CrickAlike

Capstone Project Report

End-Semester Evaluation

Submitted By:

(101903397) Sameer Alam

BE Fourth Year - COE

CPG No. 281

Under the Mentorship of

Dr. Simranjit Kaur

Assistant Professor

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

ABSTRACT

Over 93% of viewership in Cricket comes from India which explains how passionate people are about cricket in India . Wherever feasible, they spend hours in front of the TV enjoying cricket matches and occasionally playing with bat and ball. Gully cricket is played on the streets by young people and adults inspired by the record breaking performances by players in matches to show off their enthusiasm for this game. But with the increase in subscription cost for streaming platforms it is getting difficult for people to watch live cricket matches. With CrickAlike, an attempt has been made to recreate live matches in the form of animated content that closely imitates live events for free of cost to the users.

DECLARATION

I hereby declare that the design principles and working prototype model of the project entitled **"CrickAlike"** is an authentic record of my own work carried out in the Computer Science and Engineering Department, TIET, Patiala, under the guidance of **Dr. Simranjit Kaur** during 6th and 7th semester (2022).

Date: 26-12-2022

Roll No.	**Name**	**Signatures**
101903397	Sameer Alam	

Counter Signed By:

Faculty Mentor: **Dr. Simranjit Kaur**

Dr. _____

Assistant Professor,

CSED,

TIET, Patiala

ACKNOWLEDGEMENT

We would like to express our thanks to our mentor **Dr. Simranjit Kaur**. She has been of great help in our venture, and an indispensable resource of technical knowledge. She is truly an amazing mentor to have.

We are also thankful to **Dr. Shalini Batra**, Head, Computer Science and Engineering Department, entire faculty and staff of Computer Science and Engineering Department, and also our friends who devoted their valuable time and helped us in all possible ways towards successful completion of this project. We thank all those who have contributed either directly or indirectly towards this project.

Lastly, we would also like to thank our families for their unyielding love and encouragement. They always wanted the best for us and we admire their determination and sacrifice.

Date:

Roll No.	**Name**	**Signatures**
101903397	Sameer Alam	

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

S No.	Definition	Abbreviation
1.	Machine Learning	**Ml**
2.	User Interface	**UI**
3.	Natural Language Processing	**NLP**
4.	Logistic Regression	**LI**
5.	Support Vector Machine	**SVM**
6.	Software Requirement Specification	**SRS**
7.	Database Management System	**DBMS**
8.	Figure	**Fig**

1. Introduction

1.1 Project Overview

Cricket is the most popular sport on the planet after football, with a billion-dollar economy. There is much interest in emulating cricket and football, especially in displaying real-time scores from cricket matches, one-day international and T20 cricket matches, and different formats of a cricket match.

1.1.1 Technical Terminology

- **Machine Learning:** Machine learning is a subfield of artificial intelligence, broadly defined as a machine's capability to imitate intelligent human behavior. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems.
- **Natural Language Processing:** Natural Language Processing (NLP) is a subfield of artificial intelligence and linguistics that focuses on the interaction between computers and human (natural) languages. It involves developing algorithms and models that can understand, interpret, and generate human language.
- **Software Engineering:** Software engineering is the process of designing, building, testing, and maintaining software applications and systems. It involves using a systematic and disciplined approach to the development of software, with the goal of creating high-quality, reliable, and maintainable software that meets the needs of its users.
- **Web Development:** Web development is the process of building and maintaining websites. It involves a variety of tasks and technologies, including design, content creation, programming, and security.

1.1.2 Problem Statement

It is a fact that not everyone can watch cricket matches because of the platform subscription fees, and that many fans are subjected to using scorecard apps such as

Cricbuzz as their only means of observing the game, which deprives them of the opportunity to actually watch the game in person.

In order to meet the desired experience for users, we would like to provide a product that falls between the tedious comments and the live game in order to allow everyone to share the same intensity and passion throughout the entire game.

1.1.3 Goal/Motivation

With 93 percent of all sports viewers in 2018 turning in to cricket content [5], cricket is the country's most popular sport. Despite this, this number is fast declining as a result of recent increases in subscription prices.

Instead of text-based information, which is now the most preferred format for information following a live match, there is a need to upgrade to more interactive video that showcases everything as a live cricket match.

Our objective is to create a platform that can grant access to people who do not have the right or means to view the subscription material.

We aim to produce animated video that closely resembles live events to mimic sporting events like matches.

We aim to produce animated video that closely resembles live events to mimic sporting events like matches. That would help numerous people from jeopardizing their privacy by stopping them from using third-party software to view live cricket matches.

1.1.4 Solution/Contribution

Cricket is a direct sport in both its natural and good aspects. Two players, a bat, and a ball are all required. Twenty-two players play professional cricket, yet even two people can enjoy a fast-paced cricket match off the field. Furthermore, cricket does not require a large amount of space to play. One might see people playing cricket sideways on India's busy roadways on the highway. Indians, according to Davis, the world's foremost mathematical educators, do not just watch and enjoy cricket; they actively participate in it [1]. Numerous websites are now available that display live cricket scores [2]. We have proposed a more robust experience by providing free live cricket

streaming without copyright issues. This would be accomplished by decoding a live streaming match into an understandable format using video processing and then creating a video containing all relevant information in real-time, such as which ball is being bowled and which shot is being played by the player, as shown in Fig 1.

Fig 1 - showing transition of information from live match to a new live video.

1.2 Need Analysis

There are already multiple websites which provide live scores, summary during cricket matches which are very convenient when we just want to get an update about the match. But these websites fail to provide users with the same experience of the enthusiasm which they feel during the cricket match. There is a need to upgrade from these scoreboard style information to more interactive video showcasing everything as a live cricket match rather than text based information which is a most popular format as of today after a live match.

In light of these considerations, an attempt is made to design something that solves these issues and can be used by a vast number of individuals.

1.3 Research Gap

After taking many research papers and journals into consideration we found that most of the methods used for depicting the cricket techniques take many assumptions. These methods do not try to cover all aspects but they just took the text for their study. We could not find enough research papers which focused on this idea. There is also minimum work done to include the time or order of the text as a parameter.

Furthermore, there is a gap in research regarding how to optimize models using unstructured and unlabelled databases. Most of the databases that we have collected are unlabelled and contain much unnecessary content. A well-established and maintained database could not be found during our initial research for our data source.

So we had to explore various social media data sources to find databases , structure it and give a label on our own accord and then combine to make a dataset for our model.

1.4 Problem Definition and Scope

Over 61 percent of Indians have access to the Internet as of February 1, 2022, thanks to telecom providers' consistent efforts in previous years [3]. Just as education is a basic human right, every Indian should have access to their favorite sports without need of paying subscription fees.

Not everyone has access to cricket matches due to platform subscription fees, and many fans are forced to rely on scorecard apps like Cricbuzz to follow the game, depriving them of the opportunity to see a real match.

Our Product would attempt to give users an experience that falls halfway between the tedious comments and the live game, so that everyone can share the same level of excitement and passion throughout the match. This will mostly affect people who cannot afford monthly subscriptions and must make do with reading commentary, which takes away from the heart of the live game, or going to third-party websites, which puts their privacy at risk [4].

1.5 Assumptions and Constraints

1.5.1 Assumptions:

S.NO	ASSUMPTIONS
1.	We are assuming that the user has access to electronic gadgets like smartphones, laptops etc.
2.	We are assuming that users can understand cricket terminology.
3.	We are assuming that users have basic understanding of using web applications

Table 1. Assumptions

1.5.2 Constraints:

S.NO	CONSTRAINTS
1.	It should take less than one minute to display live animations.
2.	Internet speed must be greater than 500 kbps.

Table 2. Constraints

1.6 Standards

1. **Web 2.0.**
2. **IEEE 830-1998:** The content and qualities of a good software requirements specification (SRS) are described and several sample SRS outlines are presented. This recommended practice is aimed at specifying requirements of software to be developed but also can be applied to assist in the selection of in-house and commercial software products.
3. **IEEE/ISO/IEC P23026:** This document defines system engineering and management requirements for the life cycle of websites. This document applies to those using web technology to present information and communications technology (ICT) information, such as information for users of systems and services, plans and reports for systems and software engineering projects, and documentation of policies, plans, and procedures for IT service management.
4. **IEEE SA - P3123:** The standard defines specific terminology utilized in artificial intelligence and machine learning (AI/ML). The standard provides clear definitions for relevant terms in AI/ML. Furthermore, the standard defines requirements for data formats. This will help to ensure that data is accurately interpreted by AI/ML systems, resulting in more accurate predictions and insights

1.7 Approved Objectives

1. To extract semantic structures from a video, allowing for summary, browsing, and indexing of the video material using video mining.

2. To categorize a series of frames into one of the following groups: GROUND, PITCH, BOUNDARY, PLAYER, and CROWD using shot categorization.
3. To analyze four occurrences for categorization: FOUR, SIX, OUT, and RUN using an appropriate identification algorithm using event classification from video and text extraction from audio features.
4. To recreate a given event by choosing the most appropriate and similar video from our video library using recreating events.
5. To develop a live dashboard featuring live scores of the match and providing text summarization of ball-to-ball events.

1.8 Methodology

The methodology of proposed work is depicted in Fig 2.

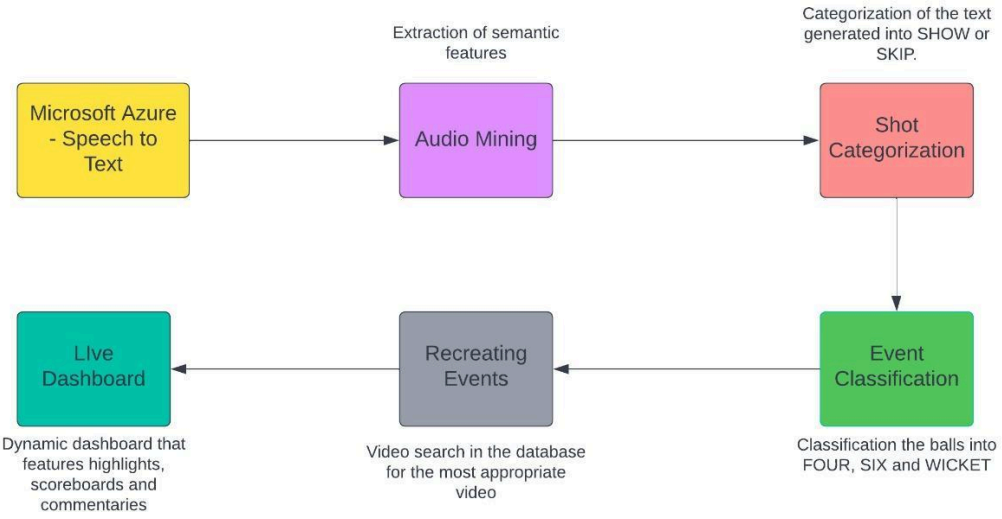

Fig 2 - Showing the Workflow for the methodology

1.8.1 Video and Audio Mining

For mining a sports video, audio and visual characteristics must be retrieved from the frame sequence. After relevant characteristics are obtained, finding events of interest in the video reduces to a challenge of temporal or sequential classification.

1.8.2 Shot Categorization:

We would be employing audio features for the extraction of texts through the voice of the commentators using the “Speech Recognition” API and the “PyAudio” library.

The ball is classified into shot or non-shot boundaries by finding patterns with the text extracted from the voice of commentators. For example, during the boundary more active and lively words can be expected from the voice of commentators.

1.8.3 Event Classification:

The ball is further classified into one of the few major events during the game such as SIX, FOUR, OUT, and RUN taking into account the shot categorization results as well. This is done by mapping a few most used slangs or words related to that event. For example, the keyword “huge” can be attributed to six.
Support Vector Machines is employed for classification into the respective events based as it most closely relates to our requirements. The exact path of the ball can also be traced using the text which would help in choosing the most relevant video.

The accuracy of our model employing audio features discussed previously significantly increased the accuracy for event classification.

1.8.4 Recreating Events:

We have a database of videos corresponding to each of the four events in the match, and it selects and displays the one that is most relevant to the current event. A video search is done in the entire database using Machine Learning to find the most identical video.

1.8.5 Live Dashboard:

To design the dashboard, we will utilize React for Interactive Dashboard along with socket io to pass information in real time from live cricket matches to streaming video. We will also try to build the backend of our website and integrate it with the video database.

1.9 Project Outcomes and Deliverables

Our application displays live scores and ball-by-ball commentary and also provides live animation matches for registered users

We anticipate the following notable outcomes of our project.

- Users will have an alternative to the monotonous cricket commentaries.
- Users who cannot afford hefty subscription fees can now have access to live matches.
- Users do not have to put their privacy at risk by visiting the third-party websites for watching the cricket live.
- Streaming other applications and websites will require more internet bandwidth than our application which uses animations primarily.
- During the match, users may communicate with one another using the live chat feature.

1.10 Novelty of Work

The project is mostly inspired from already accomplished work but it tends to fulfill the research gaps given out by the existing work in the field of our interest. We made our applications that display live scorecards with commentary and also play live animation of matches . Increasing our novelty and concerns for actual requirements we also made our own scorecard and whole database.

2. Requirement Analysis

2.1 Literature Survey

2.1.1 Theory Associated With Problem Area

Cricket is the sport that the Indians enjoy watching the most. Cricket, on the other hand, has progressed beyond a game to become a business model. Individuals must pay high subscription fees in order to watch and enjoy cricket because of this.

As discussed above, users of many platforms, like Hotstar, SonyLIV, and Amazon Prime, are charging hefty monthly fees. Cricket is the crown jewel of the country's growing sports audience, with 93 percent of all sports viewers tuning in to cricket material in 2018 [5]. However, due to expensive membership rates, this number is rapidly dwindling. In this section, we'll examine these solutions in depth, as well as their shortcomings and what we can learn from them.

