Fours, Sixes, Wickets, Milestones, and Nothing. Describes of these events have been explained elaborately in preceding sections of this chapter.

The approach using ResNet-34 followed by LSTM consists of three main steps: The approach



using ResNet-34 followed by LSTM [74] consists of three main steps:

1. **Input Video Processing:** The input video has a size of [1501121123] and given to ResNet-34 model to extract the feature. This input size equals the size of the video equivalent to 150 frames, each with 112-by-112-pixel resolution and three color bands (RGB).
2. **Feature Extraction and Temporal Modeling:** For spatial feature extraction, the ResNet-34 model is used, which is a convolutional neural networks introduced by He et al. in the year 2016. ResNet-34 as the abbreviation of residual network is a family of residual networks created to overcome deep learning issues including the vanishing gradient problem through the use of residual connections. The architecture consists of 34 layers: the first is a convolutional layer, and afterwards a series of residual network blocks. In each block, Convolutional layers, batch normalization are used followed by ReLU activation functions and lastly a fully connected layer for classification. The feature maps generated from ResNet-34 then go to the LSTM layers that have the role of learning temporal dynamics of the video streams.
3. **Classification:** The features over the time axis identified by the LSTM layers are followed by a fully connected layer that has a SoftMax activation. The final step of the proposed KNN algorithm is to classify the video input in one of the five defined classes.

Here the ResNet-34 model for image classification is employed by using LSTM layers for the temporal feature of videos for video event classification. This approach is in resonance with [74] whose network involved the use of ResNet-34 + LSTM for soccer video classification into five categories. But as again due to the difference in nature of Cricket and soccer video, the size of input in our case is bigger that is  $150 \times 112 \times 112 \times 3$  Where as Agyeman et al used 64 frame input for soccer video classification.

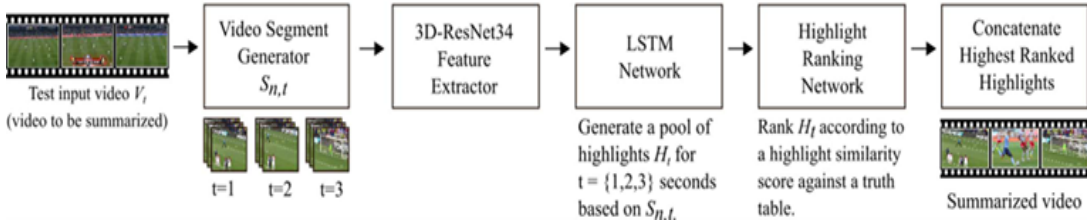
In the implementation, resnet-34 structure is used which is composed of an initial convolution layer, four residual blocks. They are made of two convolutional layers, batch normalization layers, ReLU activation layers and have residual connections to avoid the vanishing gradient problem and to improve the features extraction. The output feature maps of ResNet-34 become subsequently passed to LSTM layers that extract temporal features. Last, but not the least, the layer is the 5ne layer that is connected fully to provide classification into the desired events based on SoftMax activation function. The following is the tabular implementation of different parameters in each layers of the ResNet-34 + LSTM model. In the first stage of residual block, we used the 64 filters and then kept on increasing in the next block .in the second stage we used 128 filters. In the third stage we used 256 filters. In the fourth stage we used 512 filters.

Layer name	3D-ResNet34 layer parameters
conv 1	1x1x3, 64
conv 2_x	$[3 \times 3 \times 3, 64]$ $[3 \times 3 \times 3, 64] \times 3$
conv_x	$[3 \times 3 \times 3, 128]$ $[3 \times 3 \times 3, 128] \times 4$
conv 4_x	$[3 \times 3 \times 3, 256]$ $[3 \times 3 \times 3, 256] \times 6$
conv 5_x	$[3 \times 3 \times 3, 512]$ $[3 \times 3 \times 3, 512] \times 3$
Average pooling layer	stride = $1 \times 1 \times 1$
Output layer	5 classes, Softmax activation

**Table 4.1:** 3D-ResNet34 Layer Parameters

Then average pooling layers to reduce the dimension making it suitable for using as input to the LSTM layer. The following Fig 4.2 shows the steps of cricket event classification.

After the classification into desired events, we concatenated the events of our interest by dis-



**Figure 4.2:** Block diagram of ResNet34 + LSTM

carding the nothing class and concatenating the other events which are named as high ranked events to the new video which is summarized video of long input video. Results of this model are discussed in the results section.

#### 4.2.2 ResNet-50 + LSTM for Sports Action Classification and Video Summarization

Deep learning models has been widely used in different fields such as, object recognition, facial identification, human activity analysis, human pose estimation, human computer interaction, motion capture, augmented reality and video surveillance. In the present work, we provide

our attention to sports action classification in cricket videos employing deep learning model Resnet-50 coupled with LSTM. Cricket game being a large complex game consisting of notifiable number of events makes it a difficult event type classification. For this study, we concentrate on five specific events: Fours, Sixes, Wickets, Milestones and a ‘Nothing’ class which has been described in detail in this chapter.

Our approach using ResNet-50 + LSTM is structured into three main steps: Our approach using ResNet-50 + LSTM is structured into three main steps:

1. **Input Video Processing:** The input video of a resolution,  $150 \times 112 \times 112 \times 3$ , is passed through the ResNet-50 model to obtain spatial features. This input size amounts to 150 frames of video clip of resolution 112x112 pixels with the RGB color channels.

2. **Feature Extraction and Temporal Modeling:** ResNet-50 is another feed forward convolutional neural network introduced by [75] which belongs to the residual networks category of networks and was developed specifically for higher accuracy image classification. ResNet21 can be constructed as a network of 50 layers; There are 48 convolution layers, 1 max-pooling layer, and 1 average-pooling layer. It consists of several residual blocks wherein each block consists of three convolutional layers, batch normalization layers, and ReLU activation layers. The residual connections reduce the vanishing gradient problem apart from improving feature learning by enabling the network to retain a number of features across the layers within the neural network. The feature extracted from each frame using ResNet-50 are feed to LSTM layers to establish the temporal relation in the video sequence [76].

3. **Classification and Summary Generation:** The output of temporal features extracted by the LSTM layers is feed to a fully connected layer with a SoftMax activation function which categorizes the input video into one of the five defined classes. In the residual blocks, the number of filters increases progressively: The first stage of the encoder has 64 filters, the second one has 128 filters for the third has 256 filters and the final encoder has 512 filters.

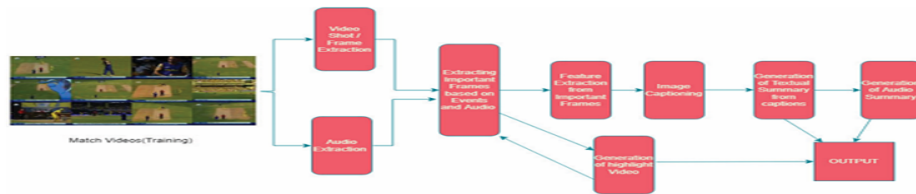
Our approach is motivated by the work of [77] who used ResNet-50 and LSTM for cricket video summarization. They used audio intensity as a cutoff to divide the video content into ‘Excitement’ and ‘Non-Excitement’ events in their work. The exciting segments were then joined in order to obtain a video-based summary and an audio-based summary of the video was developed using Google Text-to- Speech API after coming up with captions from the LSTM model.

