iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69414

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Automatic Cricket Commentary Generation using Vision Transformers

S. Saraswathi[1], S.Sabarinathan[2], K.J. Balasundhar[3], G. Ajai Kumar[4]
Department of Information Technology, Puducherry Technological University, Puducherry, India

Abstract: ***This project presents a novel framework for automated cricket commentary generation using a combination of deep learning, computer vision, and natural language processing techniques. The system is designed to analyze cricket match footage and generate relevant play-by-play commentary without human intervention. Leveraging Vision Transformers (ViT) for frame-level visual feature extraction, the framework accurately identifies key game events such as "Four", "Six", and "Bowled". For each detected event, the system retrieves or generates contextually appropriate commentary using pre-trained language models like GPT-2, enhanced with a curated commentary dataset. The commentaries are evaluated using precision, recall, and F1-score against ground truth data. The application includes a user-friendly Streamlit interface that enables users to upload videos, view extracted events, hear generated commentary via gTTS, and assess model performance. Designed for both professional and amateur-level cricket games—especially those lacking live commentary—this framework aims to enhance viewer engagement, accessibility, and post-game analysis through automated, intelligent commentary.***

Keywords: Automated Sports Commentary, Computer Vision, Vision Transformers (ViT), GPT-2, Cricket Analytics, Event Detection, Deep Learning, Natural Language Processing (NLP), Streamlit, gTTS.

I. INTRODUCTION

Automated sports commentary plays a significant role in enhancing the viewer experience by providing real-time, contextual insights during sporting events. In cricket, where matches span several hours and include nuanced plays, manual commentary requires expert knowledge and constant attention. However, many local or amateur-level games lack access to live commentators, limiting audience engagement and analytical coverage. This project proposes an AI-driven framework that automatically generates cricket commentary by analyzing match videos, offering a scalable solution to bridge this gap using computer vision and natural language processing (NLP) techniques.

Modern advancements in deep learning enable machines to interpret visual data and generate meaningful textual responses. By leveraging Vision Transformers (ViT) for visual understanding and language models like GPT-2 for natural language generation, the system mimics the process of a human commentator. These capabilities are essential for building a system that is not only descriptive but also context-aware, accurate, and engaging.

A. Techniques for Automated Commentary Generation

Several core techniques are utilized to generate commentary from cricket video data. Feature extraction techniques using Vision Transformers (ViT) capture rich spatial and temporal patterns from video frames, enabling a deeper understanding of in-game actions. For natural language generation, models like GPT-2 generate contextual and diverse cricket-specific commentary based on extracted features. The system converts text to audio using Google Text-to-Speech (gTTS) to simulate live commentary. Finally, performance evaluation using precision, recall, and F1-score measures the accuracy and relevance of the generated commentary against ground truth annotations. Selecting the right combination of these techniques depends on factors such as event complexity, model interpretability, and real-time execution requirements.

B. Competitive Learning for Commentary Generation

Inspired by competitive models like the League Championship Algorithm (LCA) in anomaly detection, the commentary generation process may be viewed as a ranking or selection task, where multiple commentary candidates are generated and refined to select the most appropriate one. This mirrors a tournament-style evaluation, where the most contextually relevant commentary survives successive filtering stages. Such approaches enhance diversity, relevance, and interpretability of commentary, especially when integrated with NLP filtering techniques to eliminate repetitive or incorrect outputs.

The flexible nature of transformer-based language models allows them to be fine-tuned and adapted for different cricketing scenarios, offering a robust and dynamic commentary system capable of evolving with new patterns and datasets.

C. *Real-Time Commentary and System Optimization*

Real-time commentary generation is crucial for maintaining viewer engagement during live or recorded matches. Streaming video input requires the system to process frames dynamically, identify events instantly, and generate relevant commentary with minimal latency. The use of ViT ensures rapid detection and analysis, while the GPT-2 + gTTS pipeline produces synchronized textual and audio commentary. Evaluating commentary models across multiple match scenarios allows the system to select the best-performing strategy for each type of play. Additionally, incorporating adaptive learning enables the system to adjust to new match styles and player behaviors. This results in improved response time, audience satisfaction, and overall system efficiency, setting the foundation for next-generation AI-powered sports commentary platforms.