2.1.2 Existing Systems and Solutions

Ramalatha Marimuthu et al.(2022) develops a novel Taylor Gradient Descent Political Optimizer (Taylor GDPO) based deep learning model for speech recognition [6]. A developed Taylor GDPO is obtained by integrating Taylor series, Gradient Descent (GD) and Political Optimizer (PO). Firstly, pre-processing of an input signal is done and the features are extracted. Then, the extracted features are given as input to the Deep Residual Network (DRN), which is trained by the developed Taylor GDPO

Metin Turan et al.(2022) focused on the extraction of inferential written summaries of communications that occur in oral environments such as meetings [7], lectures and conferences. However, since this type of problem requires conversion from audio to text, it also includes issues such as the human factor, sound recording environments, and language-specific problems

In this paper Banoth Thulasya Naika et al.(2022) discussed computer vision techniques for various applications [8] .Detection and classification of players,tracking player or ball in sports, predicting the trajectories of player or ball,detection and

classification of each player based on their team in every frame or also by recognizing the jersey number and classify gestures, recognize the actions of umpire.

Vladyslav Tsap et al.(2021) explored the AdaBoost algorithm, one of the most popular ensemble boosting methods [9]. The main ensemble methods and their advantages and disadvantages are considered and focused on the AdaBoost algorithm. The dataset of the University of Belgium compiled is used. The AdaBoost algorithm was experimentally applied to the data set, and its effectiveness was tested.

In this paper R Rahman et al.(2021) proposed a novel strategy to identify the type of delivery from the finger grip of a bowler while the bowler makes a delivery [10]. To identify the grip of a bowler while bowling we need the help of different convolutional architectures and transfer learning models. About 5573 images from Real-Time videos were prepared and the group of these images are named as GRIP DATASET.

Pinal Shah et al. (2018) observed different techniques and algorithms that are applied to achieve functionalities like Speech to Text, Text to speech [11], Speech recognition, communication, Machine translation etc.

In order to categorize six types of cricket shots, Md. Ferdouse Ahmed et al.(2018) suggested a 13-layer CNN method [12]. The suggested model has a high accuracy while having a low cross-entropy rate..It also discusses published works in a variety of application-specific tasks related to sports and the present researchers views regarding them.

M.H. Kolekar et al.(2008) revolved around developing a novel temporal classifier [13].Which is based on sequential pattern mining and machines learning algorithms for identifying different events, but the work of this research paper was only limited to the two major events in cricket, wickets and hits.

In this paper K.P. Sankar et al.(2006) researched primarily on the sports video annotation and summarization of a variety of sports videos including baseball, tennis, and cricket [14]. Cricket videos have received less attention than other sports, owing to the game's higher complexity and duration.

I. Koprinska et al.(2001) has proposed a more sophisticated thresholding scheme for the detection of the gradual shot transitions since fades and wipes are frequent during replays in cricket videos [15]. This review identifies the research directions, probable challenges, and future trends in the area of visual recognition in sports.

J. Platt et al.(1998) built a training corpus that consisted of the tuples associated with shot sequences such as GROUND, PITCH, BOUNDARY, CROWD, PLAYER and REPLAY and corresponding events [16]. Support Vector Machine was trained on the training corpus using the sequential minimal optimization algorithm for the classification of the events .

2.1.3 Research Findings for Existing Literature

Author	Tools/Dataset Used	Methodology	Result/Accuracy
Ramalatha Marimuthu	Taylor-gradient Descent political optimization based Deep residual network	Taylor Gradient Descent Political Optimizer (Taylor GDPO) based deep learning model for speech recognition. A developed Taylor GDPO is obtained by integrating Taylor series, Gradient Descent (GD) and Political Optimizer (PO).	The proposed Taylor GDPO-based DRN offered effective performance with highest accuracy of 96.93%, smallest FAR of 2.438%, smallest FRR of 2.101%, smallest MSE of 0.038.
Metin Turan	Speech Recognition Python library, Google's cloud Application Programming Interface (API) etc.	The extraction of inferential written summaries of communications that occur in oral environments such as meetings, lectures and conferences.	The results obtained are remarkable and it is seen that approximately 71% success was achieved.
Banoth Thulasya Naika	Deep Learning and AI techniques used.	Detection and classification of players,tracking player or ball in sports, predicting the trajectories of player or ball,detection and classification of each player based on their team in every frame.	Got Maximum accuracy of about 0.614 by using the classifiers.

Vladyslav Tsap	The dataset of the University of Belgium compiled is used.	The AdaBoost algorithm, one of the most popular ensemble boosting methods. The main ensemble methods and their advantages and disadvantages are considered and focused on the AdaBoost algorithm.	This algorithm can improve the result compared to "strong" classifiers upto 30%. However, there are cases when even a difference of 3-5% is significant.
R Rahman	Collected their data from ICC GRIP DATASET.	Identify the type of delivery from the finger grip of a bowler while the bowler makes a delivery. For identify the grip of a bowler while bowling we need the help of different convolutional architectures and transfer learning models	The results obtained are remarkable and it is seen that approximately 76% success was achieved with Accuracy of about 0.767.
Pinal Shah	IBM Watson Developer Cloud, Google Translate	Techniques and algorithms that are applied to achieve functionalities like Speech to Text, Text to speech, Speech recognition, communication, Machine translation etc.	Got Maximum accuracy of about 0.835 by using the classifiers.
Md. Ferdouse Ahmed	Machine learning model training using CNN method.	In order to categorize six types of cricket shots, a 13-layer CNN method suggested	The suggested model has a high accuracy while having a low cross-entropy rat
M.H. Kolekar	SVM and sequential pattern mining algorithms are used.	Developing a novel temporal classifier based on sequential pattern mining and support vector machines for identifying different events	The work of this research paper was only limited to the two major events in cricket, wickets and hits
K.P. Sankar	Collected their dataset from volunteers phone and manually labeled data	The sports video annotation and summarization of a variety of sports videos including baseball, tennis, and cricket	Achieved an accuracy of 0.83, with precision 0.90 and recall of 0.8-

I. Koprinska	-	A more sophisticated thresholding scheme for the detection of the gradual shot transitions.	
J. Platt et al	Manually collected their dataset	a training corpus that consisted of the tuples associated with shot sequences such as GROUND, PITCH, BOUNDARY, CROWD, PLAYER and REPLAY and corresponding events	Support Vector Machine was trained on the training corpus with accuracy upto 75%

Table 3: Literature review

2.1.4 Problem Identified

In case of the sports like Cricket, the jerseys are massive and have so much branding, due to which jersey number recognition is quite hard and also Identification of the player is difficult when they're move out of the frame and to retain the identity when the players reappear in subsequent frames, the player must be recognized this is the most difficult recognition challenge.

Other major issues are similar appearance of subjects, complex occlusions, unconstrained field environment, background, unpredictable movements, unstable camera motion, issues with calibration of low textured fields and the editing done to broadcast video, lower pixel resolution of players who are distant and smaller in the frame, motion blur etc.

2.1.5 Survey of Tools and Technologies Used

The software tools and packages used by the system are:

- Library for Support Vector Machine (SVM)
- Library for Generalized Sequential Pattern (GSP)
- Multi-Layer Perceptron (MLP)
- OpenCV [Open Source Python library for Computer Vision]
- Anaconda Spyder
- Mongo Database (MongoDB)
- React JS (for Front-end)

2.1.5 Summary

Very less work has been done on the game of cricket on the account that it usually lasts much longer than other games like football and basketball. Most of the existing research work and literature primarily focus on the video aspects employing frames. Very less work has been done on the audio features of the cricket commentaries. The existing video solutions have a lot of drawbacks and are very inefficient.We aim to make a software based approach employing audio features and classifying the commentaries to display the major events.

2.2 Software Requirement Specification

2.2.1 Introduction

2.2.1.1 Purpose

Our project aims to build a platform that can provide access to users who do not have the privilege to access the subscription content. As a result, many users tend to rely on third-party applications to watch live cricket matches, endangering their privacy. We recreate live matches in the form of animated content that closely imitates live events.

2.2.1.2 Intended Audience and Reading Suggestion

- Anyone who loves cricket
- Moreover , who can’t afford to buy premium can watch live match in a processed form with live scoreboard
- Read each heading and its content written in it as per the index
- The document is intended to be read by project managers, developers, testers and users.
- The document is organized into 5 parts.
    1. Introduction,
    2. Overall Description,
    3. System Features,
    4. External Interface Requirements
    5. Quality Attribute Requirements.
- All the parts are independent however reading the document sequentially helps the reader understand the Web Order system better.

2.2.1.3 Project Scope

- Our objective is to provide the facility for free to those who are not able to afford to watch live cricket.
- We are working on a software-based approach for voice recognition.
- It will understand the audio(from input video database) and convert it from voice to text.
- And the scoreboard will get updated .
- The project will be displayed using a website.
- This will make it more approachable and feasible for the general public.

2.2.2 Overall Description

2.2.2.1 Product Perspective

Our Product can be mainly divided into 2 Components involving Match Summary and showcasing Live Match. If Users want to have a quick recap about a Match Undergoing then they can view Match Summary and if they want to experience a full Match then they can explore Live Match Component. Per Ball Match Summary would be made available in Match Summary Component and Users would have access to Live Scoreboard, Video created using Live Match Undergoing and also to make sessions more interactive users can chat with each other about events undergoing in a cricket match. Fig 3 Representing the product perspective diagram of CrickAlike

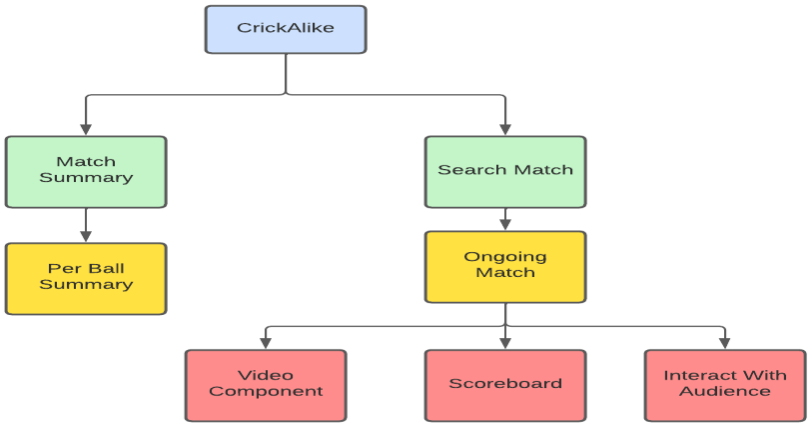

Fig 3 - Product Perspective Diagram of CrickAlike

2.2.2.2 Product Features

Major Features of the website are:-

1. Users can Register/Login into the website.
2. Users can view Live Cricket Match undergoing.
3. Users can talk with each other about the event happening live in cricket match.
4. Users can view Match Summary for quick brief about the Match.

2.2.3 External Interface Requirements

2.2.3.1 User Interface

- Browser Based for user

2.2.3.2 Hardware Interface

- Laptop or PC

2.2.3.3 Software Interface

- Web Server Based Interface.

2.2.4 Non-functional Requirements

2.2.4.1 Performance Requirements

- Since this software is going to be web – based, it does require a powerful server machine with high band internet access.
- Server machine should have a powerful CPU and high speed internet access so that it can handle multiple users at the same time. Another performance requirement is the storage space. Higher storage space means more users and bigger workspace per user so higher the storage, better the performance.
- Performance requirement by the user side is, web application should be developed as a lightweight web app so that it can work on almost any platform even with slower internet connections.
- Expected number of simultaneous users should be at least 100. System should be able to deal with 100 users at the same time. Also the database of the system

should handle at least a thousand of users at any period.

2.2.4.2 Safety Requirements

- The users should not share their passwords with others.
- Users should remember to logout after each session.
- The DB should be managed by administered personnel only.

2.2.4.3 Security Requirements

- During user registration, the given email address is validated.
- The password should be at least 8 characters, containing at least a small character and one capital, a number and a special character.
- Password is stored as a hash value in the DB.
- We are transferring all data via HTTPS i.e. via SSL so that the data is encrypted during the transit. Thus safeguarding the user information.
- In the near future we will try to integrate reCAPTCHA service which is used during registration for human identification to avoid attacks on the website as well as fake entries.

2.2.4.4 Software Quality Attributes

- Usability : User interface should be simple and easy to use for a new user.
- Availability : The system should be available at all times. It should be ensured that there should be minimum or no downtime to ensure better user experience. The system should be reliable. It should yield correct results if a user performs searches for a post/discussion or topic . Also, if the user writes an opinion or creates a topic, the system should ensure that the correct opinion is entered in the DB without data loss and data redundancy.
- Testability : The application should be testable. A separate test environment should be set up where testers and the Quality Assurance engineers can test the application for bugs and/or incomplete or missed requirements
- Maintainability : The system should be developed in such a way that it is

extensible. It should be easy to incorporate new features requirements or accommodate a change in the existing requirements.

2.3 Cost Analysis

We have purchased a Cricket 22 game for extracting animation. It will save us from copyright.The game cost us Rs. 2199 Other than we are using open source tools and products so they are free to use.

2.4 Risk Analysis

Data Theft: Chances are the personal data of any person is stolen and used for wrong purposes.

3. Methodology Adopted

3.1 Investigative Techniques

Investigative Projects Techniques	Investigative Techniques Description
Descriptive	Tracing the direction of the ball using the visual features was not possible. Shortcomings from the video classification only were dealt with to classify events based on the voice of commentators as well by extracting text from the voice.
Comparative	For simulating sports of different types like football, and tennis, very little work has been done in regard to cricket as the duration of the game is enormous in cricket. Also another restriction is that in cricket, there are frequent replays and videos from the crowd, which makes the task even more challenging to build a proper algorithm.

Table 4: Investigative techniques

3.2 Proposed Solution

The idea is to provide users who cannot afford hefty subscription fees a platform to watch live matches in the form of animated content that closely resembles the live match. The solution proposed works by extracting meaningful features from the audio and video and recreating a few major events during the match. The working of the model can be explained in the following steps:

- Audio characteristics are retrieved from the video samples from the voice of commentators.
- Shot categorization is employed to classify the ball as a shot and a non-shot boundary.

- Text extraction is carried out further from the voice of commentators to classify it further into one of the four major cricketing events that is, SIX, FOUR, OUT, and RUN.
- The video that most closely resembles the current event, is taken out from the database fetched and displayed, and scores are updated correspondingly.

3.3 Work Breakdown Structure

The Work Breakdown Structure of CrickAlike is Explained below in Figure 4.