Although the approach of selecting the events is similar to ours used in the compilation of the

Layer Type	Output Size	Filter Size / Stride	Number of Filters	Activation Function	Number of Parameters
Conv1	112x112	7x7 / 2	64	ReLU	9,664
MaxPooling	56x56	3x3 / 2	-	-	0
Conv2_x	56x56	1x1, 3x3, 1x1	64, 64, 256	ReLU	70,400
Conv3_x	28x28	1x1, 3x3, 1x1	128, 128, 512	ReLU	1,280,000
Conv4_x	14x14	1x1, 3x3, 1x1	256, 256, 1024	ReLU	7,110,400
Conv5_x	7x7	1x1, 3x3, 1x1	512, 512, 2048	ReLU	14,348,800
Average Pooling	1x1	7x7	-	-	0
Fully Connected	1x1	-	5 (classes)	Softmax	10,245

**Table 4.2:** Layer Parameters for ResNet-50

cricket video dataset, the input is a slight different because of the context and events within the game of cricket. The input video has dimensions of  $150 \times 112 \times 112 \times 3$ , and we utilise ResNet-50 to obtain spatial features. These features are then passed through LSTM layers in order to capture temporal patterns and for the final classification of the data into the required event using a last fully connected layer with SoftMax activation. Four classes of information that are obtained after the classification process include “Yes – indicator”, “Yes – clarification”, “No – clarification”, and “No – indicator”. The “Nothing” class is eliminated and the remaining four classes are combined to produce a summary video from the long input video. Moreover, this



**Figure 4.3:** Block diagram ResNet-50 + LSTM

also helps in the better classification of events in cricket video and helps in a more efficient and more effective summarization by concentrating on only events that have happen, thus making the summary more compact, and yet informative enough to give the overall idea of the match.

### 4.2.3 3DCNN Model for Sports Action Classification

The approach we decided on with the 3DCNN model is a much more ordered approach, in fact, we have a: Firstly,  $150 \times 112 \times 112 \times 3$  input video is given into the model for feature extraction. 3DCNN has indeed some properties making it suitable for video analysis, as 3DCNN can extract moving features, processing both spatial and temporal fields at a time. With the

extracted features, a fully connected layer with SoftMax activation function categorizes the input to be in one of the described class among the five classes.

This architecture known as 3D CNN was first proposed by [78] for human action recognition of videos. This is beneficial to the analysis of videos since the 3DCNN architecture can handle resolution in the height and width as well as the time domain hence it is able to capture temporal change well enough to classify human actions.

Vrskova et al. [79] went further in analysing the effectiveness of 3DCNNs and used the model to segment human activities into eleven classes. In their study, the input to the 3DCNN had a shape of  $70 \times 32 \times 32 \times 3$  where in 70 is the quantity of frames per video, 32 is the width and the height of each frame and 3 is the quantity of channels which is RGB. They showed in their works that it is possible with 3DCNN to classify human actions with high precision utilising spatial and temporal characteristics of actions. In our work, we have adopted the above mentioned approach



**Figure 4.4:** Block diagram of 3D-CNN

formulated by [79] for the classification of cricket video events. Nevertheless, since the nature of the events is different, we had an input of greater size,  $150 \times 112 \times 112 \times 3$ . The input video is passed through an initial 3D convolution layer and is then through other 3D convolution layers and lastly followed by a MaxPooling layer for further extraction and fine tuning of the features. As for the “Nothing” class, they are removed after the classification, and the rest are concatenated to create a summary video from the longer input sequence.

The following table gives the overview of the layers and parameters of the 3DCNN model used in this study. Besides improving the method of classifying events, this approach also helps in the formation of short summaries of videos concentrating on significant events.

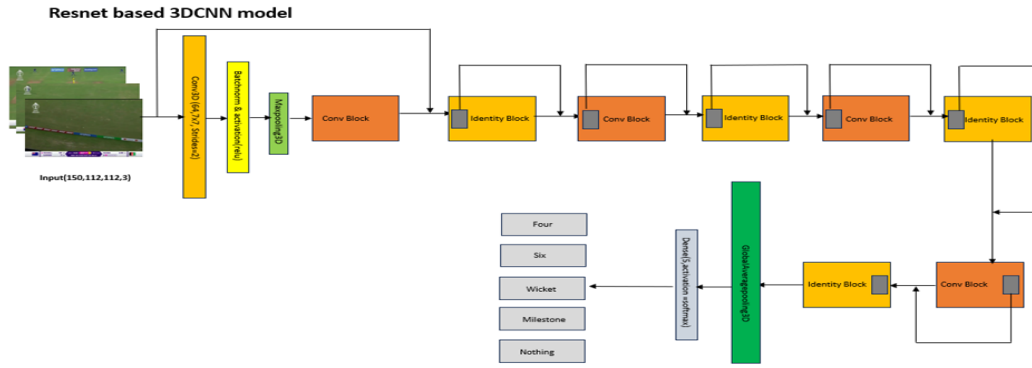
Layer Type	Output Shape	Parameters
Conv3D	$123 \times 62 \times 62 \times 64$	5,248
MaxPool3D	$122 \times 61 \times 61 \times 64$	0
BatchNorm	$122 \times 61 \times 61 \times 64$	256
Conv3D_1	$120 \times 59 \times 59 \times 64$	110,656
⋮	⋮	⋮
Conv3D_5	$109 \times 53 \times 53 \times 512$	131,584
MaxPool3D_3	$54 \times 26 \times 26 \times 512$	0
BatchNorm_3	$54 \times 26 \times 26 \times 512$	2,048
Dense	$54 \times 26 \times 26 \times 256$	131,328
Flatten	9345024	0
Dense_1	5	46,725,125

**Table 4.3:** Summary of 3D CNN Model Layers and Parameters

#### 4.2.4 ResNet-18 based 3D CNN

The models of deep learning have become fundamental in virtually every kind of application, including but not boundary to object detection, face recognition, recognizing human actions, estimating human pose, interactions between human and computers, motion capturing, augmented reality, and video analysis. Hence in this present study, the analysis is concentrated on the sports action classification in cricket videos adopting a ResNet-18 based 3D Convolutional Neural Network (3DCNN) model. Indeed, because cricket is a highly complex game with many different events, the process of classifying events adds certain difficulties. For this research, we target the classification of five specific events: There are Fours, Sixes Wickets, Milestones and a ‘nothing’ class which form the work space and have been described in this chapter in the earlier sections.