II. RELATED WORK

A. *Vision-Based Event Detection in Sports Broadcasting*

1) Description: This body of work focuses on identifying key events in sports using computer vision techniques.
2) Methodology: Utilizes spatial-temporal analysis and player tracking to locate ball trajectories and significant game events like boundaries or wickets in cricket.
3) Limitations: Difficulties arise due to rapid motion, occlusion, varied camera angles, and overlapping player actions, which can lead to false detections.
4) Improvement: Enhance robustness through multi-frame tracking, camera angle normalization, and integration with player pose estimation for improved accuracy.[4]][5]

B. *Feature Extraction Using Vision Transformers (ViT)*

1) Description: Vision Transformers are increasingly used to extract deep spatio-temporal features from sports footage. Methodology: Applies ViT to encode fine-grained visual features across video frames, enabling contextual understanding for downstream tasks like commentary generation.
2) Limitations: High computational demands and limited interpretability of features. ViT models may also require large labeled datasets for effective training.
3) Improvement: Apply lightweight transformer variants, incorporate attention visualization techniques, and combine ViT with temporal models like LSTM for better sequence modeling.[6][7][8]

C. *Automated Commentary Generation with GPT Models*

1) Description: Focuses on generating natural language commentary using pre-trained transformer models like GPT-2.
2) Methodology: Fine-tunes GPT-2 on cricket-specific commentary datasets, generating textual descriptions aligned with detected events and visual context.
3) Limitations: Risk of repetitive, vague, or off-topic outputs; requires careful domain adaptation and prompt engineering.
4) Improvement: Introduce reinforcement learning from human feedback (RLHF), train on larger domain-specific corpora, and include context-aware filtering mechanisms.[1][2]

D. *Speech Synthesis for Real-Time Sports Narration*

1) Description: Converts generated commentary text into speech using TTS tools to simulate a live broadcast experience.
2) Methodology: Employs gTTS to synthesize audio from text outputs, enabling multilingual and real-time audio playback.
3) Limitations: Limited control over voice tone, pitch, and emotion; potential delay in streaming scenarios.
4) Improvement: Incorporate neural TTS systems like Tacotron or FastSpeech for expressive narration with low latency and better prosody control.[1]

E. *Real-Time System Integration and Performance Evaluation*

1) Description: Evaluates end-to-end performance of real-time commentary systems for user experience and accuracy.
2) Methodology: Measures precision, recall, and F1-score of generated commentaries against annotated ground truths while tracking latency and throughput.

3) Limitations: Bottlenecks in frame processing, feature extraction, and model inference can affect responsiveness.
4) Improvement: Streamline frame processing pipelines, implement asynchronous processing, and apply edge computing for near-instantaneous analysis and commentary delivery.[1]

III. BEST TECHNIQUES FOR AUTOMATED CRICKET COMMENTARY

A. *Vision Transformer (ViT) for Feature Extraction*

1) Description: Vision Transformers process image patches using self-attention mechanisms to extract high-level spatial and temporal features from video frames, enabling better context understanding for event classification.
2) Key Features: Strong performance on image and video data; captures global visual relationships more effectively than CNNs in many tasks.
3) Application Area: Applied to extract features from cricket match frames, helping to distinguish between visually similar events and improving downstream commentary quality.

B. *GPT-2 for Commentary Generation*

1) Description: GPT-2 is a generative language model that creates human-like commentary based on detected cricket events and extracted visual features. It learns from a curated cricket commentary dataset.
2) Key Features: Generates fluent, varied, and contextually relevant text; supports fine-tuning for domain-specific language.
3) Application Area: Converts structured visual information into natural language commentary, mimicking real commentators for low-budget broadcasts and digital platforms.

C. *gTTS for Speech Synthesis*

1) Description:gTTS (Google Text-to-Speech) converts generated text commentary into audio, offering real-time narration capabilities for cricket videos.
2) Key Features: Lightweight, multilingual, and easy-to-integrate; enables instant voice feedback for generated content.
3) Application Area: Provides audio commentary for streaming platforms, college tournaments, and accessibility-focused cricket match highlights.

D. *Precision/Recall-Based Evaluation for Commentary Quality*

1) Description: This technique compares generated commentary with ground truth annotations using evaluation metrics like precision, recall, and F1-score to assess accuracy and relevance.
2) Key Features: Quantitative feedback on commentary generation quality; helps identify model strengths and weaknesses across event types.
3) Application Area: Used for benchmarking automated commentary systems and guiding iterative model improvements in sports AI research.