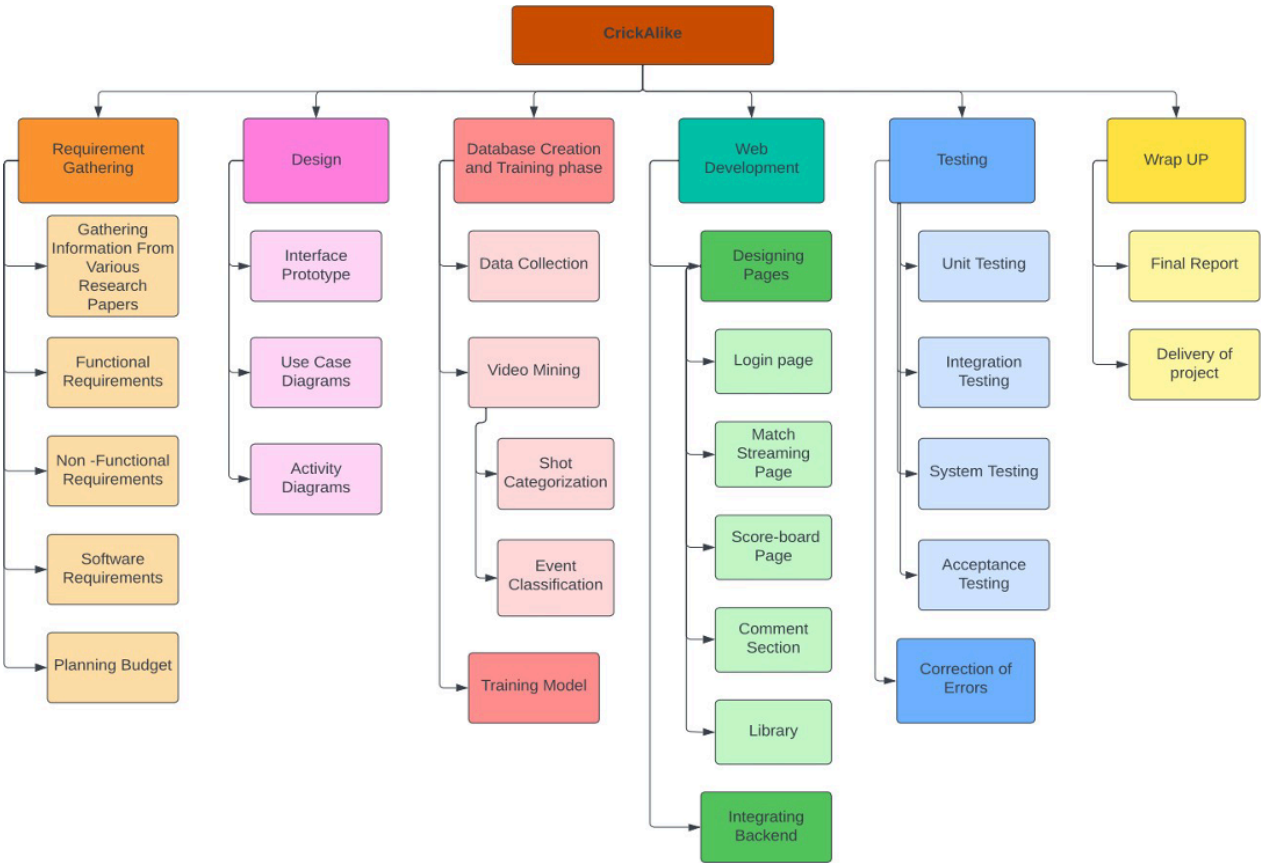

Fig 4 - Work Breakdown Structure.

3.3.1 Work Plan

No.	Activity	Month	March				April				May				June				July			
		Week	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1.	Team Formulation And Project Planning	Plan																				
		Actual																				
2.	Require ment Analysis	Plan																				
		Actual																				
3.	Reading Research Papers	Plan																				
		Actual																				
4.	Project Proposal And Finalizing Analysis Model	Plan																				
		Actual																				
5.	Data Collection	Plan																				
		Actual																				
6.	Web Develop ment	Plan																				
		Actual																				
5.	Audio Mining	Plan																				
		Actual																				

Table 5 - Gantt Chart depicting work plan from March to July

No.	Activity	Month	August				September				October				November				December			
		Week	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
6.	Shot Categorization	Plan																				
		Actual																				
7.	Event Classification	Plan																				
		Actual																				
8.	Training Model	Plan																				
		Actual																				
9.	Performing Modifications	Plan																				
		Actual																				
10.	Results Evaluation	Plan																				
		Actual																				
11.	Final Report	Plan																				
		Actual																				

Table 6 - Gantt Chart depicting work plan from August to December

3.4 Tools and Technology

The software tools and packages used by the system are:

- Scikit Library for **XGBClassifier** (XGB)
- Scikit Library for AdaBoost algorithm (AdaBoost)
- OpenCV [Open Source Python library for Computer Vision]
- Anaconda Spyder
- Mongo Database (MongoDB)
- React JS (for Front-end)

4. Design Specifications

4.1 System Architecture

4.1.1 Block Diagram

Figure 5 shows the Block Diagram.

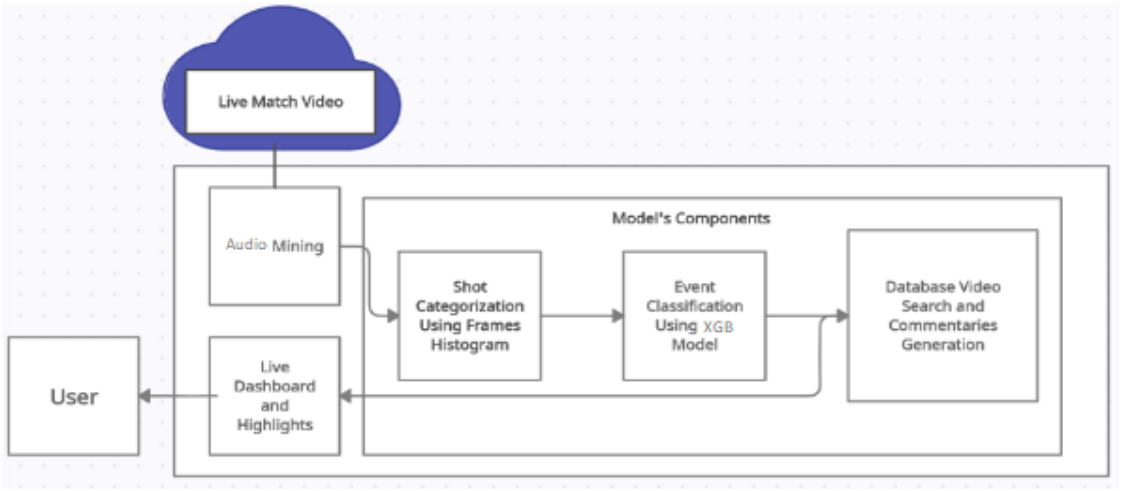

Figure 5. Block Diagram

4.1.1 MVC Diagram

Figure 6 shows the MVC Diagram.

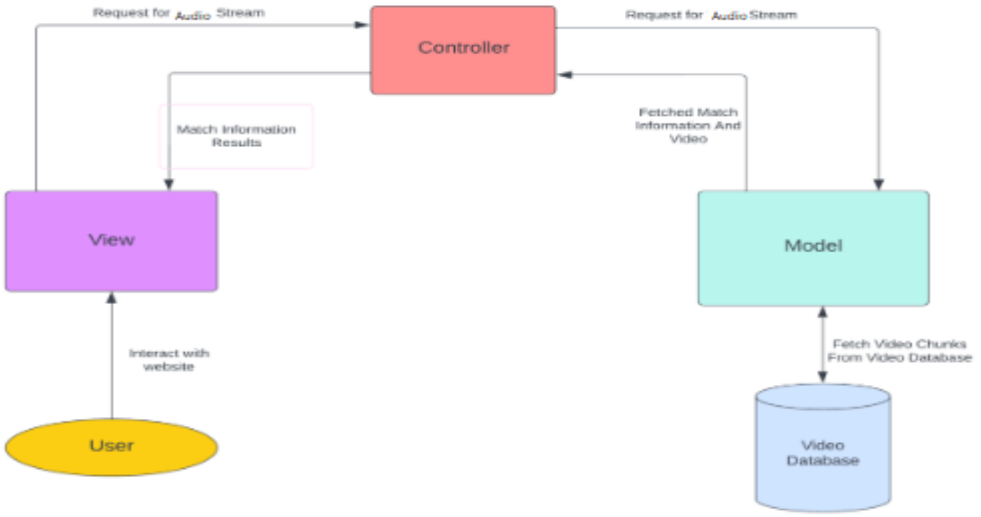

Figure 6. MVC Diagram.

4.1.3 Tier Diagram

Figure 7 showing Tier Diagram

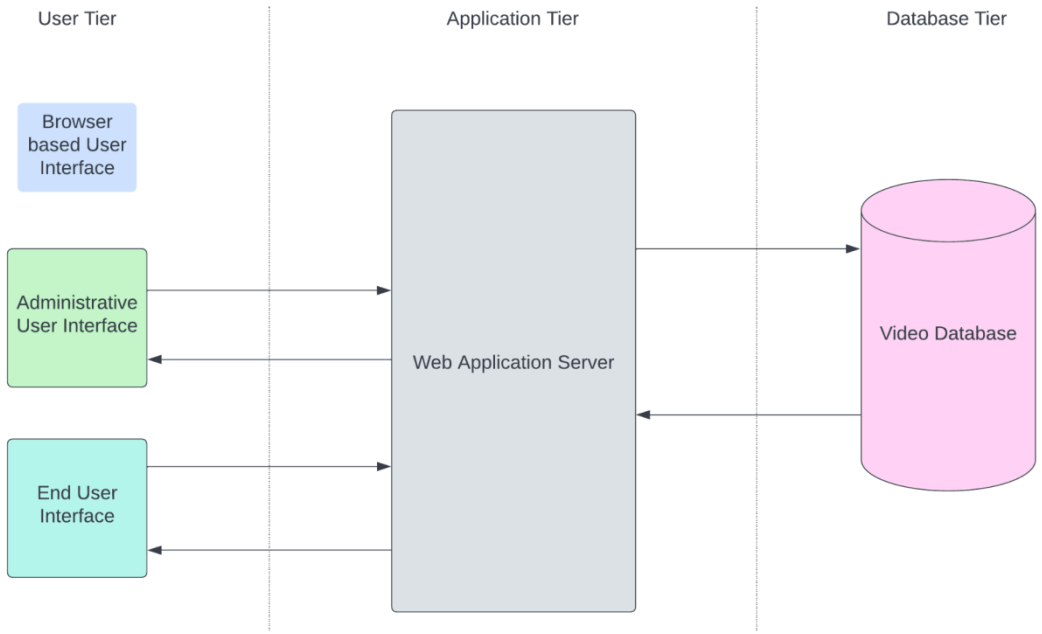

Figure 7. Tier Diagram

4.2 Design Level Diagrams

4.2.1 ER Diagram

Figure 8 showing ER diagram depicting relationships among people, objects, places, concepts or events

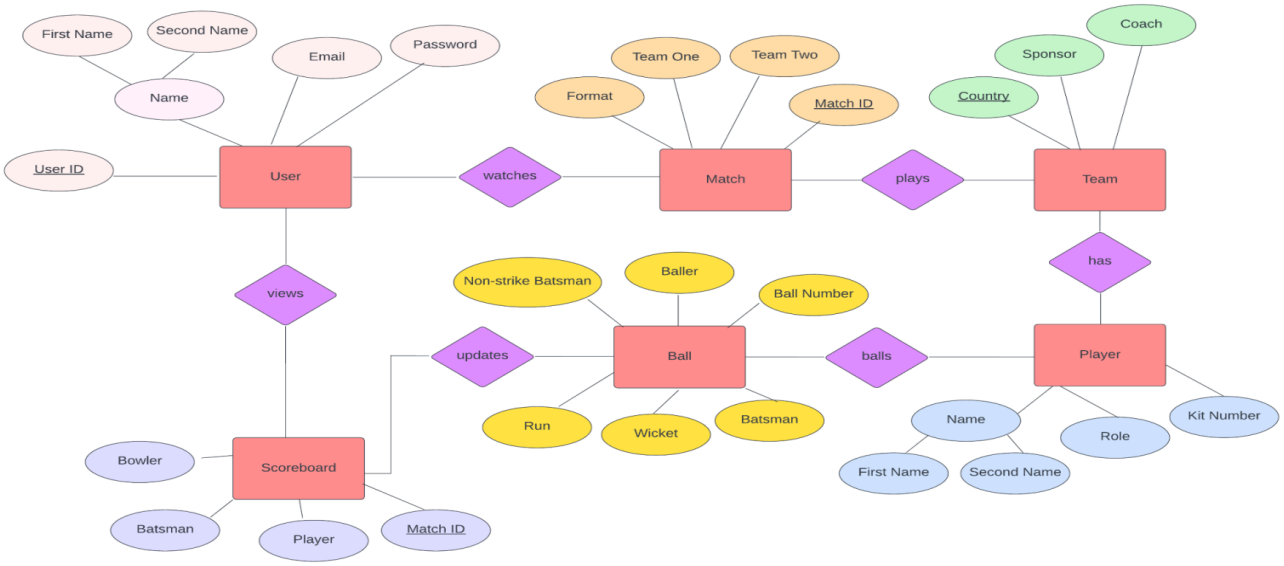

Figure 8. ER Diagram.