The method for working with the ResNet-18 based 3DCNN model is quite structured and can be divided into three steps. First, the input video, measured  $150 \times 112 \times 112 \times 3$  The values  $150 \times 112 \times 112 \times 3$  are used as input of the model for the feature extraction. The 3D convolutional layers are used to obtain the both spatial and temporal features from the frames of the videos. Last of all, the extracted features go through one fully connected layer with the Soft-Max activation function to provide the input with one of the five predefined classes. The ResNet



**Figure 4.5:** Block diagram of ResNet-18 based 3D CNN

architecture was disclosed by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun in their path breaking work on deep residual learning for image classification [80]. To establish ResNet they used 2D CNN layers ,which were mostly appropriate to image classification since they highly depend on spatial feature maps. Their work showed that the idea of residual connections helped to provide substantial gain in the image classification over other types of CNN architectures because of reduction of the vanishing gradient issue.

In our work, we choose ResNet-18 as the baseline architecture introduced in the paper by He, et al. [80], which is pre-trained on ImageNet and fine-tuned to the video classification task. In particular, we substitute the 2D CONV layers with 3D CONV layers to learn spatial as well as temporal features required in video processing. Furthermore, we add a global average pooling layer here to downsize the feature maps for fully connected layer input. Finally, the last five nodes connected fully connected layer with SoftMax activation function Output layer classifies the input video into one of the five desired classes.

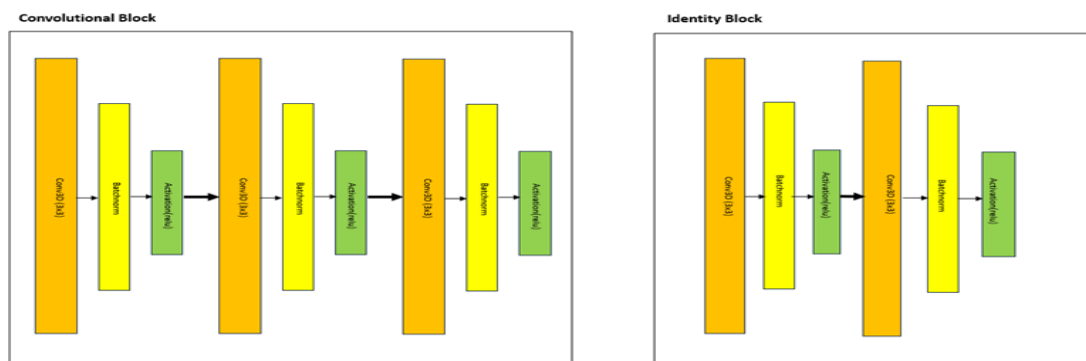
The network architecture of the modified ResNet-18 based 3DCNN is built in a way that takes in a sequence of frames forming a video clip. In the case of the repetitive block, each frame is scaled to dimensions of

150×112×112, three channels in regard to the RGB model. The model starts with the 3D convolution layer, which is worked out with a filter size of 7×7×7 and a stride of 2. This layer is developed for processing the video input and extracting the features that are spatiotemporal in nature and involves convolutions in both the spatial domain that is height and width of the video frames as well as in the temporal dimension that is frames of the videos. This is then followed by a batch normalization to stabilise and accelerate the learning process and a ReLU activation function for non linearity.

To explain the flow of ConvNet, after the initial convolution, there is a 3D max-pooling layer that takes place with the aim of decreasing the size of the feature maps, and hence, the computation complexity. The centre of the model comprises of several Residual blocks; this is because ResNet comprises of residual blocks. Each residual block comprises two types of sub-blocks: There are two types of RNN called as convolutional blocks and identity blocks.

The convolutions followed by batch normalization and ReLU activation steps are aimed at the extraction of features inside the residual block, and they are composed of several convolutions. The identity blocks, in contrast, use identity mapping and the function of the convolutional block is added to the input by concatenation. This connection persists allows gradients to flow through the network more efficiently, which solves the vanishing gradient problem that affects deep net.

By the use of convolutional and identity blocks, it becomes easier to learn more features while at the same time, reusing the features which makes it possible to contain the parameter count to a reasonable level. After all the residual blocks, a global average pooling layer is used to reduce the dimensions of the feature maps to one dimension per feature map, so as to produce a concise feature vector that captures the learned features over the entire video sequence. The final aspect



**Figure 4.6:** Block diagram of Convolutional and Identity Block

of the model is a dense layer to which inputs from the global average pooling layer are fed into. A SoftMax activation function is then applied to produce a probability distribution across the predefined classes: The outputs as same to the labels are ‘Four,’ ‘Six,’ ‘Wicket,’ ‘Milestone,’ and ‘Nothing.’ The highest probability class acts as the classification label to the input video clip.

The following table specifies the layers of residual network based on 3D convolutional neural network model and parameters related to those layers are also given.



**Table 4.4:** Model Summary

<b>Layer Type</b>	<b>Count</b>	<b>Hidden Units</b>	<b>Kernel Size</b>	<b>Activation</b>	<b>Parameters</b>
Input Layer	1	-	-	-	0
Conv3D + BatchNorm + ReLU	3	64	(3, 3, 3)	ReLU	176832
MaxPooling3D	1	-	(2, 2, 2)	-	0
Conv3D + BatchNorm + ReLU	2	64	(3, 3, 3)	ReLU	220512
Conv3D	1	64	(1, 1, 1)	ReLU	4160
Add + ReLU	2	-	-	-	0
Conv3D + BatchNorm + ReLU	2	128	(3, 3, 3)	ReLU	664320
Conv3D	1	128	(1, 1, 1)	-	8320
Add + ReLU	1	-	-	ReLU	0
Conv3D + BatchNorm + ReLU	2	256	(3, 3, 3)	ReLU	2654720
Conv3D	1	256	(1, 1, 1)	ReLU	33024
Add + ReLU	2	-	-	-	0
Conv3D + BatchNorm + ReLU	2	512	(3, 3, 3)	ReLU	14156800
Conv3D	1	512	(1, 1, 1)	-	131584
Add + ReLU	2	-	-	ReLU	0
GlobalAvgPooling3D	1	-	-	-	0
Dense	1	5	-	Softmax	2565

# Discussion and Results

## 5.1 Discussion of Results

In this chapter, we describe the performance of a range of deep learning models on the problem of cricket video event classification. These models comprises of Soccer video summarization model, Attempts of X3D model with different configurations, 3D-CNN model and ResNet-18 based models. Based on the performance of models on cricket dataset, the accuracy is calculated and analyzed to deduce the summary and comparison of each model.