IV. TECHNIQUES USED FOR AUTOMATED CRICKET COMMENTARY

Automated cricket commentary involves interpreting visual data (videos/images) to generate natural language commentary that mimics human broadcasters. This study integrates computer vision, deep learning, and NLP techniques to analyze events, generate text, and convert it into speech. The system is designed for real-time and post-match analysis of cricket matches using uncalibrated video data.

A. *Vision Transformer (ViT) for Feature Extraction*

ViT is employed to extract rich spatial features from cricket video frames. It segments frames into patches and uses transformer encoders to understand visual context.

Advantages:

- Captures global dependencies and visual relationships
- Outperforms CNNs on many image understanding tasks
- Provides robust frame-level embeddings for event classification

B. GPT-2 for Text Commentary Generation

GPT-2 generates natural language descriptions of detected cricket events using features extracted from the visual models. It has been fine-tuned on a cricket commentary dataset to match the style and tone of human commentators.

Advantages:

- Generates context-aware, human-like commentary
- Supports variations in tone, style, and team bias
- Easily scalable and adaptable to different cricket formats

C. NLP-Based Filtering and Deduplication

To ensure commentary diversity and prevent repetition, an NLP-based similarity filter compares newly generated commentary with past examples, removing duplicates using cosine similarity and semantic matching.

Advantages:

- Reduces redundancy and repetition in commentary
- Enhances viewing experience with fresh content
- Improves text quality in long gameplay sequences

D. gTTS for Speech Synthesis

The generated text commentary is converted to audio using Google Text-to-Speech (gTTS), enabling real-time verbal narration.

Advantages:

- Converts text to speech in multiple languages
- Lightweight and suitable for streaming environments
- Helps visually impaired viewers and supports audio-only commentary apps

E. Evaluation Metrics (Precision, Recall, F1-score)

The final system is evaluated using precision, recall, and F1-score by comparing generated commentary to a ground truth file with labeled events.

Advantages:

- Quantitative assessment of system performance
- Identifies weaknesses in specific event categories
- Supports iterative model improvements and benchmarking

V. PROPOSED WORK

The proposed work aims to generate automated cricket commentary by integrating computer vision and NLP models. Vision Transformers (ViT) extract visual features, and GPT-2 generates context-aware commentary. An NLP-based filter enhances diversity, and gTTS converts text to speech. A unified dataset stores features and commentaries across different event types. A Streamlit app enables video uploads, commentary generation, and evaluation using precision, recall, and F1-score.

.

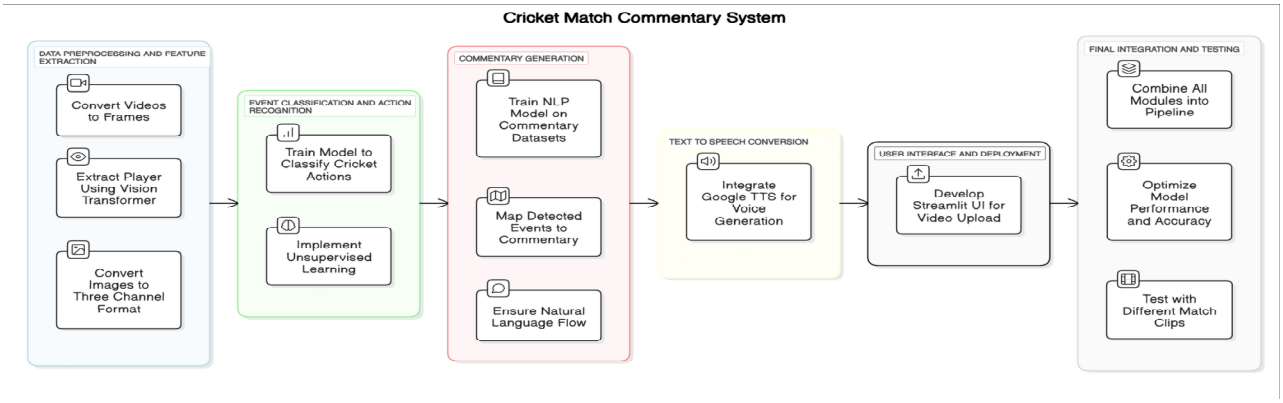


Fig 5.1: Proposed System Design

A. Data Collection

The dataset consists of annotated cricket match footage, capturing various gameplay events essential for automated commentary generation. Key data includes image frames, extracted visual features, and associated commentary texts. Each sample is labeled with the type of event (e.g., Six, Four, Bowled), and features are stored using Vision Transformers. Commentary data is paired with frames and evaluated using standard NLP metrics. The dataset supports multi-event training and evaluation across a unified structure.