4.2.2 Data Flow Diagram

Figure 9 is showing the Context Diagram of CrickAlike

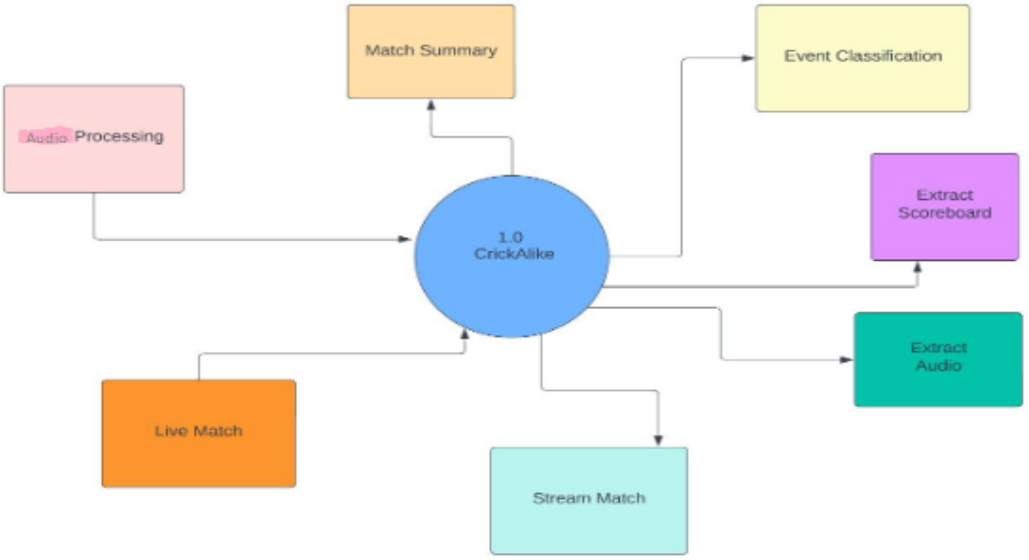

Figure 9. DFD Level 0

Figure 10 shows the DFD Level 1

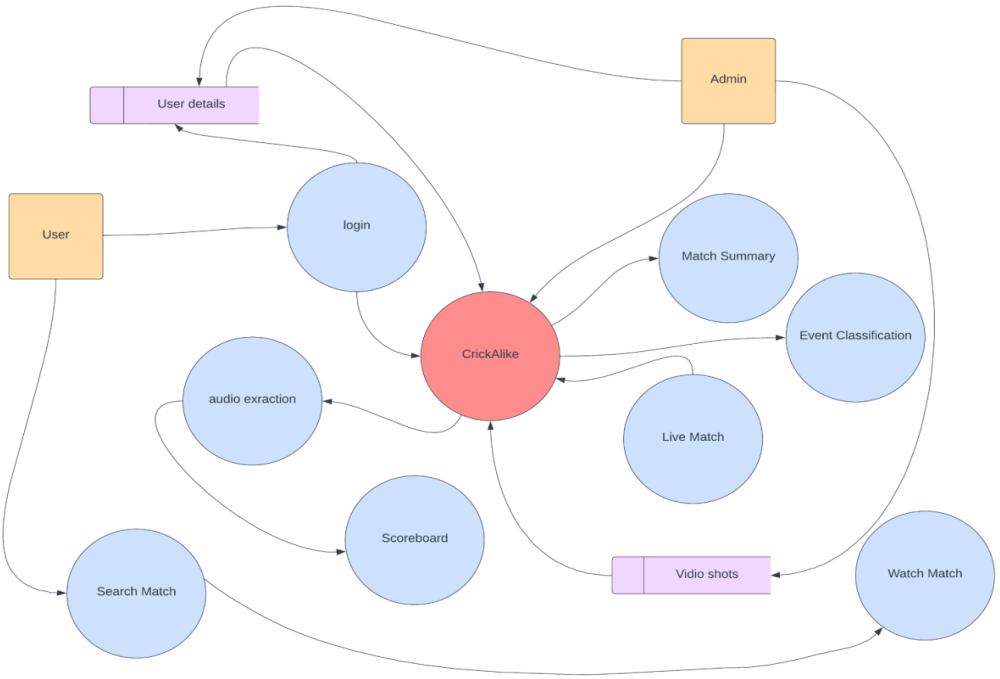

Figure 10. DFD Level 1

4.2.3 Class Diagram

Figure 11 describes the blueprint of the CrickAlike and also shows the relationship between the objects and the services they provide

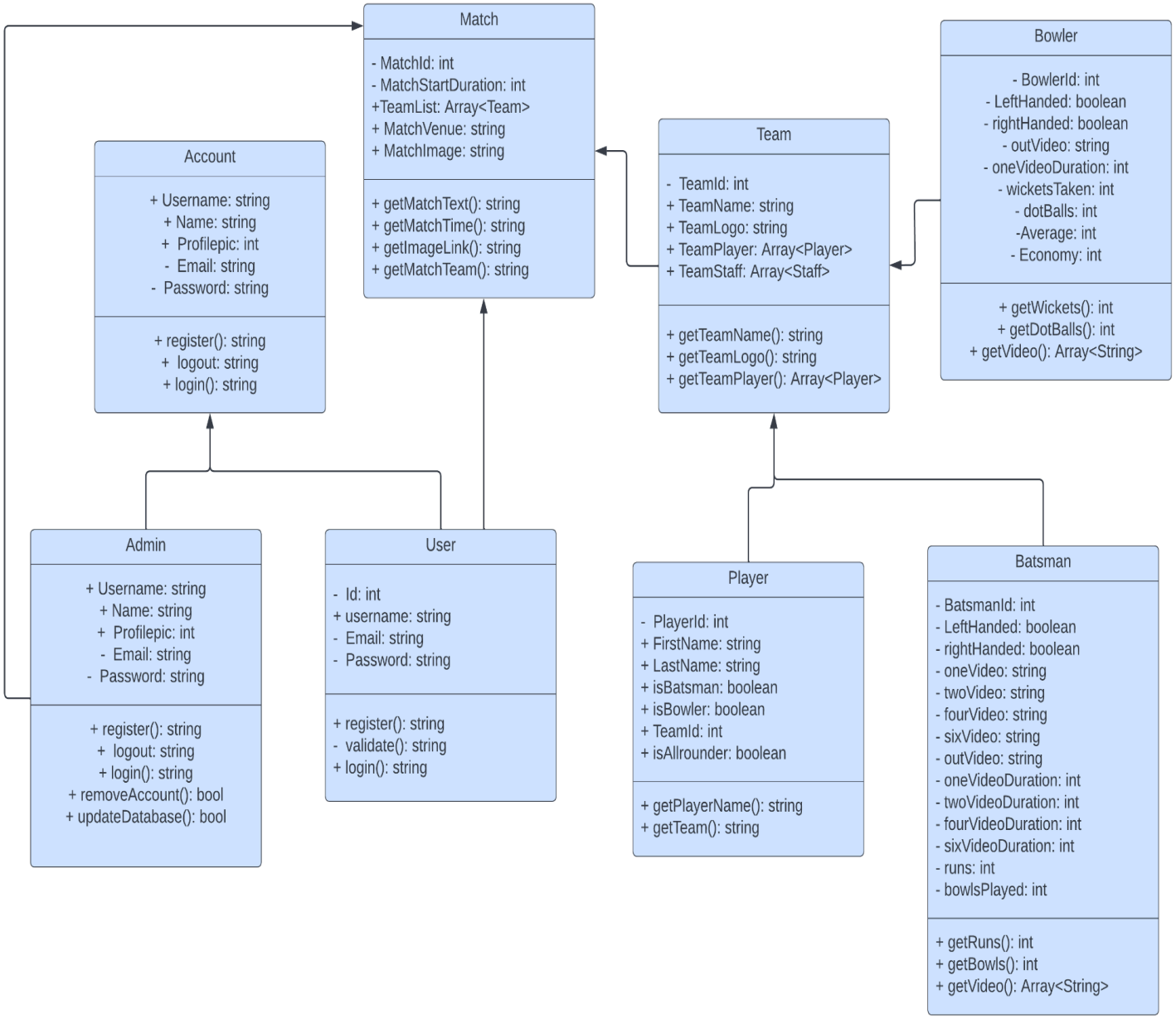

Figure 11. Class Diagram

4.2.4 Use Case Diagram

Figure 12 of Use Case Diagram identifies the interactions between the system and its actors.

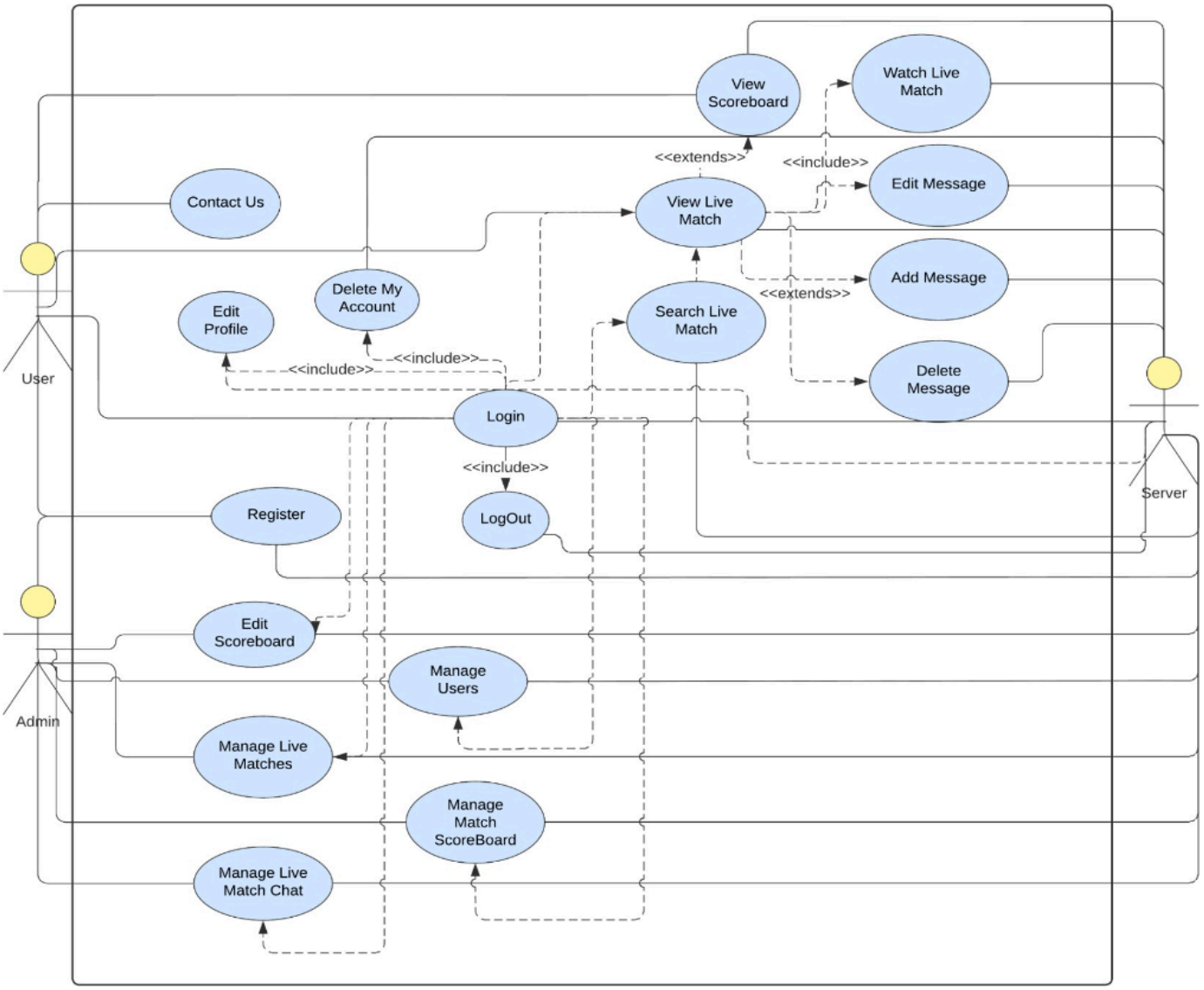

Figure 12. Use Case Diagram

4.2.5 Activity Diagram

This figure 13 of the Activity Diagram shows the business and software processes as a progression of actions.

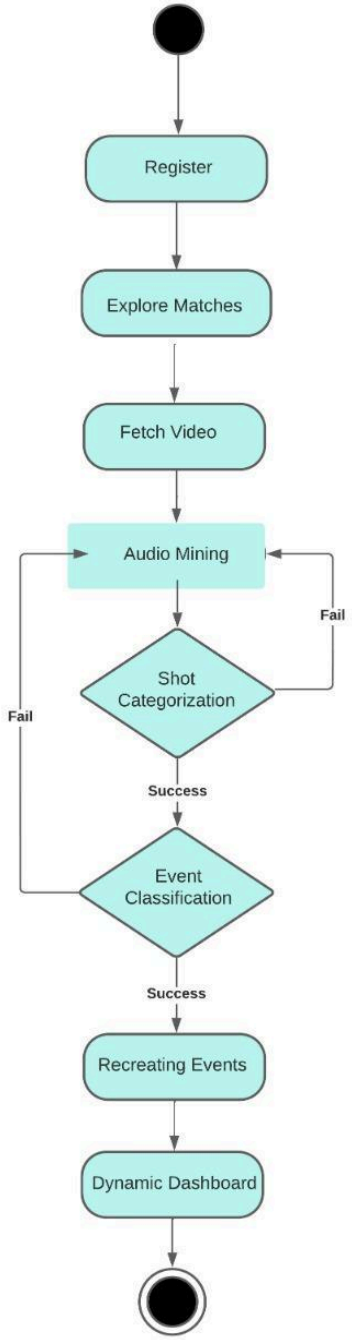

Figure 13. Activity Diagram

4.2.6 State Diagram

Figure 14 in Statechart Diagram is an illustration of the states an object can attain and the transitions between those states.