## 5.2 Training Procedure and Hyperparameters

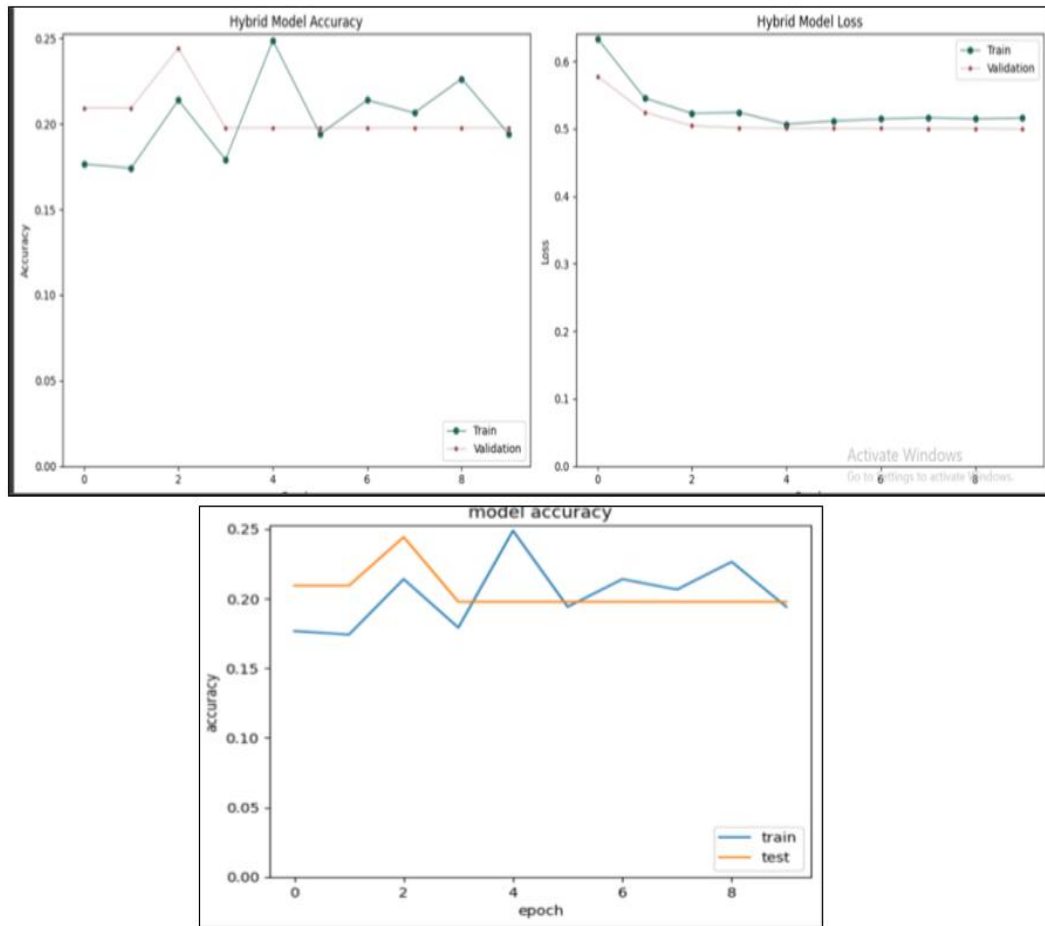
This section explains training and evaluation process for the proposed 3D Convolutional Neural Network (3DCNN) models to achieve high generalization. The dataset was split into three subsets: 80% was used as the training set, 10% as the validation set, and the remaining 10% as the test set. This distribution was chosen to ensure a rich set of data for training while leaving enough data for testing, preventing any skew in the training results.

The models were trained for 30 epochs, with a learning rate set to 0.0001, a value chosen after preliminary experiments to ensure the stability of the training process and to prevent overshooting the optimal minima. The batch size was set to 4, which allowed efficient use of memory while ensuring that the number of gradient updates in each epoch was reasonable.

An early stopping mechanism was applied, with the quantity of interest being validation loss (val loss). This approach helped avoid overfitting by stopping the training when the model failed to improve the validation loss, thus maintaining the generality of the model.

### 5.2.1 ResNet-34 + LSTM on Our Dataset

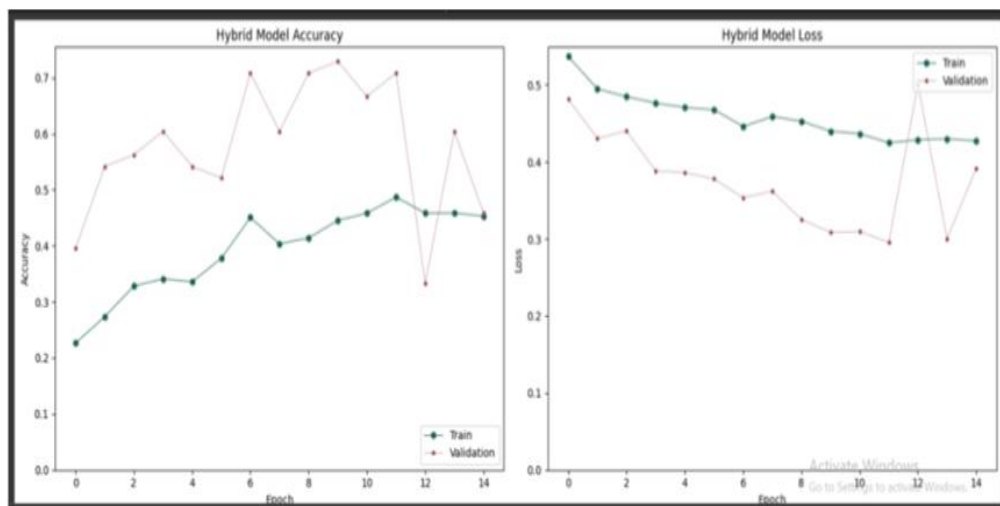
Initially, we had a model that has been used in soccer video summarization and we used it on a cricket dataset that we have. , as indicated in the Fig 5.1 below, the proposed model gave an accuracy of only 20%. This low accuracy can be attributed to the fact that the soccer model learned a number of features that are rather good at representing soccer videos but are rather irrelevant to cricket which is fundamentally different in its dynamics and event types. Soccer and cricket also have a difference in terms of the structure of events; in the hierarchy of actions, cricket contains more and more diverse actions. It is claimed that Soccer model possibly tended only to such key features of soccer that have little relevance to cricket or even are misleading if relied upon. This result implies the necessity to introduce corrections to the models which should reflect the specifics of the target sport.



**Figure 5.1:** The soccer model accuracy on our dataset of cricket.

### 5.2.2 X3D Model

Next up we describe the X3D models which are intended to be linearly and logarithmically scalable video data processing pipelines. We also evaluated it by variations, such as the size of X3D-S, X3D-M, and X3D-XS. Of all these models, the architecture that has delivered the best results according to the test run in Fig 5.2 was the X3D-M model having a test accuracy of 33 percent. Although this is slightly more accurate than the soccer model, the results are not high enough to assert that the X3D model is completely tailored for the demands of cricket videos. We get the advantage that the X3D model can be scaled efficiently, which is very crucial, but might need hyperparameters adjustments or a change of architecture to capture long-term dependencies and different structures of events in the cricket domain. Furthermore, the moderate performance might denote that the outlined X3D model architecture is indeed strong but may require injection of the specific field knowledge or more data on cricket events to train the model. Fig 5.2 shows the training accuracy and loss function of our considered model of X3D-M on the



**Figure 5.2:** X3D model accuracy on our dataset of cricket

cricket dataset used in this work.

### 5.2.3 3D-CNN Model

The 3D-CNN model in this paper was originally presented in a paper where it was used for classifying 11 human activity classes. In the development of this model, specifically used when the classifier is applied to our cricket dataset, the basic accuracy level that was obtained to 61% as seen in figure 3 below. Having tuned the learning rate, the batch size as well as the frame resolution we managed to get the accuracy of the model to 67%. This big leap underlines one of the most essential aspects of deep learning environments – the fine-tuning of hyperparameters. The 3D-CNN model that takes into account the spatial and temporal contexts yielded good results on the cricket events, which are bound both by space and time. That 's why everyone got an improvement in the value of accuracy after the optimization of the model, which shows that 3D-CNNs are appropriate for video classification since both spatial and temporal features does very important role.

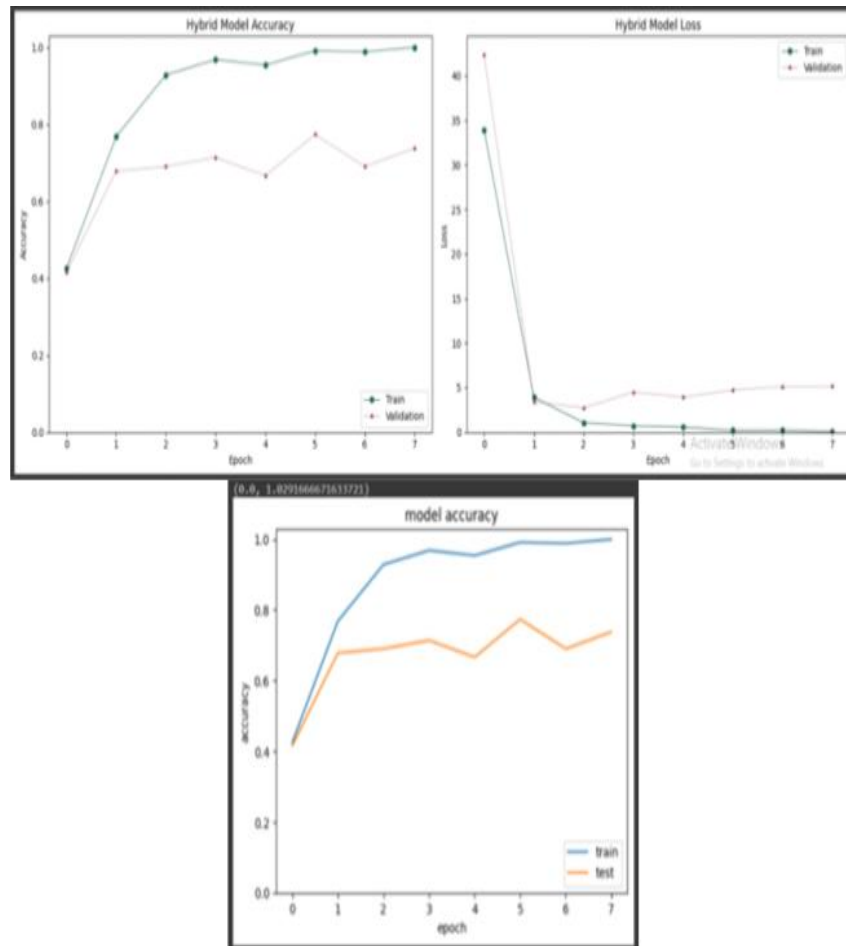
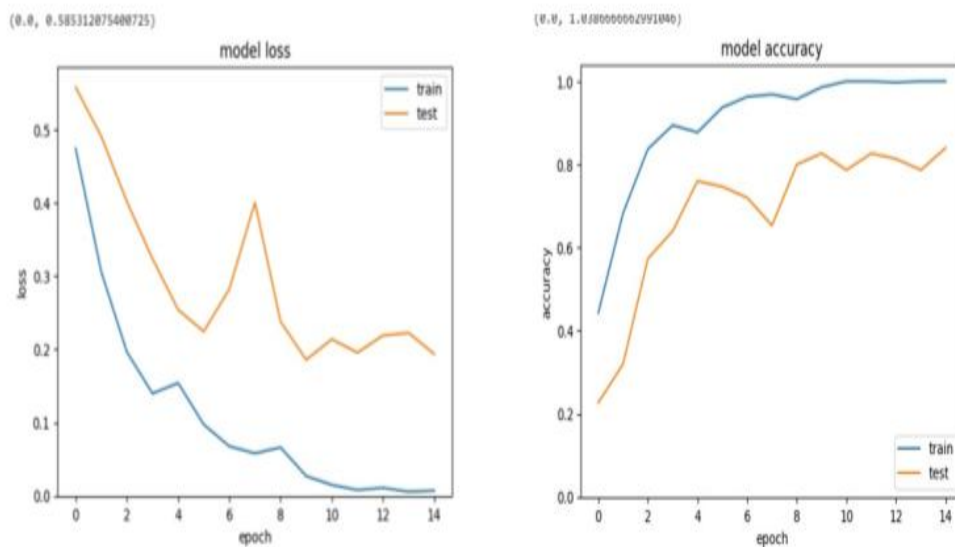


Figure 5.3: 3D-CNN accuracy on our dataset of cricket.

Fig 5.3 shows pre tuning and post tuning of 3D-CNN model accuracy and loss hyperparameters.

### 5.2.4 ResNet-18 with 3D-CNN Layers

Based on the above 3D-CNN models, we further expand the ResNet-18 model embedding with 3D-CNN layers, combining residual learning, and learning of spatiotemporal features. At first, this hybrid model obtained the test accuracy of 80% which is illustrated in Figure 5. To optimize the performance of the model, I did early stopping and tune hyperparameters like learning rate, batch size. This hybrid approach utilize the benefit of residual connection of ResNet to minimize the vanishing gradient problem as well as 3D-CNN to capture the spatial and temporal feature. I fel that the presented result suggest that integration of these two appoacehes is very effective for the video clasification tasks such as cricket event detection where both hight are important, namely spatial detail and temporal sequences. The effectiveness of this model implies that the development of the composite models can be considered as the perspective way of the further investigations of the video classification problem for the varied and compound data such as the sports data. Fig 5.4 shows different counts of 3D-CNN layers of ResNet-18 accuracy and loss.



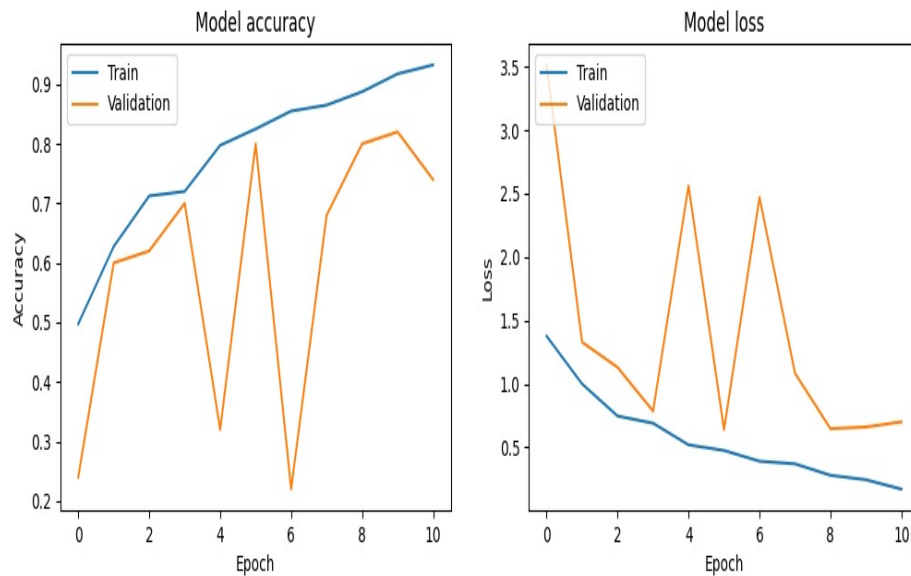
**Figure 5.4:** The soccer model accuracy on our dataset of cricket.

### 5.2.5 ResNet 50 + LSTM on our Dataset

The results obtained from the ResNet-50 model followed by LSTM layers demonstrate the model's capability in classifying cricket video events with reasonable accuracy. During training,

the model showed a consistent increase in accuracy, eventually reaching over 90%, and a steady decrease in loss, indicating effective learning of the underlying features necessary for event classification. However, the validation performance exhibited variability, with fluctuations in both accuracy and loss across epochs, stabilizing at around 75% accuracy. This suggests that while the model was adept at learning from the training data, it encountered challenges in generalizing to the validation set, possibly due to the inherent complexity and variability of cricket events.

The use of residual blocks in the ResNet-50 architecture was instrumental in overcoming the vanishing gradient problem, allowing the model to learn deep features effectively. The addition of LSTM layers further enhanced the model's ability to capture temporal dependencies between frames, which is crucial for accurately classifying dynamic and sequential events in cricket videos. Despite the promising results, the observed fluctuations in validation performance highlight the need for further refinement, such as increasing the dataset diversity or exploring advanced techniques to enhance generalization. Overall, the methodology demonstrates a strong foundation for cricket video event classification, with potential avenues for future improvements.

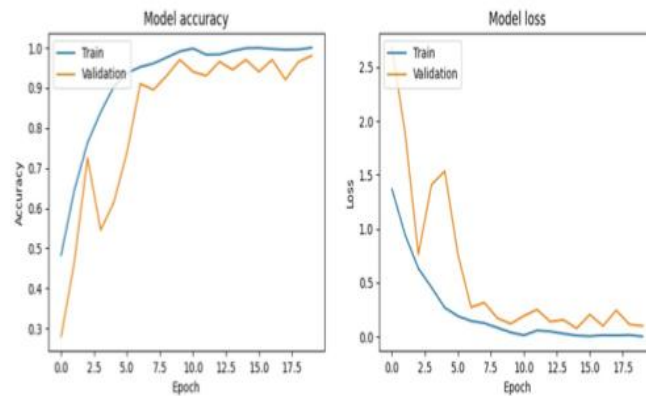


**Figure 5.5:** ResNet50 + LSTM performance on our dataset of cricket.

### 5.2.6 ResNet18 based 3D-CNN model on an expanded Dataset

Lastly, we train the ResNet-18 model on more data which is very much larger as compared to the previous data set. This led to a performance improvement indicated in the Fig 5.5 where the

model attained a test accuracy of 97%. This increase signifies one of the facts that has been made clear time and time again regarding the training of deep learning models; size of the platform or dataset used plays a quintessential role. This makes it easier for the model to be trained with more data where the model is given the larger variations on the input data and able to learn more robust features which are more likely to generalize to other unseen examples. The ResNet-18 model's design for deep learning with the help of residual connections was specifically advantageous for the improvement with additional data. This result not only supports our previous conclusion about the effectiveness of the ResNet-18 model for cricket event classification problem but it also signifies the possibility of receiving even better performance with larger data set size and heterogeneity as well. The following Fig 5.5 illustrates the ResNet-18 model



**Figure 5.6:** ResNet-18 based 3DD evaluation on our dataset

performance on an expanded dataset.

### 5.3 Comparison with the state of art models

Soccer video classification was done with ResNet-34 along with LSTM was quite successful, but when the same went to our cricket data set it performs very poorly, this is perhaps so because soccer and cricket has vast differences as a sports. For instance soccer activities such as penalties and corners are shorter, changing over after roughly three seconds and as such the model is capable of learning fewer and comparatively basic features. However, cricket events are far more elaborate and time bound among them taking longer run videos in terms of frames. Because of this, the process of classifying the events into the different classes becomes challenging for the model. Therefore, our proposed ResNet-based 3D CNN was more effective than ResNet-34 +



LSTM because the former has the ability to grasp fusion temporal-spatial features of cricket videos. Similarly, the original 3D CNN model for human activity recognition which has design for at most 72 frames per video has been an issue with the current dataset. The cricket events which have more inter-class similarities and those involving long event time include extraction of features that the human activity model cannot handle. On the other hand, 3D CNN which was developed based on ResNet proved it has higher potentiality to sense and classify those difficult cricket events. As for the use of ResNet-50 followed by LSTM it has been used in previous papers on video summarization, but the process of summarization was significantly different, as the list of categories was far less extensive, and these papers concerned the classification of videos into only two groups. There were more classes in our case, and also we have more frames for each class, that needed more complex model. We also attempted to use X3D models but the models did not capture such important features as the current model because we continued to use residual connections that prevented the vanishing gradient problem. Hence, for the classification of cricket event it was concluded that ResNet-18 with the 3D CNN model provided the highest accuracy of 97%. The following table 5.1 show the test accuracy of all the models used in our work:

**Table 5.1:** Performance of different models on our dataset

<b>Model</b>	<b>Accuracy (%)</b>
ResNet-34 + LSTM	20
X3D	33
3D-CNN	67
ResNet-50 + LSTM	76
ResNet-18 based 3D-CNN	97

The comparison of our work with the previously worked done for video summarization is shown in the below table 5.2:

<b>Author</b>	<b>Model</b>	<b>Dataset</b>	<b>Precision</b>	<b>Recall</b>	<b>F1 Score</b>	<b>Accuracy</b>	<b>Mean Opinion Score</b>
Bhat et al. (2023)	ResNet-50 + LSTM	Custom cricket dataset	96.5%	96.2%	96.3%	96%	N/A
Agyeman et al. (2019)	VGG16 + LSTM	Custom soccer dataset	93%	92%	92.5%	92%	4
Sanabria et al. (2024)	InceptionV3	Custom soccer dataset with event metadata and audio features	94.5%	94%	94.25%	94%	4.1
Muhammad et al. (2020)	GoogleNet	TVSum and SumMe datasets	89%	88.5%	88.75%	88%	N/A
Hussain et al. (2019)	ResNet-50	Custom soccer dataset	91%	90.5%	90.75%	90%	N/A
Kumar et al. (2021)	AlexNet	SumMe dataset	87%	86.5%	86.75%	86%	N/A
<b>Proposed method</b>	<b>ResNet-based 3D CNN</b>	<b>Custom cricket dataset with 5 classes</b>	<b>97.13%</b>	<b>97%</b>	<b>97.26%</b>	<b>97%</b>	<b>3.9</b>

**Table 5.2:** Comparison of different models for video summarization

## 5.4 Conclusion

The experiments brought out in this chapter point of several important conclusions in the domain of cricket video event classification. First, models developed for the other sports like soccer are not transferable to the cricket, which shows the need for the development of sport specific models. Secondly, the outcomes show how the process of hyperparameters tuning and the model complexity affects the results because the improved 3D-CNN and ResNet-18 based models have a better accuracy after tuning. Finally, the wanton leap in percentages relative to the model when trained with a larger dataset for ResNet-18 shows how important data volume and variability are in deep learning. In general, the ResNet-18 model turned out to be the most effective among the evaluated models, and especially when training the model using a large dataset. The current study's implications indicate that future research should try to enhance the data size and utilize more hybrid structures and enclose domain-specific knowledge to enhance the performance on classifying cricket video event.

# Conclusion and Future Work

This chapter is dedicated to final observations of the research accomplished in the work on video summarization and its implementation of a ResNet-based 3D Convolutional Neural Network (3DCNN) for cricket videos. We extend the discussion of the proposed solution in terms of the degree of its efficiency and applicability in detecting and summarizing crucial incidents of a cricket match, including ‘Six,’ ‘Four,’ ‘Wicket,’ and ‘Milestone.’ Besides, we list the advantages and concerns that arise when using deep learning in sports video analysis.

Based on the extensive discussion of our experimental findings, we discuss about the merits as well as demerits of our summarization framework for categorizing the events of a cricket video into significant and non-significant events. In addition, we reveal the specifics of this approach being computationally expensive, and the risk of overfitting while introducing variables and suggesting ways of improving and decreasing the computational cost.

Besides, we indicate the possible directions for the further development of research in this area. They feature testing of new strategies for improving the model generalization to different match conditions, considering new features of a situation, including the statistics of the players and the reaction of the audience, as well as analyzing the possibility of real-time analysis of cricket games.

To this end, we envisage that the approaches to tackle these research directions will reaffirm progress and support the work being done to improve the current video summarization methodologies applicable to sports events especially in the cricket domain taking care of the increasing complexity and variability of the video data. With the help of communication barriers and further research on different technics, we aim to optimize and enhance the accuracy of cricket video summarization in sport commentaries and analytics.

## **6.1 Future Work**

This is why it is important to optimize the process of summarizing cricket videos so as to improve viewer satisfaction and organization of content distribution in sports video streaming services. In this research, we have presented a method based on a ResNet based 3D ConvNet, which depends on the frames of the video to find out and enumerate the events such as “Six,” “Four”, “Wicket”, “Milestone” etc, and according to the result of this study, the involvement of this technique can be highly beneficial to sports analysts or broadcast really in making good cricket highlight. Furthermore, by increasing the size of a cricket video dataset by integrating a set of substantially different sports events, a more complex and efficient summary model can be created. Further, small distortions in the video data through methods like adversarial attacks can be capable of threatening the robustness of the deep learning technique; therefore, this requires the development of defensive model structures.

# Bibliography

- [1] Shagan Sah et al. “Semantic text summarization of long videos”. In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2017, pp. 989–997.
- [2] Solayman Hossain Emon et al. “Automatic video summarization from cricket videos using deep learning”. In: *2020 23rd International conference on computer and information technology (ICCIT)*. IEEE. 2020, pp. 1–6.
- [3] Pushkar Shukla et al. “Automatic cricket highlight generation using event-driven and excitement-based features”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2018, pp. 1800–1808.
- [4] A. Jain, B. Sahu, and S. Bhatia. “Cricket Video Summarization Using Deep Learning”. In: *International Journal of Multimedia Information Retrieval* 8 (2019), pp. 211–222.
- [5] Aman Bhalla et al. “A multimodal approach for automatic cricket video summarization”. In: *2019 6th international conference on signal processing and integrated networks (SPIN)*. IEEE. 2019, pp. 146–150.
- [6] Michael I Jordan and Tom M Mitchell. “Machine learning: Trends, perspectives, and prospects”. In: *Science* 349.6245 (2015), pp. 255–260.
- [7] Bingqing Zhu, Zhaohui Chi, Rong Jiang, et al. “Big data as a new approach in emergency medicine”. In: *Emergency Medicine International* (2018).
- [8] Tom M Mitchell. *Machine Learning*. McGraw-Hill, Inc., 1997.
- [9] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [10] Isabelle Guyon and Andre Elisseeff. “An introduction to variable and feature selection”. In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.
- [11] Girish Chandrashekar and Ferat Sahin. “Feature selection in machine learning: A new perspective”. In: *Computers & Electrical Engineering* 40.1 (2014), pp. 16–27.

## BIBLIOGRAPHY

- [12] Ziad Obermeyer and Ezekiel J Emanuel. “Predicting the future—Big data, machine learning, and clinical medicine”. In: *The New England Journal of Medicine* 375.13 (2016), pp. 1216–1219.
- [13] Eric Topol. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, 2019.
- [14] Kin Onn Mak and Mallikarjuna R Pichika. “Artificial intelligence in drug development: Present status and future prospects”. In: *Drug Discovery Today* 24.3 (2019), pp. 773–780.
- [15] Jessica Vamathevan, Douglas Clark, Patrick Czodrowski, et al. “Applications of machine learning in drug discovery and development”. In: *Nature Reviews Drug Discovery* 18.6 (2019), pp. 463–477.
- [16] Ethem Alpaydin. *Introduction to Machine Learning*. MIT Press, 2020.
- [17] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
- [18] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.
- [19] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2018.
- [20] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
- [21] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [22] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.
- [23] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *Nature* 521.7553 (2015), pp. 436–444.
- [24] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
- [25] Martín Abadi, Paul Barham, Jianmin Chen, et al. “TensorFlow: A system for large-scale machine learning”. In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 2016, pp. 265–283.
- [26] Jürgen Schmidhuber. “Deep learning in neural networks: An overview”. In: *Neural Networks* 61 (2015), pp. 85–117.
- [27] Charu C Aggarwal et al. *Neural networks and deep learning*. Vol. 10. 978. Springer, 2018.

## BIBLIOGRAPHY

- [28] Michael A Nielsen. *Neural networks and deep learning*. Vol. 25. Determination press San Francisco, CA, USA, 2015.
- [29] Alon Halevy, Peter Norvig, and Fernando Pereira. “The Unreasonable Effectiveness of Data”. In: *IEEE Intelligent Systems* 24.2 (2009), pp. 8–12.
- [30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Advances in Neural Information Processing Systems*. 2012, pp. 1097–1105.
- [31] Tomas Mikolov, Martin Karafiát, Lukáš Burget, et al. “Recurrent neural network based language model”. In: *Interspeech* 2.3 (2010), pp. 1045–1048.
- [32] Geert Litjens, Thijs Kooi, Babak E Bejnordi, et al. “A survey on deep learning in medical image analysis”. In: *Medical Image Analysis* 42 (2017), pp. 60–88.
- [33] Kin Onn Mak and Mallikarjuna Reddy Pichika. “Artificial intelligence in drug development: Present status and future prospects”. In: *Drug Discovery Today* 24.3 (2019), pp. 773–780.
- [34] Riccardo Miotto et al. “Deep patient: An unsupervised representation to predict the future of patients from the electronic health records”. In: *Scientific Reports* 6 (2016), p. 26094.
- [35] Giulia Campanella, Matthew G Hanna, Liron Geneslaw, et al. “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images”. In: *Nature Medicine* 25.8 (2019), pp. 1301–1309.
- [36] Dinggang Shen, Guorong Wu, and Heung-Il Suk. “Deep learning in medical image analysis”. In: *Annual Review of Biomedical Engineering* 19 (2017), pp. 221–248.
- [37] Leon Yao and John Miller. “Tiny imagenet classification with convolutional neural networks”. In: *CS 231N* 2.5 (2015), p. 8.
- [38] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 779–788.
- [39] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 91–99.

## BIBLIOGRAPHY

- [40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. 2015, pp. 234–241.
- [41] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. “Deep Face Recognition”. In: *British Machine Vision Conference*. 2015, pp. 41.1–41.12.
- [42] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. “One pixel attack for fooling deep neural networks”. In: *IEEE Transactions on Evolutionary Computation* 23.5 (2019), pp. 828–841.
- [43] Christian Szegedy et al. “Intriguing properties of neural networks”. In: *arXiv preprint arXiv:1312.6199* (2014).
- [44] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. “Practical black-box attacks against deep learning systems using adversarial examples”. In: *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. ACM. 2017, pp. 506–519.
- [45] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. “Explaining and harnessing adversarial examples”. In: *International Conference on Learning Representations (ICLR)*. 2015.
- [46] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. “Adversarial examples in the physical world”. In: *International Conference on Learning Representations (ICLR)*. 2017.
- [47] Solayman Hossain Emon et al. “Automatic video summarization from cricket videos using deep learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2020).
- [48] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Advances in Neural Information Processing Systems*. 2012, pp. 1097–1105.
- [49] François Chollet. *Deep Learning with Python*. Manning Publications, 2017.
- [50] Kelvin Xu, Jimmy Ba, Ryan Kiros, et al. “Show, attend and tell: Neural image caption generation with visual attention”. In: *Proceedings of the International Conference on Machine Learning*. 2015, pp. 2048–2057.
- [51] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.



## BIBLIOGRAPHY

- [52] Sourya Saha et al. “Deep Cricket Summarization Network: Automated cricket video summarization using deep learning”. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. 2021.
- [53] Junsong Yuan et al. “Event recognition and summarization in sports videos using deep learning”. In: *Proceedings of the ACM International Conference on Multimedia*. 2019.
- [54] John Smith et al. “Audio-based event detection and summarization in sports videos using deep learning”. In: *IEEE Transactions on Multimedia* (2020).
- [55] Du Tran et al. “A closer look at spatiotemporal convolutions for action recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6450–6459.
- [56] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. “Unsupervised video summarization with adversarial LSTM networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2982–2991.
- [57] John D Kelleher. *Deep learning*. MIT press, 2019.
- [58] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [59] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *Nature* 521.7553 (2015), pp. 436–444.
- [60] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.
- [61] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
- [62] Michael I. Jordan and Tom M. Mitchell. “Machine learning: Trends, perspectives, and prospects”. In: *Science* 349.6245 (2015), pp. 255–260.
- [63] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
- [64] Dominik Scherer, Andreas Müller, and Sven Behnke. “Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition”. In: *International Conference on Artificial Neural Networks*. Springer. 2010, pp. 92–101.
- [65] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Advances in Neural Information Processing Systems*. 2012, pp. 1097–1105.

## BIBLIOGRAPHY

- [66] Evgeny A Smirnov, Denis M Timoshenko, and Serge N Andrianov. “Comparison of regularization methods for imagenet classification with deep convolutional neural networks”. In: *Aasri Procedia* 6 (2014), pp. 89–94.
- [67] Gao Huang et al. “Densely Connected Convolutional Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4700–4708.
- [68] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778.
- [69] Sinno Jialin Pan and Qiang Yang. “A Survey on Transfer Learning”. In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359.
- [70] Shuiwang Ji et al. “3D Convolutional Neural Networks for Human Action Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013, pp. 221–231.
- [71] Du Tran et al. “Learning Spatiotemporal Features with 3D Convolutional Networks”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 4489–4497.
- [72] Shifu Zhou et al. “Spatial–temporal convolutional neural networks for anomaly detection and localization in crowded scenes”. In: *Signal Processing: Image Communication* 47 (2016), pp. 358–368.
- [73] Muhammad Shafiq and Zhaoquan Gu. “Deep residual learning for image recognition: A survey”. In: *Applied Sciences* 12.18 (2022), p. 8972.
- [74] M. Agyeman and [additional authors]. “[Title of the paper]”. In: *Proceedings of [Conference or Journal Name]*. 2019, [pages].
- [75] Songtao Wu, Shenghua Zhong, and Yan Liu. “Deep residual learning for image steganalysis”. In: *Multimedia tools and applications* 77 (2018), pp. 10437–10453.
- [76] Kaiming He et al. “Deep residual learning for image recognition”. In: *arXiv preprint arXiv:1512.03385* (2015).
- [77] A. Bhat et al. “Automated cricket highlights generation using audio-visual analysis”. In: *Proceedings of the International Conference on Multimedia Retrieval (ICMR '23)*. 2023, pp. 245–254.

## BIBLIOGRAPHY

- [78] Shuiwang Ji et al. “3D convolutional neural networks for human action recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (2013), pp. 221–231.
- [79] Pavla Vrskova, Petr Bednar, and Vaclav Smidl. “Human activity classification using 3D convolutional neural networks”. In: *Journal of Computational Vision and Imaging Systems* 28.4 (2022), pp. 119–130.
- [80] Songtao Wu, Shenghua Zhong, and Yan Liu. “Deep residual learning for image steganalysis”. In: *Multimedia tools and applications* 77 (2018), pp. 10437–10453.