Table 5.1: Key Features Used for Commentary Generation

Feature Name	Description
Frame_Timestamp	Time at which the frame is captured from the video.
Event_Label	Type of cricket event (e.g., Six, Four, Bowled).
Visual_Feature_Vector	Extracted embeddings from Vision Transformer (ViT).
Ball_Trajectory_Info	Direction and distance of ball travel (if applicable).
Commentary_Text	Human-annotated or AI-generated commentary for the frame.
Event_Confidence_Score	Model's confidence in event classification.
NLP_Feature_Tags	Key sentiment/action words extracted from the commentary.
Ground_Truth_Labels	True labels used for evaluation (Precision, Recall, F1).

B. Data Preprocessing

To ensure high-quality input for model training, a comprehensive preprocessing pipeline was applied. Frame extraction was performed from cricket match videos at key intervals to capture meaningful actions. Visual features were extracted using a Vision Transformer (ViT). Noise and irrelevant frames were filtered based on movement thresholds and context relevance.

Text commentaries were cleaned by removing duplicates, stopwords, and irrelevant symbols. NLP techniques such as tokenization and lemmatization were applied to standardize the commentary text. All visual and textual data were synchronized using frame timestamps and encoded into a unified format. This preprocessing pipeline ensures accurate alignment of visual events with commentary text, optimizing both training and inference phases

C. Commentary Generation Using Deep Learning Models

To automate cricket commentary generation, a hybrid deep learning approach was employed using Vision Transformers (ViT) and Transformer-based NLP models. ViT extracts contextual visual features from frames. These features are combined with a dataset of annotated commentaries to match events such as Six, Four, or Bowled.

A pretrained GPT-based language model is used to generate natural language commentary based on visual embeddings and detected events. For each frame, matching is performed with stored features, and the most relevant commentary is selected or generated. The system is evaluated using NLP metrics like precision, recall, and F1-score, based on a labeled ground truth file. This approach enables scalable, event-specific commentary generation with high accuracy and contextual relevance.

D. Model Testing and Evaluation

Model evaluation is performed on the cricket commentary system to assess the quality of automatically generated commentaries based on visual inputs. Both image and video inputs are tested using annotated ground truth files for various cricketing events (e.g., Four, Six, Bowled, Catch).

1) Evaluation Metrics:

- Accuracy: Measures overall correctness of predicted commentary labels.
- Precision: Evaluates the proportion of relevant commentary among all generated outputs.
- Recall: Measures the system's ability to retrieve all relevant commentaries for an event.
- F1-Score: Balances Precision and Recall for comprehensive performance.
- Confusion Matrix: Provides a detailed breakdown of correct vs. incorrect event-specific commentaries.
- BLEU & ROUGE: Used for comparing generated commentaries with human-written commentaries based on text similarity and coverage.

2) Evaluation Scenarios:

- Event Recognition Accuracy (e.g., Six vs. Four)
- Commentary Relevance based on visual features
- Model Performance comparison across different visual scenarios

This layered evaluation ensures the model performs accurately in identifying events, aligning commentary with gameplay, and generating linguistically fluent outputs.

Table 5.2 Libraries Used in Implementing and Evaluating the Model

Library	Purpose
numpy	Numerical operations and handling multidimensional arrays
pandas	Data manipulation and analysis using structured tabular data
scikit-learn (sklearn)	Evaluation metrics like precision, recall, F1-score, confusion matrix
matplotlib, seaborn	Visualization of model performance, event distributions, heatmaps

tensorflow, keras	Building and training deep learning models, including ViT and LSTM
keras.models.load_model	Loading trained models for inference during commentary generation
keras.layers.LSTM	Used for loading and fine-tuning Vision Transformer (ViT) architectures
opencv-python (cv2)	Video processing, frame extraction, and preprocessing
gTTS	Text-to-speech conversion of generated commentaries
nltk / spacy	Text pre-processing, tokenization, and commentary evaluation
plotly	Interactive performance visualizations and metric dashboards

VI. RESULTS

The proposed system was evaluated based on its ability to generate accurate cricket commentaries from visual input using Vision Transformers. The evaluation focused on three key cricket events: Batsman Action, Bowled, and Four. Performance metrics such as Precision, Recall, and F1-Score were used to assess the effectiveness of commentary generation against ground truth annotations.