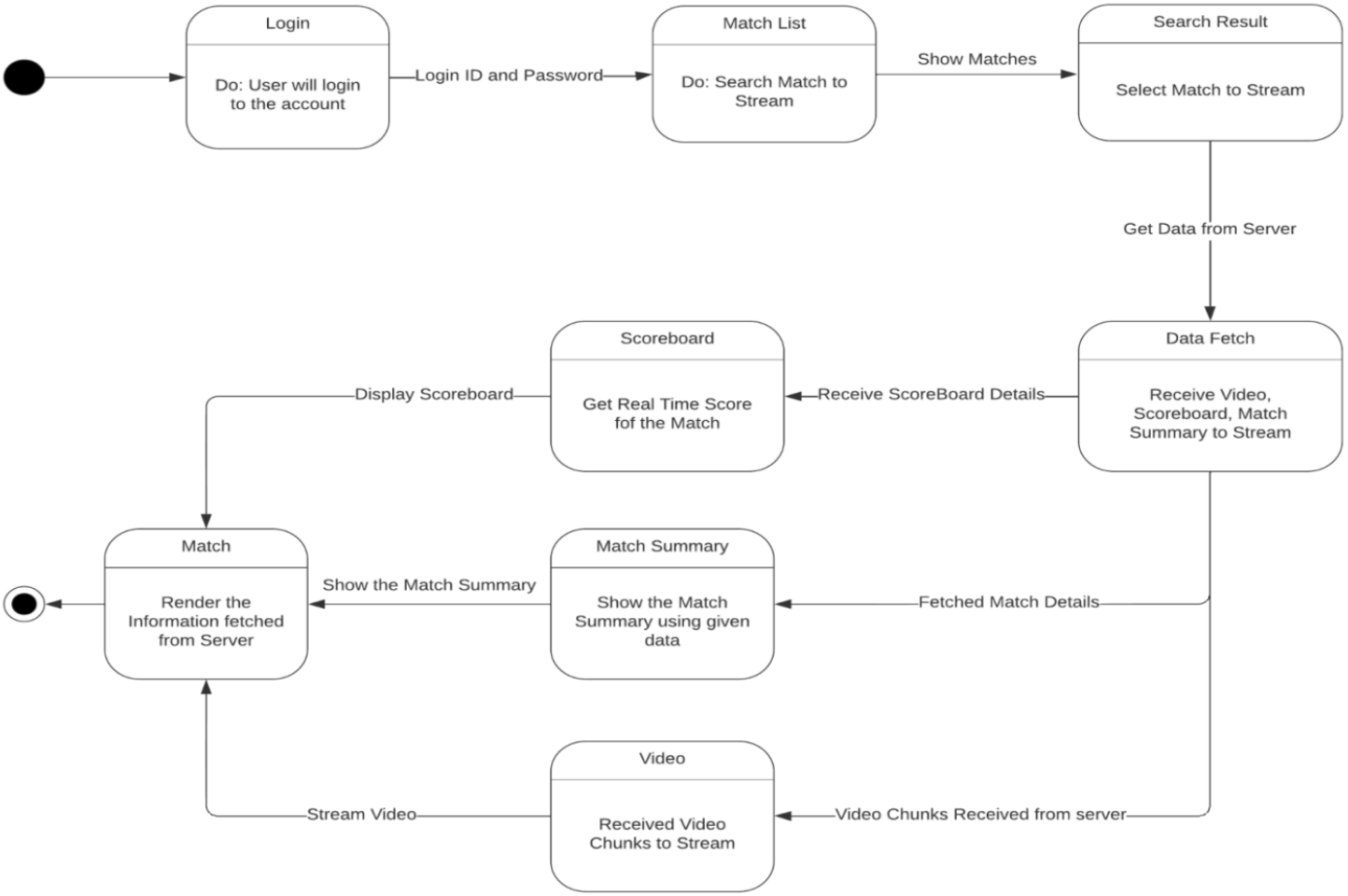

Figure 14. State Diagram

4.2.6 Sequence Diagram

Sequence Diagram in Figure 15 illustrating the sequence of messages between objects in an interaction.

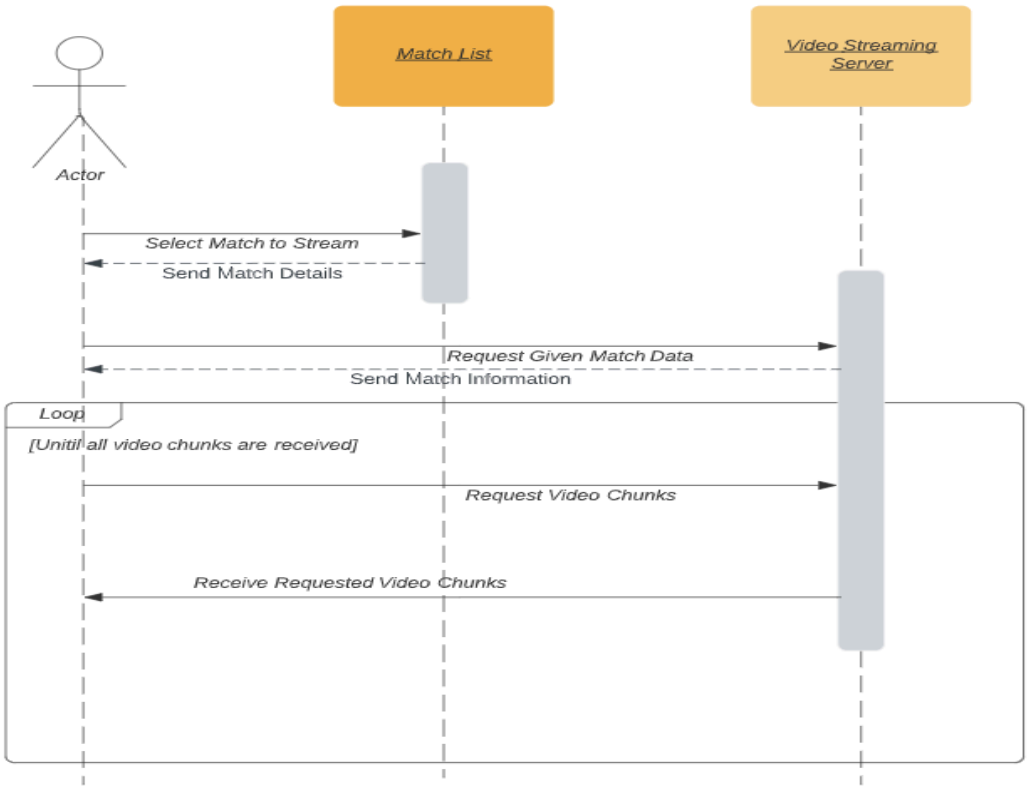

Figure 15. Sequence Diagram

4.3 User Interface Diagrams

Figure 16 showing an instance of the Database Collection.

Figure 16. Database Collection

Figure 17, 18,19 and 20 are showing the main ui.

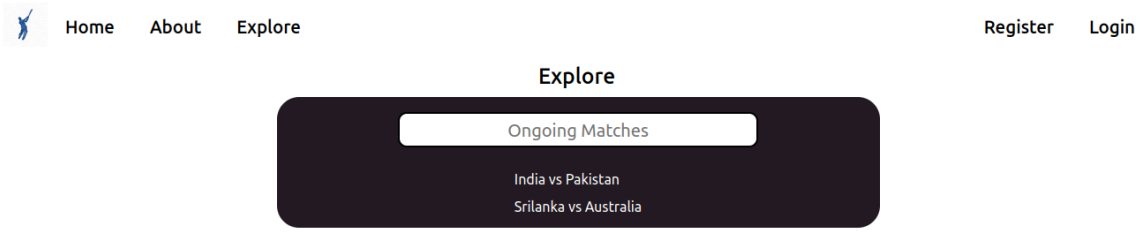

Figure 17. Main UI1

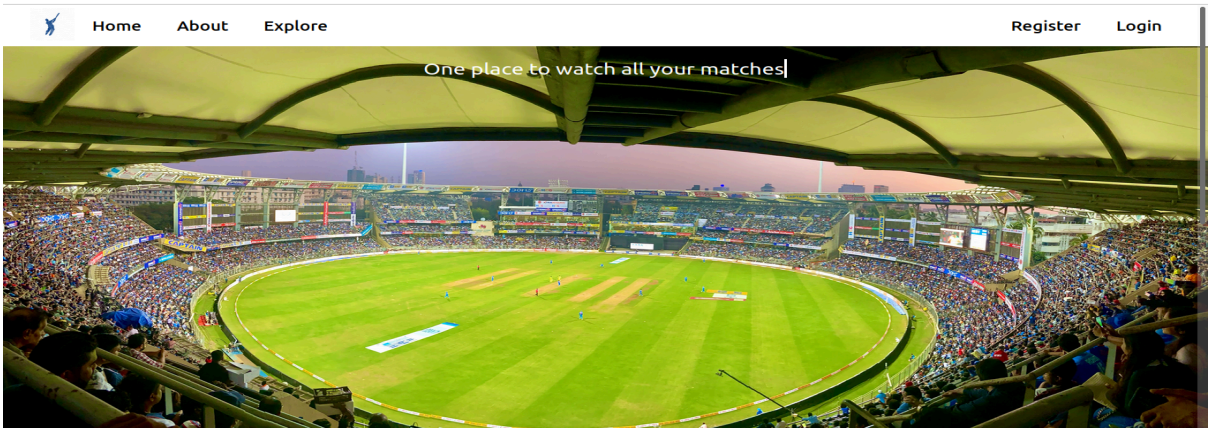

Figure 18. Main UI2.

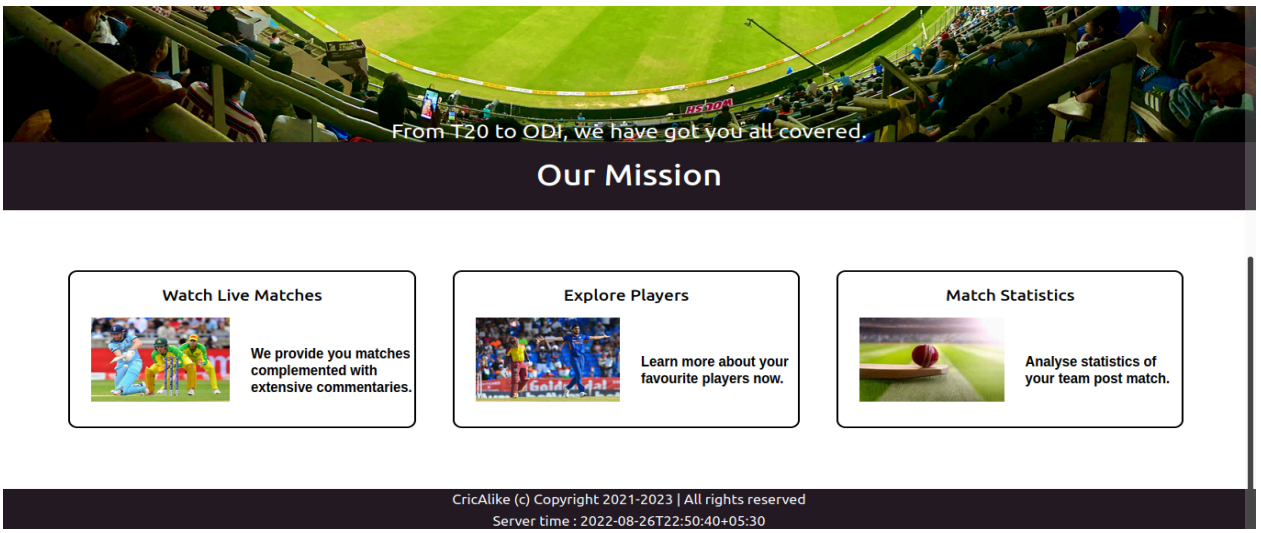

Figure 19. Main UI3.

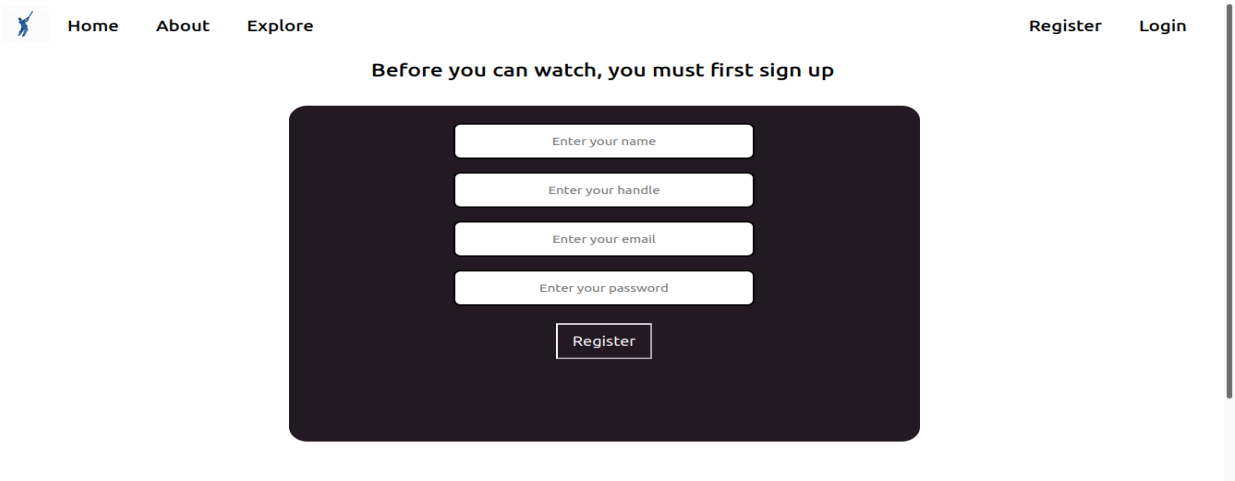

Figure 20. Main UI4.

5. Implementation And Experimental Results

5.1 Experimental Setup/Simulation

Our project's experimental setup is entirely software-based. We studied the results of different models and recognised the best performer for our final model .This experimental section was completed on kaggle using an .ipynb(Python Notebook) file and code written in Python. We used the seaborn library to plot and compare the results of various models. The visual representation of the confusion matrix aided us in determining the best model in our experiment.

5.2 Experimental Analysis

We performed experimental analysis by considering models like KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier, XGBClassifier, AdaBoostClassifier, GradientBoostingClassifier, GaussianNB, and LinearDiscriminantAnalysis for both the Shot Categorization and Event Classification models. We found that **XGBClassifier** was our best performer in the Shot Categorization model which gave the highest accuracy and minimum log loss. of **99.3** percent and **0.025** respectively. Similarly for the event classification model, XGBClassifier came out to be the best performer giving the accuracy of

5.2.1 Data

Commentary Dataset:

The dataset collection process was fully automated using Python and Selenium for web scraping to extract cricket commentaries from Cricbuzz.

The Shot Categorization model was trained over **1300** commentaries respectively extracted from several matches. Each of the commentaries was further classified into either “show” or “skip”. Show corresponds to the commentaries in which major events during the game like a Four, Six or Wicket. Skip corresponds to the

commentaries in which minor events happened that can be skipped like a single, double or a dot ball.

The Event Classification model was trained over **200** commentaries respectively extracted from several matches. Each of the commentaries was further classified into either “four”, “six” or “wicket”.

Video Dataset:

The video dataset features numerous shots of players in different directions. The video dataset was collected, bearing in mind that keeping the exact resolution with high quality of all videos provides a seamless experience to the user. The database features shots and highlights from the Cricket 2019 game. Our video datasets consist of more than 25 videos of players considering various significant events during the game.

5.2.2 Performance Parameters

Based on a variety of performance indicators, the 5 machine learning models were compared to find the best model. The most often utilized parameters are F1-score, recall, accuracy, and precision. In addition to these, it's important to take into account how long each computation takes. We were running the experiment on our individual laptops. In addition to these four factors, the models' ROC-AUC was taken into account.

1. **Accuracy** : One of the most fundamental performance metrics is accuracy, which can be defined as the number of stressed users that were predicted to be stressed , also known as the true positive, and the number of non-stressed users that were predicted to be non-stressed, also known as the true negative. The true positive and false negative parameters determine overall accuracy; the better the values of these two parameters, the better the results.

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN}$$

where TP = True Positives, TN = True Negatives, FP = False Positives FN = False Negatives.

2. **Precision** : The ratio of genuine positives to all expected positives is known as precision. Precision is the ratio of true positives to all expected positives.

$$\text{Precision} = \frac{TP}{TP+FP}$$

3. **Recall** : It is defined as the ability of a model to find all the relevant cases within a data set.

$$\text{Recall} = \frac{TP}{TP+FN}$$

4. **F1-Score** : The F1 score is the weighted average of precision and recall. In this manner, it accounts for both false positive and false negative results.

$$F1 = \frac{2*Precision*Recall}{Precision+Recall}$$

Figure 23 showing the Prototype of Confusion Matrix

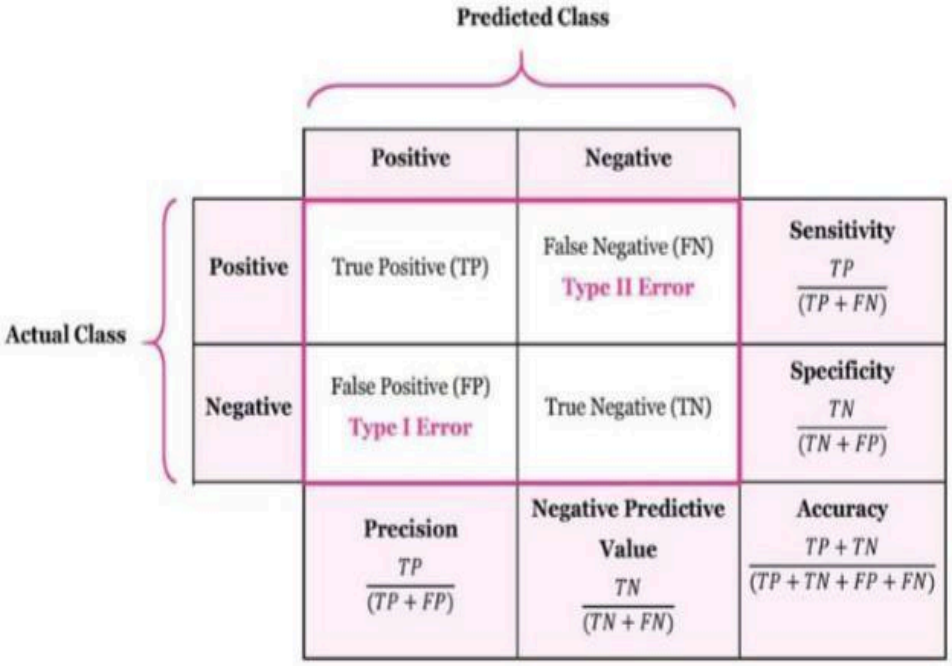

Fig 21: Confusion Matrix Prototype

5. **AUC-ROC** : A performance indicator for classification problems at various levels is the Area Under the Curve (AUC)-Receiver Operating Characteristic (ROC) curve. While ROC is a probability curve in this instance, AUC represents the degree of separability. It essentially demonstrates the model's ability to distinguish between classes; the greater the AUC, the better the model's ability to forecast whether the user is stressed or not.

5.3 Working of the Project

5.3.1 Procedural Workflow

The project works by first extracting audio features from the commentary from the match. We are utilizing one of the cognitive services provided by Microsoft for speech to text conversion. The model is currently deployed on Microsoft Azure and is secured from further access by authenticating it with a security key.

The text generated from the above model is processed further employing various features of natural language processing by removing the stopping words, special characters, tokenizing and stemming the words. The processed sentence is further segregated into a number of monograms, bigrams and trigrams.

These are further fed into the shot categorization model which uses Term Frequency-Inverse Document Frequency statistical technique. It works by evaluating the importance of a word in a document or a collection of documents (called a corpus). It is a binary classifier which categorizes the commentary into either “show” or “skip”. Show corresponds to the commentaries which we would like users to see and skip are the ones which are ignored.

Commentaries that were categorized as “show” are further fed into the event classification model which is a ternary classifier which classifies the commentary into either “four”, “six” or “wicket”.

Corresponding to the event classified from the Event Classification model appropriate video is fetched and displayed to the user.

5.3.2 Algorithmic Approaches Used

The algorithmic approach employed for the development of the classification model are as follows.

1. Generate the dataset for shot categorization and event classification by web scraping through Selenium.
2. Shot Categorization dataset is categorized into show and skip.
3. Event Classification dataset is categorized into four, six and out.
4. Synthetic Minority Oversampling technique is employed while splitting both the datasets into training and testing sets in the ratio of 3 : 1.
5. Both the models are trained over 9 different machine learning models and accuracy and log loss are computed for all.
6. XGBClassifier outperformed all the models by giving the highest accuracy and minimum log loss.
7. The commentary is generated from the audio of the commentators using Microsoft Azure - Speech to Text.
8. The text is further separated into monograms, bigrams, and trigrams. The generated fragments are processed using Term Frequency-Inverse Document Frequency technique for generating numerical features.
9. The generated features are further used to train the shot categorization and event classification model using XGBClassifier.
10. The generated models are downloaded as binary in the form of pickled files.

5.3.3 Project Deployment

The optimal Machine learning model obtained was deployed through our web application which is now currently hosted locally. The front-end of the application was made using React Js, while Flask was used to develop the back-end for the same.

Deployment diagram in Figure 24 is representing the physical architecture of our system

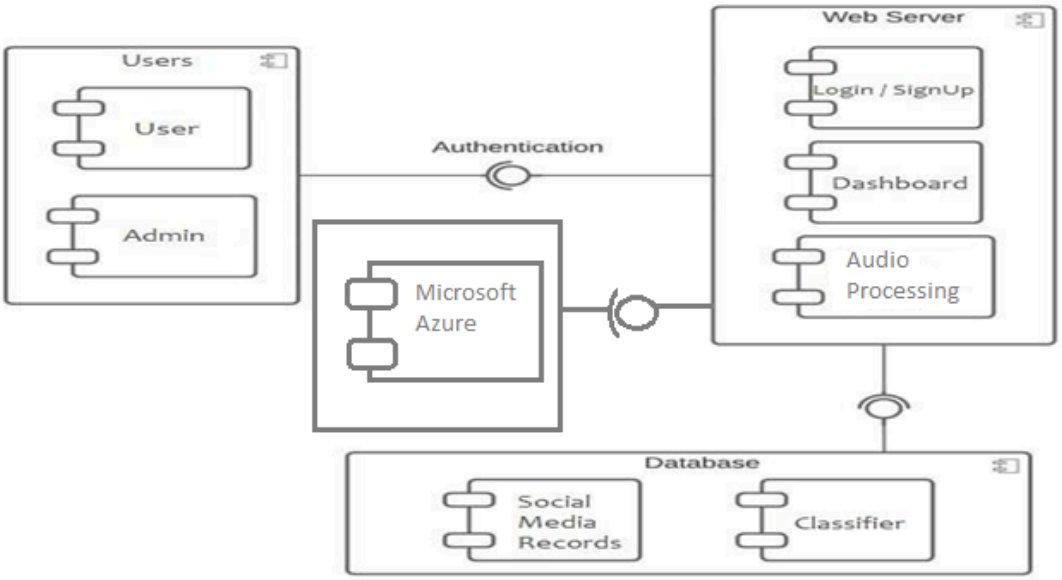

Fig 22: Project Deployment

5.3.4 System Screenshots

Figure 23 and 24 showing live match simulating

Figure 23. Match Simulation 1.

Figure 24 Match Simulation 2

Figure 25 and 26 showing final Ui

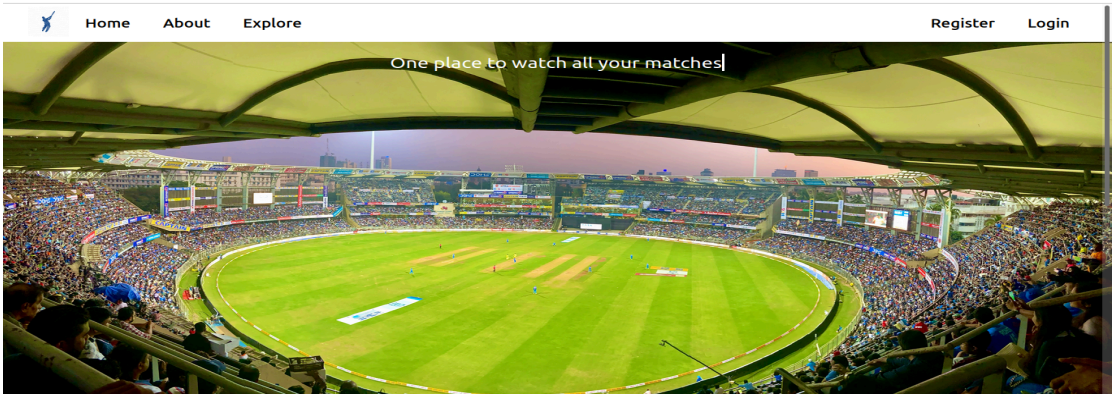

Figure 25. Final UI1.

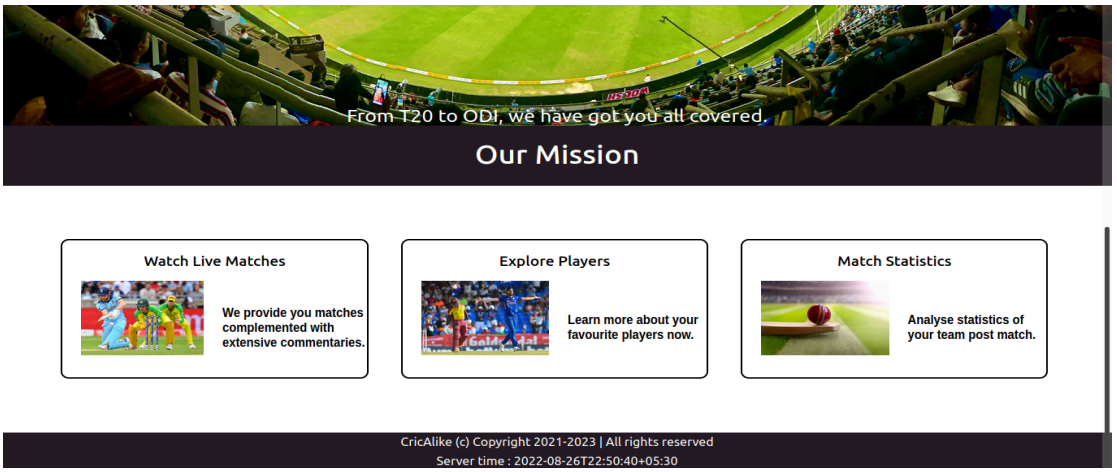

Figure 26. Final UI2.

5.4 Testing Process

5.4.1 Test Plan

With this platform, users who cannot afford hefty subscription fees can watch live matches through animated content that closely mimics the live match. Our goal is to provide users with an experience that falls somewhere between tedious comments and the live game, so everyone can enjoy the same level of excitement and passion. A few major events during the match are recreated by extracting meaningful features from the audio and video.

5.4.2 Features to be tested

1. Users can Register/Login into the website.
2. Correct Implementation of Comments Section so that users can talk with each other about the event happening live in a cricket match.
3. Users can view Live Cricket Match
4. Correct workflow of the whole website.
5. Users can view Match Summary for quick brief about the Match.
6. Successful Extraction of meaningful features from the audio and video

5.4.3 Test Strategy

The test strategy is to validate each feature one by one and thus validate the project deliverables. Manual testing was performed to test each feature one by one. If the testing produces the desired result for each feature without freezing or taking longer times, it is considered a pass, otherwise fail.

5.4.4 Test Techniques

Unit Testing : Individual components of the software are tested in unit testing. Unit testing is a software testing method by which individual units of source code-sets of one or more computer program modules together with associated control data, usage procedures, and operating procedures-are tested to determine if they can be used as intended.

Acceptance Testing : The acceptance test is a quality assurance (QA) process that determines whether or not an application meets the requirements of its end-users. A beta test, application test, field test or end-user test can all be used for acceptance testing, depending on the organization.

5.4.5 Test Cases

Test Case#: T001

Subsystem : Registration & Login

Short Description : Testing the register feature when a new user registers on the website.

Pre-Condition :

Step	Action	Expected Response	Pass / Fail
1.	Click on the Register button on the sign in page.	Register / sign up page opens up.	**PASS**
2.	User enters all the credentials	Details are filled in the input box.	**PASS**
3.	User clicks on the submit button.	A new user record is created in the database and the user is redirected to Evaluate Page with user name shown in navbar.	**PASS**

Table 7. Pre-Conditions - Test Case 1

Post Condition: The user is now registered/logged-in with our system.

Test Case#: T002

Subsystem : Correct Website URL Routing

Short Description : Testing that every link on a website goes to its specified destination.

Pre-Condition :

Step	Action	Expected Response	Pass / Fail
1.	Click on Home.	Home Page is shown.	**PASS**
2.	Click on Match Summary	The Match Summary section is shown.	**PASS**
3.	Click on Get Scoreboard details	Page shows Real time score for match	**PASS**
4.	Click on Video	Page shows Animated content	**PASS**

Table 8. Pre-Conditions - Test Case 2

Post Condition: Every link redirects to the correct page/section.

Test Case#: T003

Subsystem : Extracting Details from Audio and video

Short Description : Testing the entire feature of Animating from audio and video input.

Pre-Condition :

Step	Action	Expected Response	Pass / Fail
1.	Users Select the match and select the option for viewing.	Match Animation comes up for selected Match	**PASS**

Table 9. Pre-Conditions - Test Case 3

Post Condition: The user is able to use our CricAlike Website and is able to View the animated match based on the match recording(audio or video) given as input.

5.4.6 Test Results

Each feature of the system is validated and works correctly. The system works on an acceptable level.

5.5 Results and Discussions

Shot Categorization: A binary classifier which predicts a commentary to show or skip the current ball. The model is tested over **595** commentaries and trained over **2380** commentaries.

1. **Accuracy:** An accuracy of **99.32** is achieved.
2. **Confusion Matrix:** A confusion matrix is a table that is used to evaluate the performance of a classification model.

Figure 34 showing the performance of the model using Predicted Label.

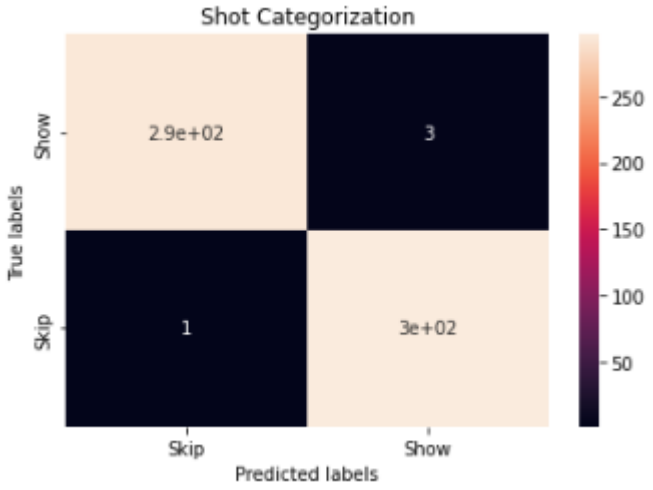

Figure 32. Predicted Label.

3. Precision: The precision scores for show and skip classes came out to respectively **0.99** and **1**.
4. Recall: The recall scores for show and skip classes came out to respectively **1** and **0.99**.
5. F1 score: The F1 scores for show and skip classes came out to respectively **0.99** and **0.99**.

Figure 33 showing the performance of the model using F1 Score.

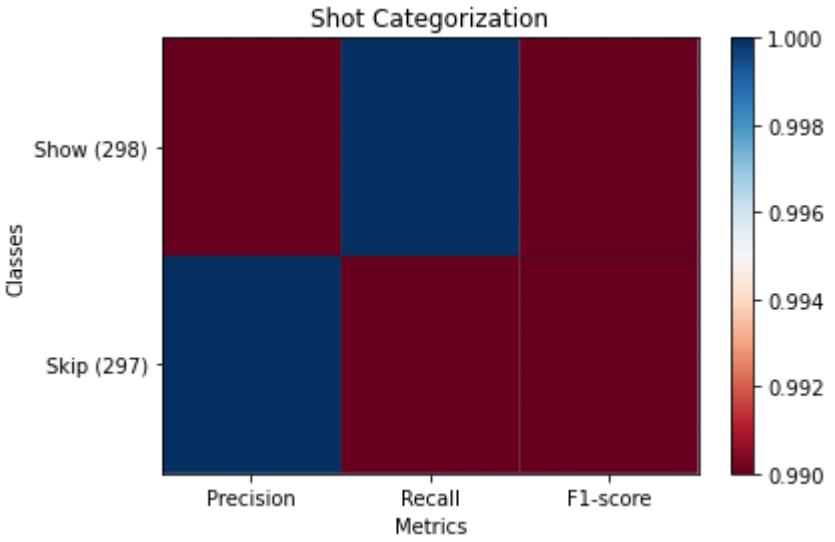

Figure 33. F1-Score.

Event Classification: A ternary classifier that predicts a commentary to show or skip the current ball. The model is tested over 73 commentaries and trained over 216 commentaries.

1. **Accuracy:** . An accuracy of **98.63** is achieved.

Figure 34 showing the performance of the model using Confusion Matrix .

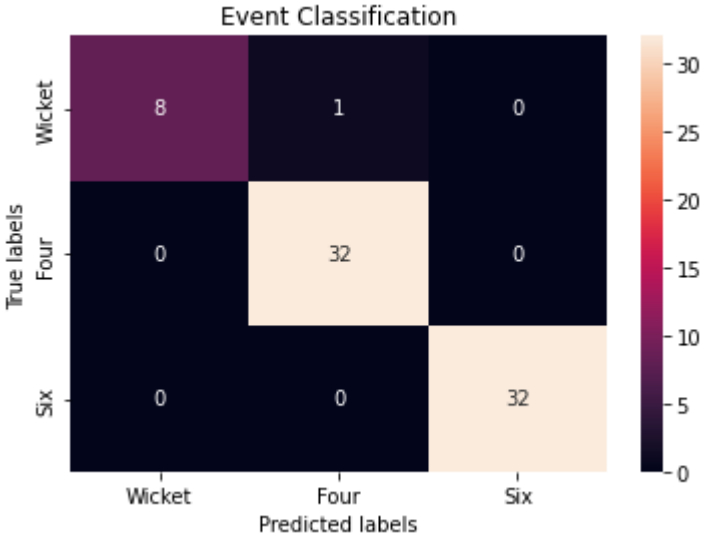

Figure 34. Confusion Matrix.

2. Precision: The precision scores for wicket, four and six classes came out to be **1**, **0.97** and **1** respectively.
3. Recall: The recall scores for wicket, four and six classes came out to be **0.89**, **1** and **1** respectively.
4. F1 score: The F1 scores for wicket, four and six classes came out to be **0.94**, **0.98** and **1** respectively.

Recall Metrics in figure 35 is showing the actual positive cases that were correctly predicted.

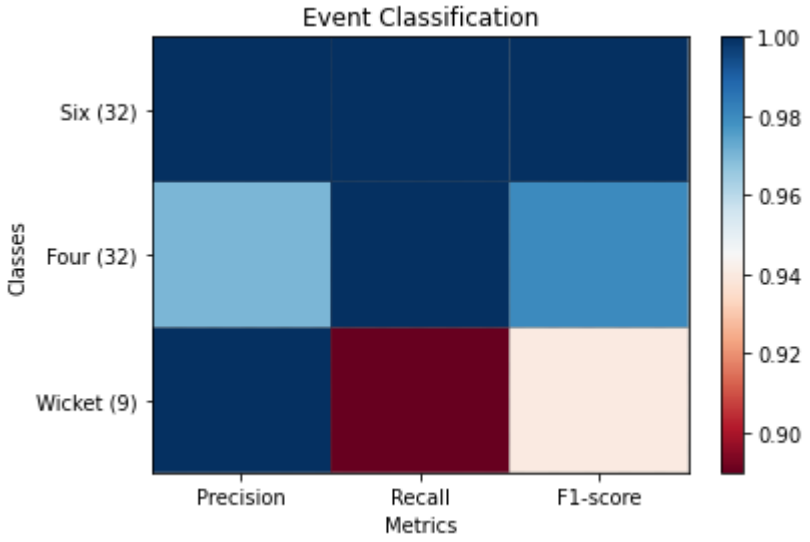

Figure 35. Recall Metrics.

5. AUC-ROC: This stands for "Area Under the Receiver Operating Characteristic curve" and is a measure of the model's ability to distinguish between positive and negative cases.

We found that **XGBClassifier** was our best performer in the Shot Categorization model which gave the highest accuracy and minimum log loss. of **99.3** percent and **0.025** respectively. Similarly for the event classification model, XGBClassifier came out to be the best performer giving the accuracy of

```
==============================
KNeighborsClassifier
Accuracy: 82.1849%
Log Loss: 2.5456567952907774
==============================
DecisionTreeClassifier
Accuracy: 98.9916%
Log Loss: 0.34829018213355406
==============================
RandomForestClassifier
Accuracy: 98.8235%
Log Loss: 0.07857162099932807
==============================
XGBClassifier
Accuracy: 99.3277%
Log Loss: 0.025336123353716312
==============================
AdaBoostClassifier
Accuracy: 98.9916%
Log Loss: 0.4311044731944466
==============================
GradientBoostingClassifier
Accuracy: 98.8235%
Log Loss: 0.03957002551500513
==============================
GaussianNB
Accuracy: 98.6555%
Log Loss: 0.4643869095114051
==============================
LinearDiscriminantAnalysis
Accuracy: 97.9832%
Log Loss: 0.6965867386273971
==============================
```

```
==============================
KNeighborsClassifier
Accuracy: 56.1644%
Log Loss: 10.581658411056834
==============================
DecisionTreeClassifier
Accuracy: 94.5205%
Log Loss: 1.8925356928718198
==============================
RandomForestClassifier
Accuracy: 93.1507%
Log Loss: 0.2514219757056789
==============================
XGBClassifier
Accuracy: 98.6301%
Log Loss: 0.06475129390569175
==============================
AdaBoostClassifier
Accuracy: 94.5205%
Log Loss: 0.14826666893571816
==============================
GradientBoostingClassifier
Accuracy: 94.5205%
Log Loss: 0.1645529670289835
==============================
GaussianNB
Accuracy: 95.8904%
Log Loss: 1.4194017696538654
==============================
LinearDiscriminantAnalysis
Accuracy: 75.3425%
Log Loss: 0.6897527584029106
==============================
```

Figure 36. Results of various models

Figure 37. Showing the Shot Categorization models accuracy.

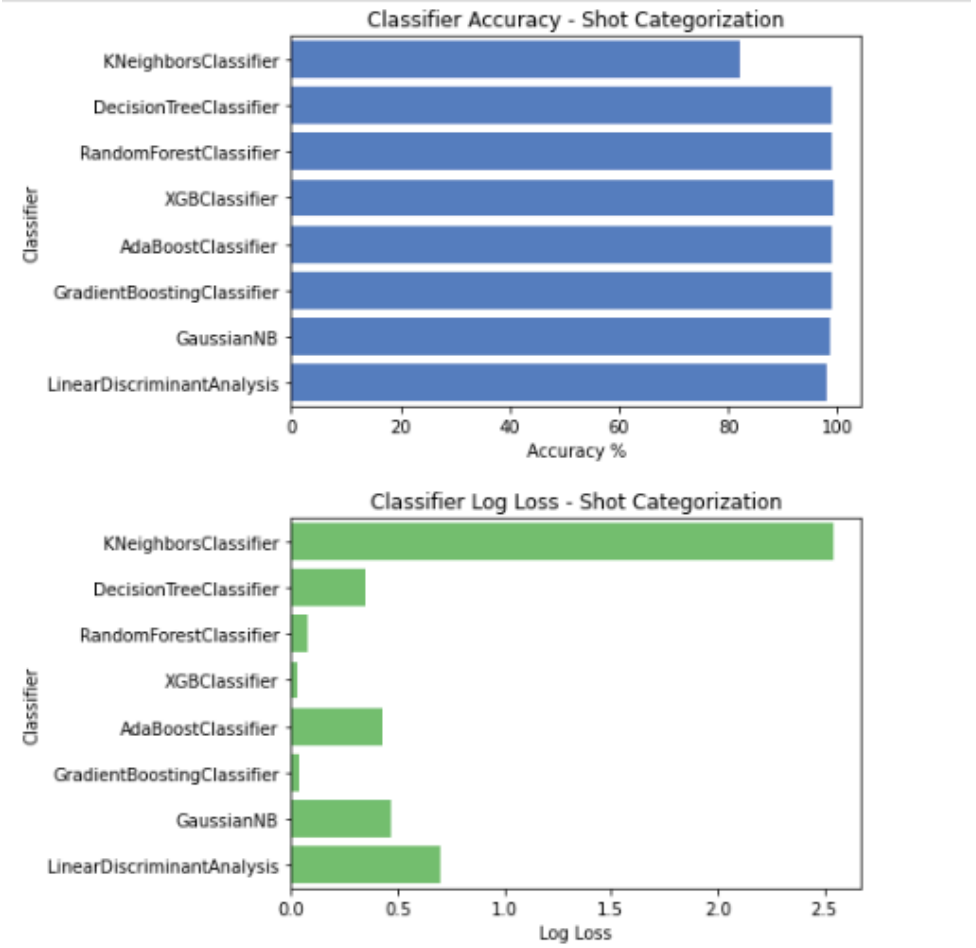

Figure 37. Classifier Accuracy and Log Loss - Shot Categorization.

Figure 37. Showing the accuracy of the Event Classification Model.

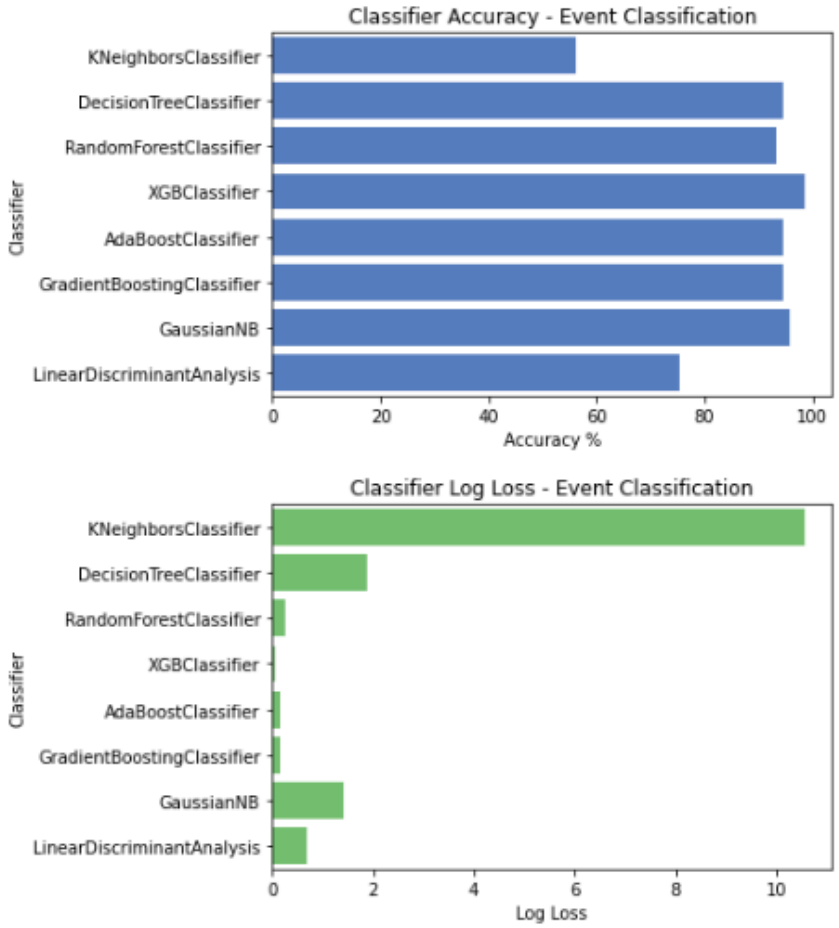

Figure 38. Classifier Accuracy and Log Loss - Event Classification.

5.6 Inference Drawn

The inference drawn from our experimental study was tabulated below in form of a table by considering different models and techniques used in different research papers during our literature survey.

Author	ML approaches used	Tools/Dataset Used	Result/Accuracy
Ramalatha Marimuthu	Deep Learning Model for Speech Recognition. GDPO PO	Taylor-gradient Descent political optimization based Deep residual network	The proposed Taylor GDPO-based DRN offered effective performance with highest accuracy of 96.93%, smallest FAR of 2.438%, smallest FRR of 2.101%, smallest MSE of 0.038.
Metin Turan	The extraction of inferential written summaries of communications that occur in oral environments such as meetings, lectures and conferences.	Speech Recognition Python library, Google's cloud Application Programming Interface (API) etc.	The results obtained are remarkable and it is seen that approximately 71% success was achieved.
Banoth Thulasya Naika	Detection and classification of players,tracking player or ball in sports, predicting the trajectories of player or ball,detection and classification of each player based on their team in every frame	Deep Learning and AI techniques used.	Got Maximum accuracy of about 0.614 by using the classifiers.

Vladyslav Tsap	SVM Random Forest Naive Bayes Deep Neural Network Best Performance – DNN	The dataset of the University of Belgium compiled is used.	This algorithm can improve the result compared to "strong" classifiers upto 30%. However, there are cases when even a difference of 3-5% is significant.
R Rahman	AdaBoost SVM KNN Random Forest Logistic Model Tree Best Performance - SVM	Collected their data from ICC GRIP DATASET.	The results obtained are remarkable and it is seen that approximately 76% success was achieved with Accuracy of about 0.767.
Pinal Shah	Techniques and algorithms that are applied to achieve functionalities like Speech to Text, Text to speech, Speech recognition, communication, Machine translation etc.	IBM Watson Developer Cloud, Google Translate	Got Maximum accuracy of about 0.835 by using the classifiers.

Table 10. Inference Drawn

5.7 Validation of Objective

Each suggested goal is mapped out with its execution outcome in the table that is included below.

S.No.	Objective	Status
1	To convert audio of the commentary to text.	Successful

2	To predict whether event occur (six,four or out) or not and after that classify the event in six, four or out	Successful
3	To evaluate performance using the best model after comparing other ML models.	Successful
4	To create a user-friendly user interface (UI) for the project that can accurately run the model and display live match with animation	Successful

Table 11: Validation of Objectives

6. Conclusions and Future Scope

6.1 Conclusion

There is nothing more Indian than watching a cricket match. Due to increasing subscription prices and media rights prices, it is unlikely that anyone will be able to watch sports for free in the near future.

Final Goal is to allow users to watch a video which gives users experience closest to the real match at free of cost. Also, to make a session more interactive Users can chat with each other about the current event happening in a cricket match. Match Summary is also provided to give users a quick recap about the current match.

6.2 Environmental / Economic Benefits

The developed system has tremendous potential to benefit people with less access to the internet(animated content would consume significantly less internet than live streaming) and less money for hefty subscriptions.

6.3 Reflections

This project gave us an opportunity to enhance many skills such as teamwork, time management, decision-making, and multitasking. As a result of these reflections from the project, we were able to develop both personally and academically.

6.4 Future Work Plan

The goal for the future is to continue improving the UI/UX of our website. Additionally, to upscale and enhance our website in case more users register. Apart from this we aim to Enhance our video processing and rendering so that we can convert into animations without any delay . This can be handled by a physical server or cloud service. Additionally, we wish to continue our support to more sports apart from Cricket.

7. Project Metrics

7.1 Challenges Faced

The dataset, the efficiency of the programming, and teamwork are the issues that were encountered while finishing this project over the course of about 10 to 11 months.

1) The hardest task we had to complete for this project was locating the relevant datasets.

2) Due to our lack of access to any high-end cloud-based GPU services, we were only able to run the experiment on our personal computers, which made it difficult to estimate the computation time for each model.

7.2 Relevant Subjects

S.No	Subject Name	Subject Code
1	Machine Learning	UML501
2	Software Engineering	UCS503
3	Building Innovative Systems	UCS757
4	Data Science Fundamentals	UCS538
5	AI-NLP, Computer Vision	UCS655
6	Database Management System	UCS310

Table 12: List of Relevant Subjects

7.3 Interdisciplinary Knowledge Sharing

1. The fundamentals of software engineering helped us make significant headway in figuring out how the project worked. Each diagram was made using key ideas related to the subject, such as use cases, data flows, and others. Additionally, the steps of software engineering were taken into consideration when developing the functioning model and its prototype.
2. We created a powerful and precise machine learning model to forecast the batsman's strokes during a live game using commentary, as well as whether it will be four, six or out using ideas from machine learning.
3. The database required for the project was easy to manage thanks to the principles learned in the Database Management System course.

7.4 Work Schedule

S.No.	Activity	Month	Marc	April				May				July	August				September				October				November				December			
		Week No.	1	2				6				10	12				16				20				24				28				
1	Identification formulation	Plan																															
	and Planning of project	Actual																															
2	Study of ML approaches	Plan																															
	and models	Actual																															
3	Implementation of prediction	Plan																															
	system without GUI	Actual																															
4	Application (front-end)	Plan																															
		Actual																															
5	Integration and GUI	Plan																															
		Actual																															
6	Design Optimisation	Plan																															
		Actual																															
7	Performing modifications	Plan																															
		Actual																															
8	Results evaluation	Plan																															
		Actual																															
9	Final Report	Plan																															
		Actual																															

Fig 39: Gantt Chart

7.5 Student Outcomes Description and Performance Indicators

SO	SO Description	Outcome
1.	Ability to identify and formulate problems in our project scope and how to handle its solution.	Identified the problem of obtaining a Machine learning based framework for predicting shots of batsmen..
2.	Applying engineering and science to formulate a solution and find an optimal ML classifier for our project.	Experimental study was conducted on 5 different models and the most optimized model was found.
3.	Prepare and present a variety of documents and report according to norms and standards given.	The entire report was done on time taking IEEE format into consideration.
4.	Able to communicate with team members and the mentor to make sure the project is going towards a positive end.	Regular meetings were done between team members and with the mentor to discuss future work plans.
5.	Ability to find and utilize resources given to us and find new ones during the development cycle.	Technologies and tech stack were identified during the initial phase and new resources were explored and used in between the development cycle.

Table 14: Student Outcomes Description and Performance Indicators

APPENDIX A: Classifier Pseudo Code

Pseudo code for Shot Categorization and Even Classification models

1. Import the necessary libraries including numpy for numerical computing, pandas for data manipulation, SMOTE for oversampling, and various machine learning and natural language processing libraries including sklearn, xgboost, hyperopt, gensim, and nltk.
2. Read the data from a CSV file using the pandas.read_csv() function and store it in a Pandas dataframe.
3. Define a function text_cleaning that takes in a dataframe and a column name, and performs text cleaning on the specified column. The function should do the following:
    - Convert the data in the specified column to an array using the np.array() function.
    - Iterate over the array, and for each element:
    - Remove stopwords from the element using the gensim.parsing.preprocessing.remove_stopwords() function.
    - Initialize a Porter stemmer object using nltk.stem.PorterStemmer().
    - Iterate over the words in the modified element. For each word:
    - Remove non-alphabetic characters using a regular expression.
    - Strip leading and trailing whitespaces from the word.
    - Convert the word to lowercase.
    - Stem the word using the Porter stemmer object.
    - Strip leading and trailing whitespaces from the stemmed word.
    - Concatenate the stemmed words into a stemmed sentence.
    - Update the element in the dataframe with the stemmed sentence.
4. Apply the text_cleaning function to the 'commentary' column of the dataframe.
5. Define a function vectorize that takes in a list of text and a TfidfVectorizer object, and returns a dataframe of Tfidf vectors for the input text. The function should do the following:

- Use the transform() method of the TfidfVectorizer object to transform the input text into Tfidf vectors.
- Get the list of words used.

6. Initialize a TfidfVectorizer object using the TfidfVectorizer() function, with the following parameters:

- analyzer set to 'word'
- stop_words set to None
- ngram_range set to (1, 2)
- max_df set to 0.5
- use_idf set to True
- smooth_idf set to True
- max_features set to 1500

7. Fit the TfidfVectorizer object on the 'commentary' column of the dataframe using the fit() method.

8. Reset the index of the dataframe using the reset_index() method with the drop parameter set to True and the inplace parameter set to True.

9. Use the vectorize function to create a Tfidf feature dataframe for the 'commentary' column, using the TfidfVectorizer object fit in step 6.

10. Reset the index of the Tfidf feature dataframe using the reset_index() method with the drop parameter set to True and the inplace parameter set to True.

11. Concatenate the original dataframe and the Tfidf feature dataframe using the pd.concat() function, with the axis parameter set to 1.

12. Drop the 'commentary' column from the dataframe using the drop() method with the columns parameter set to 'commentary' and the inplace parameter set to True.

13. Use the SMOTE library to perform oversampling on the data using the fit_resample() method.

14. Split the oversampled data into training and test sets using the train_test_split() function from sklearn.model_selection, with the stratify parameter set to the target variable and the random_state parameter set to 1.

15. Initialize a model object either as a RandomForestClassifier or an XGBClassifier.

16. Train the model on the training data using the fit() method.

17. Use the trained model to make predictions on the test data using the predict() method.

18. Calculate the accuracy of the model on the test data using the accuracy_score() function from sklearn.metrics.

19. Define a function for hyperparameter optimization using the hyperopt library. The function should use the fmin() function to search for the optimal hyperparameters using the tpe algorithm.

20. Use the hyperopt function to optimize the hyperparameters of the model.

21. Train the model with the optimized hyperparameters on the entire oversampled data.

22. Use the trained model to make predictions on a new dataset.

23. Save the trained model using the pickle library.

APPENDIX B: References

[1] Avala Durga Surya Teja Rahman, Nalajala Ganesh. "Easy Live Cricket Score Stream with Boredom Games Features-A Survey" In 2021 2th International Conference on Intelligent Engineering and Management (ICIEM), pp. 630-636. IEEE, 2021

[2] Live Cricket Scoreboard websites https://www.espncricinfo.com/live-cricket-score

[3] 61% Indians Use Internet In 2021, Up From Just 21% In 2017 Says Report https://www.indiatimes.com/technology/news/india-internet-usage-report-554181.html

[4] CricBuzz https://shorturl.at/uwGP8

[5] Cricket draws 93% of sports viewers in india https://www.shorturl.at/dpPR7

[6] RAMALATHA MARIMUTHU , Speech recognition using Taylor-gradient Descent political optimization based Deep residual network, Computer Speech & Lan-guage (2022)

[7] Neslihan Akar & Metin Turan, 2022. "A General Approach for Meeting Summarization: From Speech to Extractive Summarization," Review of Computer Engineering Research, Conscientia Beam, vol. 9(2), pages 83-95.

[8] Banoth Thulasya Naika , Mohammad Farukh Hashmia , Neeraj Dhanraj Bokdeb, Zaher Mundher Yaseenc."A Comprehensive Review of Computer Vision in Sports: Open Issues, Future Trends and Research Directions" PY - 2022/03/03

[9] MoMLeT+DS 2021: 3rd International Workshop on Modern Machine Learning Technologies and Data Science, June 5, 2021, Lviv-Shatsk, Ukraine

[10] Rahman, Rafeed, Mehfuz A. Rahman, Md Saiful Islam, and Mahady Hasan. "DeepGrip:Cricket Bowling Delivery Detection with Superior CNN Architectures." In 2021 6th International Conference on Inventive Computation Technologies (ICICT), pp. 630-636. IEEE, 2021.

[11] IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 2, Ver. I (Mar.- Apr. 2018), PP 36-43

[12] Md Ferdouse Ahmed Foysal, Mohammad Shakirul Islam, Asif Karim, Nafis Neehal. "Shot-Net: A convolutional neural network for classifying different cricket shots." In InternationalConference on Recent Trends in Image Processing and Pattern Recognition, pp. 111-120.Springer, Singapore, 2018.

[13] M.H. Kolekar, K. Palaniappan, and S. Sengupta, “Semantic event detection and classification in cricket video sequence,” in Proceedings Sixth of Indian Conference on Computer Vision, Graphics and Image Processing, Bhubaneshwar, India, Dec. 2008, pp. 382-389.

[14] K.P. Sankar, S. Pandey, and C.V. Jawahar, “Text driven temporal segmentation of cricket videos,” in ProceedingsSixth of Indian Conference on Computer Vision, Graphics and Image Processing, Madurai, India, Dec. 2006, pp.433-444.

[15] I Koprinska, and S. Carrato, “Temporal video segmentation: A survey,” Signal Processing: Image Communication, vol. 16, Jan. 2001, pp. 477 – 500.

[16] J. Platt, “Machines using Sequential Minimal Optimization,” in B. Schoelkopf and C. Burges and A. Smola,