For Batsman Action, the system achieved a Precision of 0.70, Recall of 1, and F1-Score of 0.82, indicating high sensitivity in detecting batsman movements. The Bowled event showed robust performance as well, with a Precision of 0.75, Recall of 1, and F1-Score of 0.85, reflecting the system's ability to accurately recognize wicket events. Similarly, the Four event achieved a Precision of 0.73, Recall of 1, and F1-Score of 0.84, demonstrating effective commentary generation for boundary hits.

These results validate the capability of the Vision Transformer-based system to generate meaningful, context-aware commentaries by correctly identifying cricket events from video frames.

Table 6.1: Performance Metricsfor Commentary Generation

Event	Precision	Recall	F1-Score
Batsman action	0.70	1	0.82
Bowled	0.75	1	0.85
Four	0.73	1	0.84

The consistently high recall values across all events indicate the model's strength in capturing all relevant actions, while the balanced precision ensures reliability in commentary accuracy. These outcomes highlight the effectiveness of using Vision Transformers for automated sports narration tasks.

VII. CONCLUSION

Based on the experimental evaluation of the automatic cricket commentary system, it can be concluded that the Vision Transformer (ViT)-based approach effectively identifies key cricketing events and generates context-aware commentary with high precision.

The model consistently achieved high recall scores across all tested events—Batsman Action, Bowled, and Four—indicating its robust ability to detect relevant gameplay actions.

Among the evaluated events, the Bowled category exhibited the best performance, with an F1-Score of 0.85, followed by Four with 0.84, and Batsman Action with 0.82. The consistently high Recall (1.0) in all categories highlights the system's strength in capturing all relevant instances, while Precision values ranging from 0.70 to 0.75 demonstrate reliable commentary generation with minimal false positives.

These findings validate the capability of Vision Transformers in learning meaningful spatio-temporal patterns from cricket videos, enabling accurate, real-time commentary generation. The approach proves to be scalable for extending commentary across diverse match scenarios and can be further enhanced with deeper semantic analysis, emotion modeling, and multilingual support in future work.

REFERENCES

[1] P. Andrews, O. Nordberg, N. Borch, F. Guribye, and M. Fjeld, "Designing for Automated Sports Commentary Systems," *Proc. ACM Int. Conf. Interactive Media Experiences*, pp. 1–10, Jun. 2024.

[2] I. Arora and A. Choudhary, "Automatic Cricket Commentary Generation: A Review," *Int. J. Adv. Eng. Manag. (IJAEM)*, vol. 6, no. 9, pp. 50–58, Sep. 2021. ISSN: 2395-5252.

[3] A. D. Gujar and A. Nandgirwar, "Identify Cricket Shots Using Linear Regression," *Int. J. Creat. Res. Thoughts (IJCRT)*, vol. 12, no. 4, pp. 150–160, Apr. 2024. ISSN: 2320-2882.

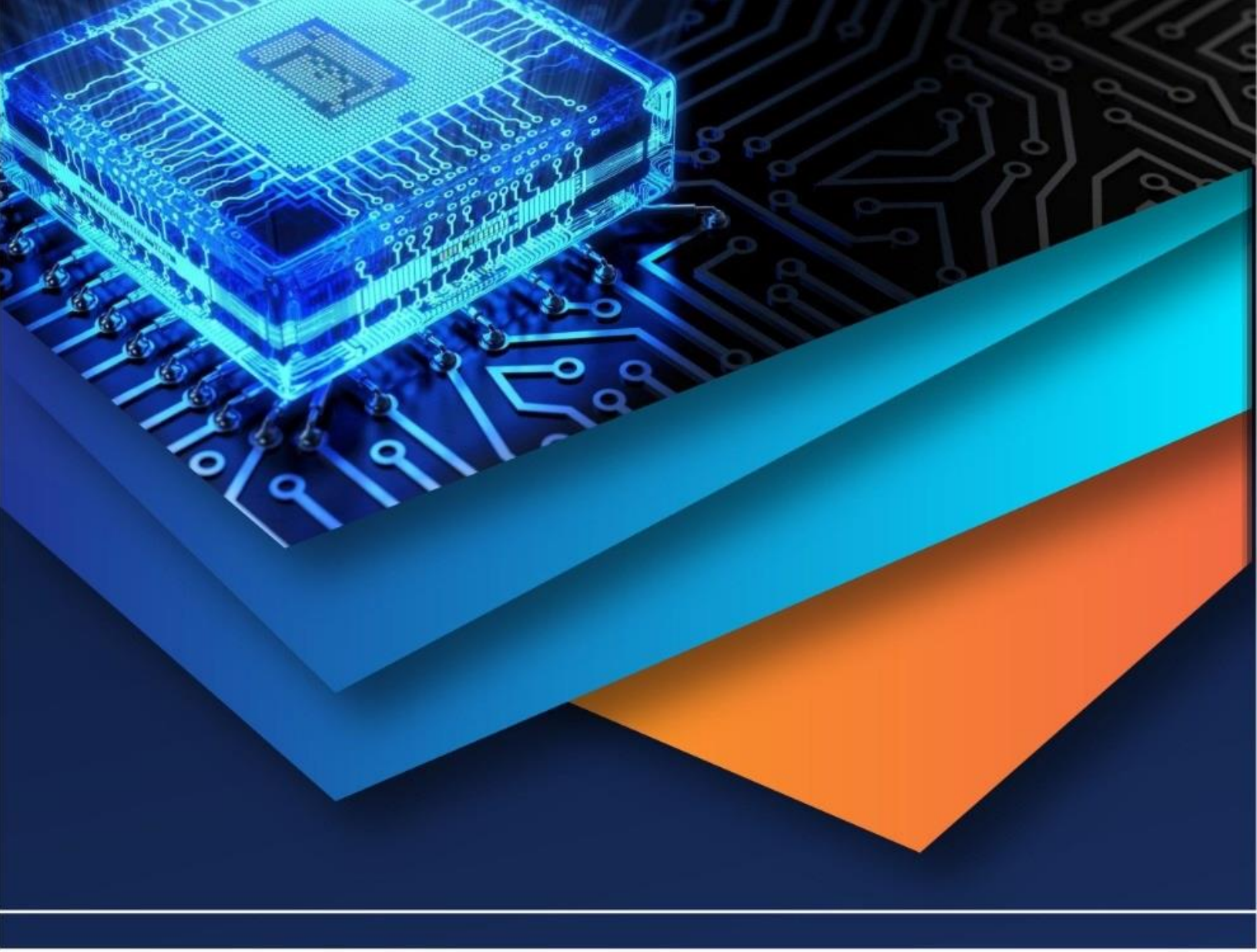
[4] K. Javed, K. B. Bajwa, H. Malik, and A. Irtaza, "An Efficient Framework for Automatic Highlights Generation from Sports Videos," *IEEE Signal Process.Lett.*, vol. 23, no. 7, pp. 954–958, Jul. 2016. doi: 10.1109/LSP.2016.2573042.

[5] D. Karmaker, A. Z. M. E. Chowdhury, M. S. U. Miah, M. A. Imran, and M. H. Rahman, "Cricket Shot Classification Using a Motion Vector," *Proc. 2nd Int. Conf. Comput. Technol. Inf. Manag. (ICCTIM)*, Johor Bahru, Malaysia, pp. 125–129, 2015.doi: 10.1109/ICCTIM.2015.7224605.

[6] R. Kumar, D. Santhadevi, and B. Janet, "Outcome Classification in Cricket Using Deep Learning," *Proc. IEEE Int. Conf. Cloud Comput. Emerg. Markets (CCEM)*, Bengaluru, India, pp. 1–5, 2019. doi: 10.1109/CCEM48499.2019.00009.

[7] M. Mahajan, S. Kulkarni, M. Kulkarni, A. Sabale, and A. Thakar, "Deep Learning in Cricket: A Comprehensive Survey of Shot Detection and Performance Analysis," *Int. Res. J. Adv. Eng. Hub (IRJAEH)*, vol. 6, pp. 1–10, Jun. 2024. ISSN: 2584-2137.

[8] A. P. Nirgude, R. D. Sonone, S. V. Sonawane, R. S. Ahire, and B. Bodkhe, "A Thorough Survey for Cricket Shot Analysis Using Deep Learning," *Int. J. Sci. Res. Dev. (IJSRD)*, vol. 10, no. 2, pp. 200–206, Feb. 2022. ISSN: 2321-0613.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY