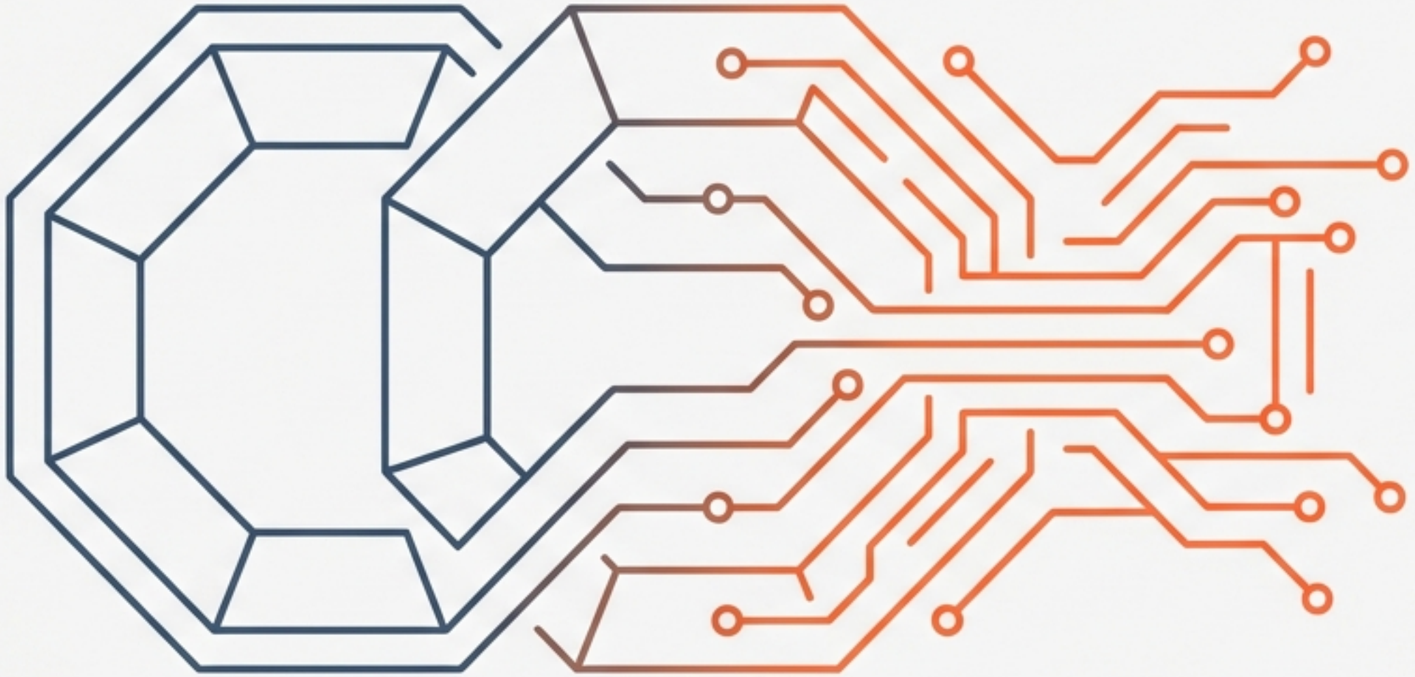
From Octagon to Algorithm

Building the Ultimate Fighting Classifier for Combat Sports Action Recognition



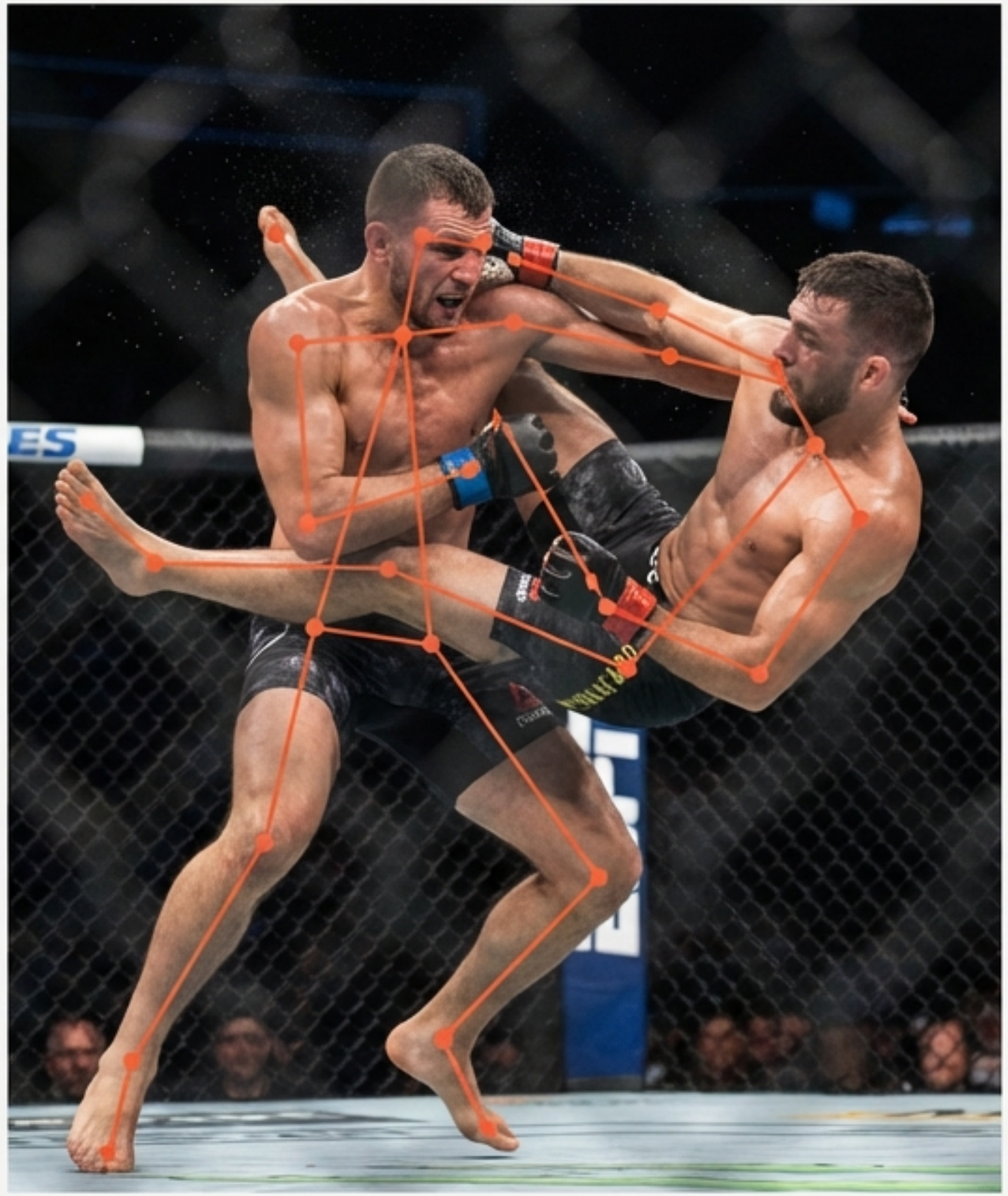
Gary Gau Ku | Master of Engineering - Data Science | University of California, Riverside

NotebookLM

The challenge is classifying motion in a world of snapshots.

Classifying Mixed Martial Arts (MMA) movements is uniquely difficult due to the sport's high speed, variability, and complexity. Traditional analysis often struggles to capture the nuances of these dynamic exchanges.

As a combat sports competitor myself, I recognize the value of an AI system capable of recognizing and analyzing actions in real-time... Such a system could democratize access to high-level coaching... and transform the way MMA is consumed by fans.



NotebookLM

Static analysis sees the fighter, but misses the fight.

STATIC VIEW

DYNAMIC REALITY

CONTEXT

The standard approach applies image classification models to individual video frames. This method has shown promise in sports analytics.

PRIOR RESULT

My own previous research using models like AlexNet on over 4,000 MMA images achieved a validation accuracy of 75%.

THE CRITICAL FLAW

Despite promising results, I hypothesized that the static nature of image classification proved insufficient... the approach proved to be **suboptimal** for classifying MMA techniques due to the dynamic movement patterns inherent in the sport.

NotebookLM

The hypothesis: To understand movement, we must analyze the sequence.

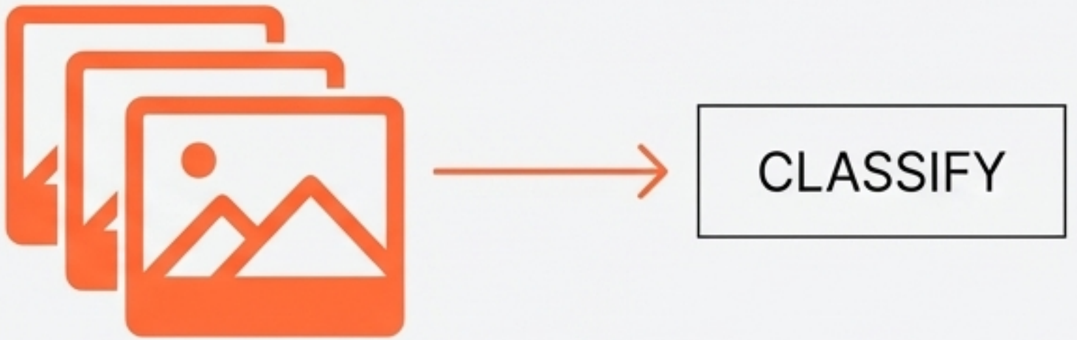
We tested this by building and comparing two distinct models on the same dataset of MMA movements.

The Static Model: Feedforward Neural Network (FNN)

Processes each video frame as an isolated, independent instance. Represents the traditional, “snapshot” approach.

The Temporal Model: Long Short-Term Memory (LSTM) Network

Processes sequences of frames, capturing temporal dependencies. Designed to understand context and the flow of motion over time.



NotebookLM

Incorporating temporal data wasn't just better. It was transformative.

41%

STATIC MODEL (FNN) ACCURACY

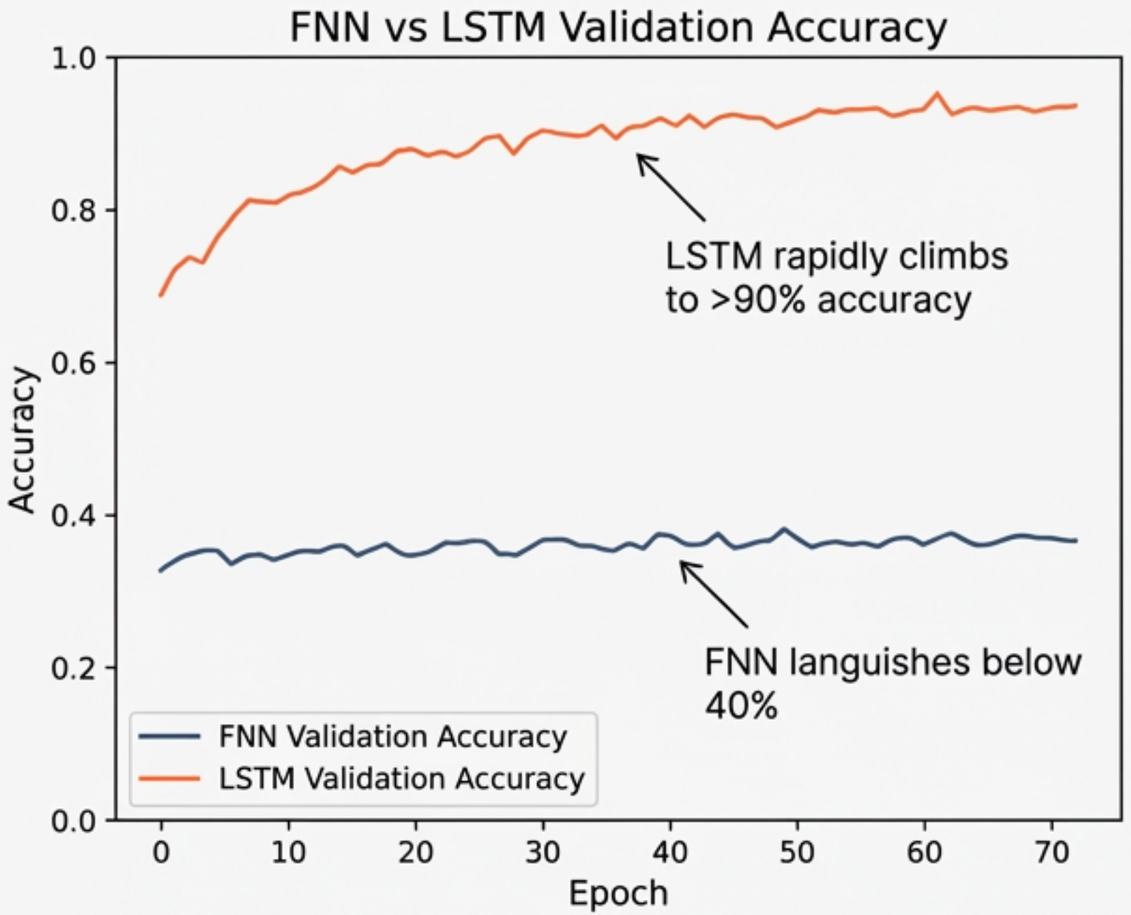
93%

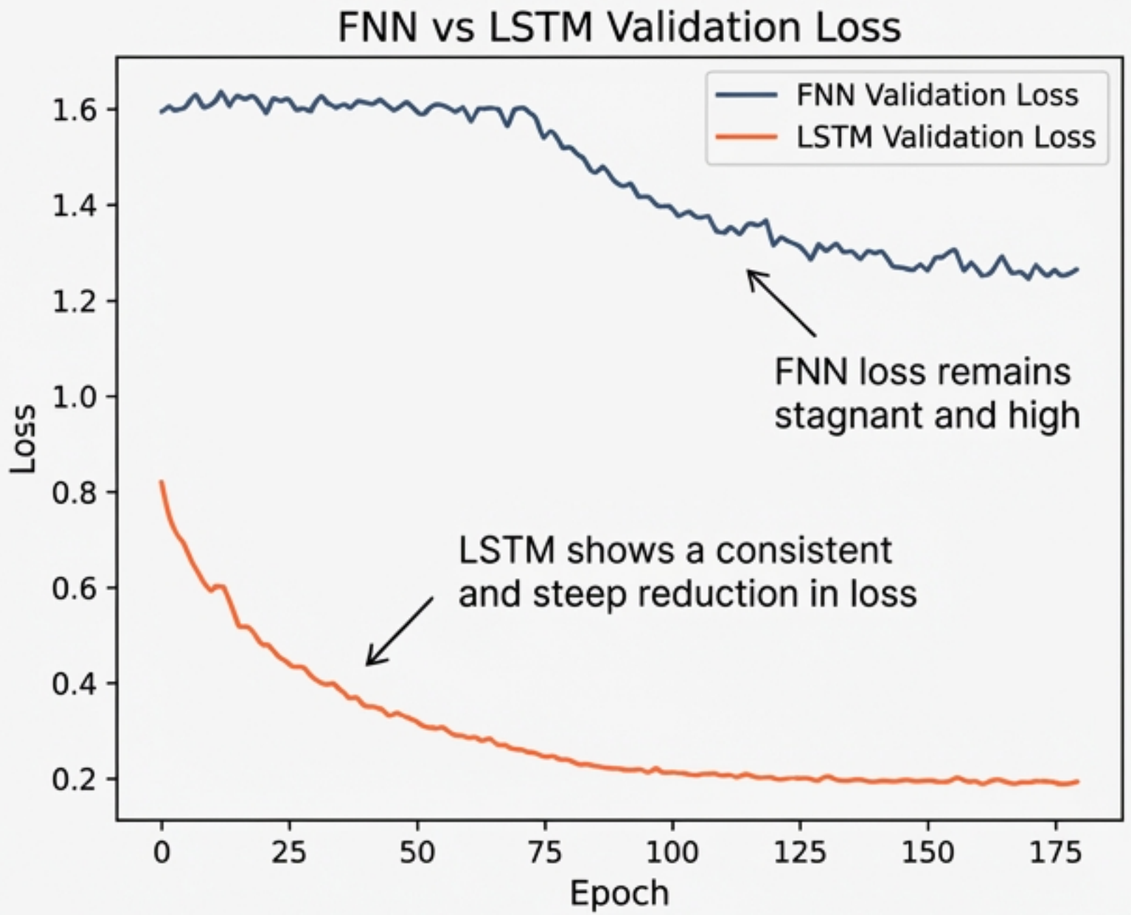
SEQUENTIAL MODEL (LSTM) ACCURACY

"The LSTM model significantly outperformed the FNN, achieving an accuracy of 93% compared to the FNN's 41%. These results validate the hypothesis that incorporating temporal data enhances model accuracy in video-based classification tasks."

NotebookLM

The learning curves tell the complete story.





NotebookLM

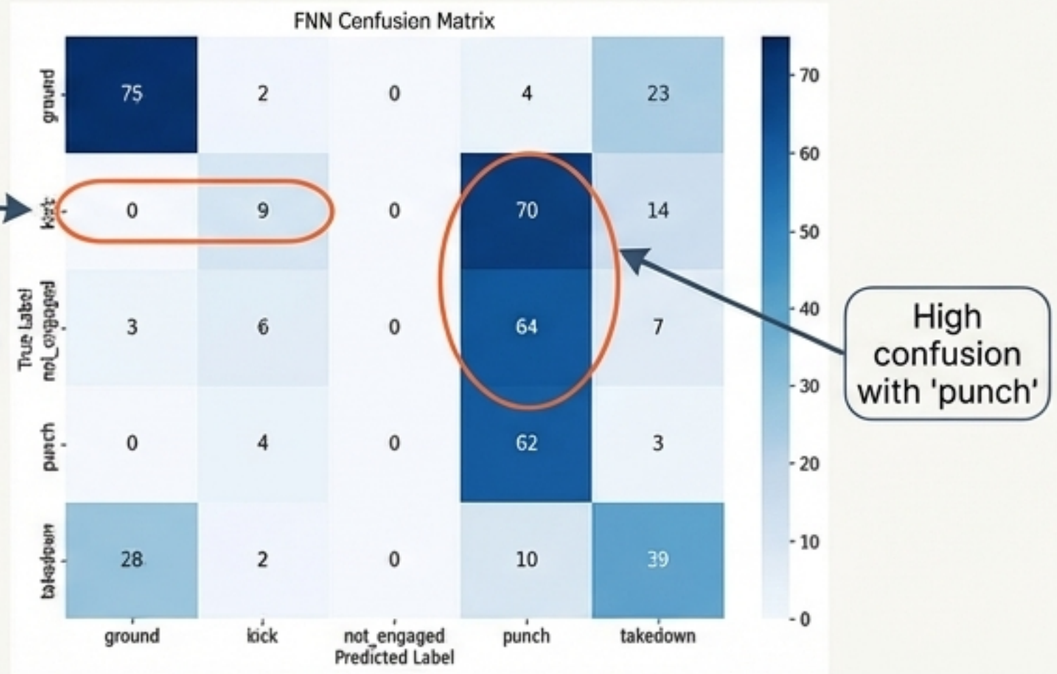
Deconstructing the results: Where the static model failed and the sequential model succeeded.

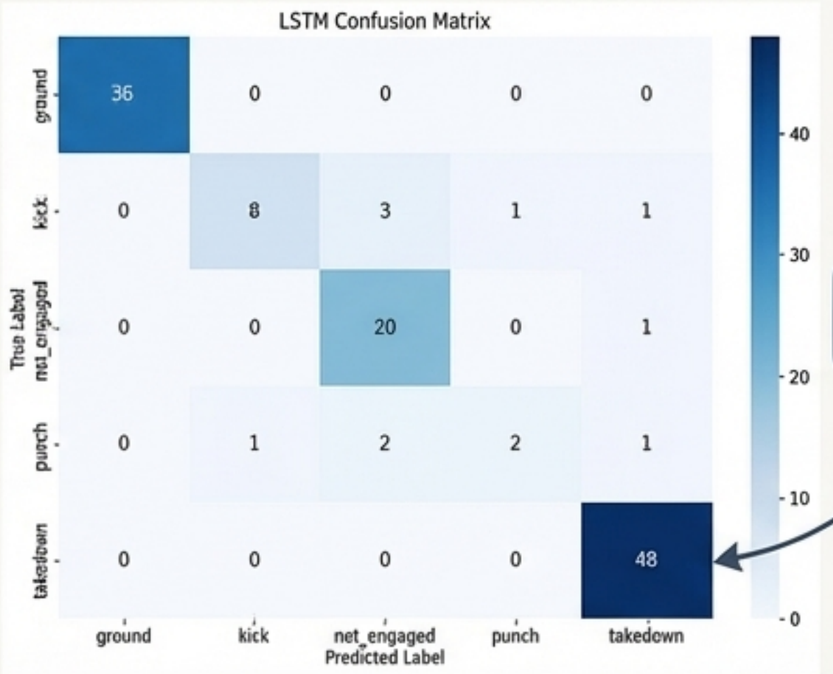
FNN (Static Model): 41% Accuracy

- **Complete failure on 'Kick":** 0 precision and 0 recall. The model could not identify a single kick correctly.
- **High Confusion:** Frequently misclassified 'kick," "not_engaged," and 'punch" as each other. Unable to discern differences within standing positions.

LSTM (Sequential Model): 93% Accuracy

- **Strong Performance:** Excellent precision and recall for 'Takedown' (0.97 F1-score) and 'Not Engaged' (0.87 F1-score).
- **Clear Distinction:** The strong diagonal pattern shows the model effectively distinguishing between all five action classes.





NotebookLM

The architecture of the winning model: From raw video to actionable features.

1. Video Capture

4 HD cameras record MMA sparring in a UFC octagon.

2. Person Detection

The **SSD MobileNet mobiletor** model identifies and creates bounding boxes for each fighter in a frame.

3. Pose Estimation

The **MediaPipe Pose Estimator Estimator** extracts 12 key anatomical landmarks for each fighter.

4. Feature Engineering

Raw coordinates are transformed into 80 features, including reiative distances and joint angles.

5. Model Training

The prepared sequences are used to train the **Bidirectional LSTM network.**

This structured pipeline was essential for converting complex human motion into a format the LSTM model could learn from effectively.

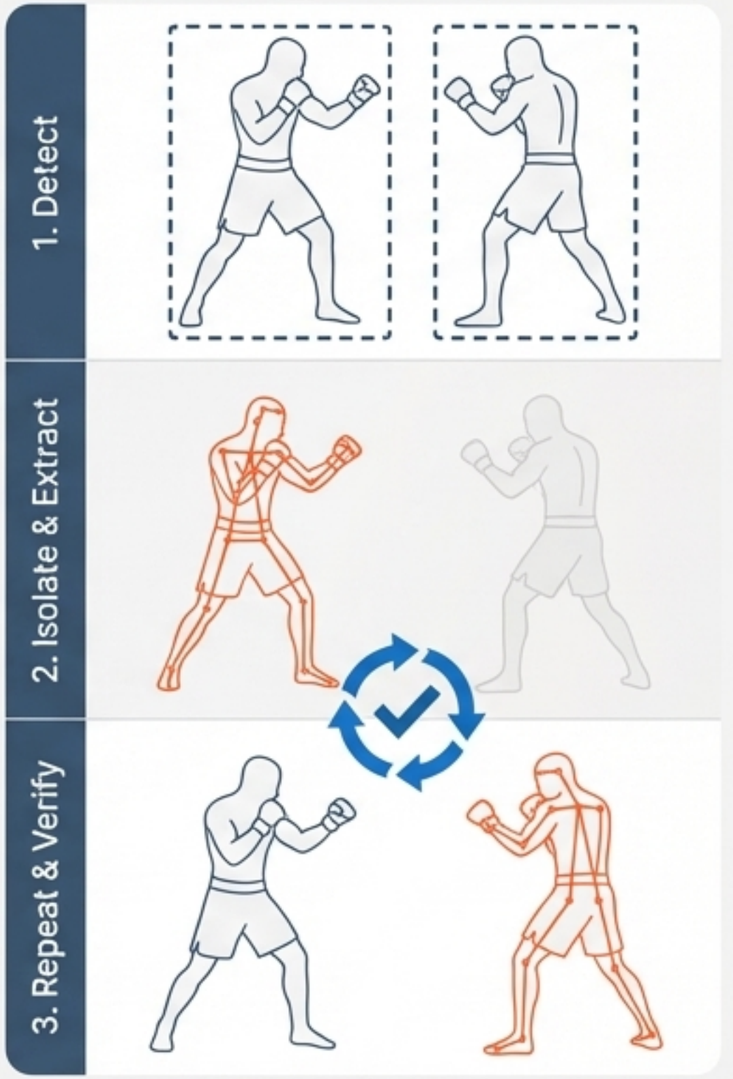
NotebookLM

A key innovation: Adapting a single-person tool for a two-person sport.

The Problem: The MediaPipe Pose Estimator is designed to work on a single person, but MMA involves two interacting fighters.

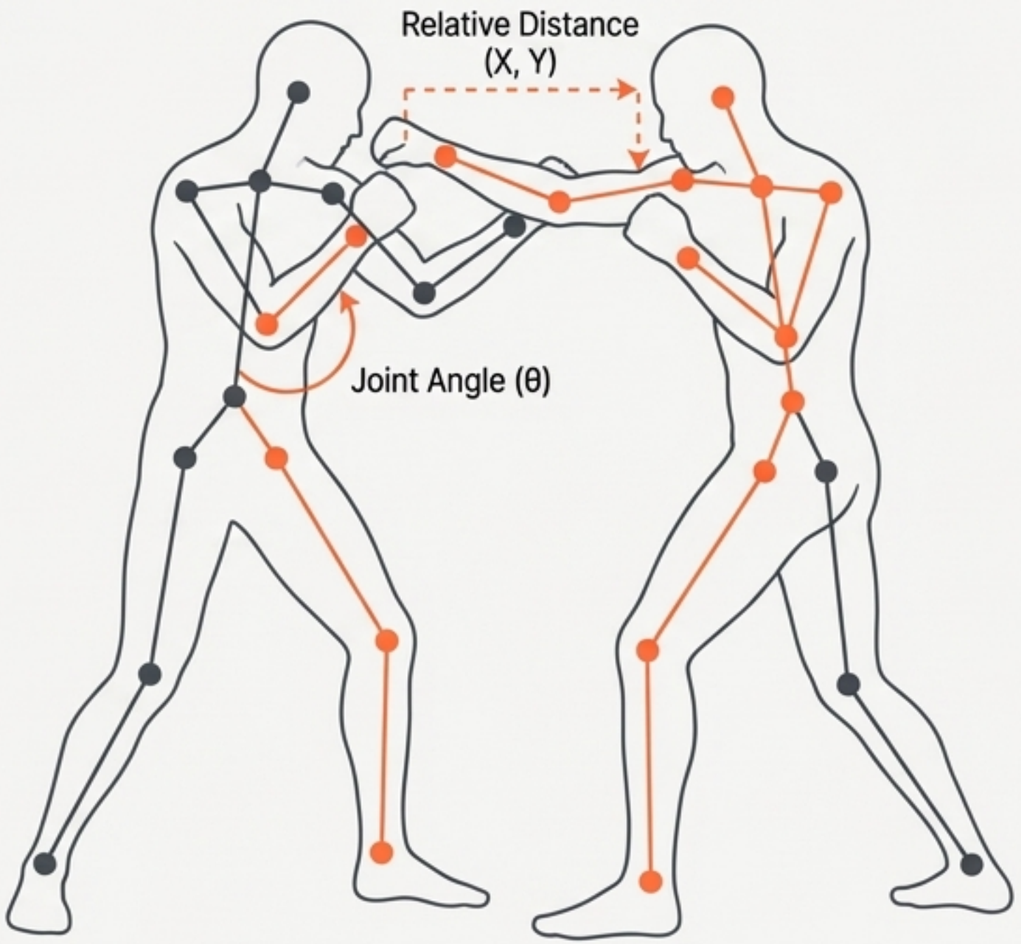
The Solution: A multi-step process was developed to ensure accurate, distinct landmark extraction for both individuals.

1. **Detect**: SSD MobileNet identifies both fighters and creates two separate bounding boxes.
2. **Isolate & Extract**: The first fighter is isolated using their bounding box, and their pose is tracked through the video clip.
3. **Repeat & Verify**: The process is repeated for the second fighter. Crucially, the system performs iterative checks for landmark overlap. If the poses become too similar (e.g., during a grapple), SSD MobileNet is re-run to re-establish distinct bounding boxes.



NotebookLM

Translating motion into math: Engineering features that capture MMA dynamics.



Raw Data

For each fighter, the X/Y coordinates of 12 key anatomical landmarks were extracted per frame (shoulders, elbows, wrists, hips, knees, ankles).

Engineered Features

To provide richer context, these coordinates were used to calculate:

- **Relative Distances**: Horizontal (X) and vertical (Y) distances between corresponding landmarks of the two fighters. (e.g., distance between Fighter A's wrist and Fighter B's shoulder).
- **Joint Angles**: Angles of the elbows, shoulders, hips, and knees for one fighter to capture limb extension and posture.

Final Feature Set: A total of **80 pose features** per frame were generated and normalized before being fed into the models. (Note: A coding oversight limited angle features to one fighter; this transparency is maintained from the source.)

NotebookLM

Creating the ground truth: A curated dataset from an authentic environment.

Data Collection

Environment: A 30-foot diameter UFC Gym Octagon.

Setup: Four high-definition cameras placed 6 feet high, 15 feet from the center.

Preprocessing: Videos downsampled (every other frame) and standardized to 720×720 resolution to balance detail with computational efficiency.

Manual Labeling: All frames were meticulously labeled by the author into five distinct action categories.

Punch: Full motion from extension to retraction.

Kick: Full motion from extension to retraction.

Takedown: Level-change maneuver to ground the opponent.

Ground: One or both fighters are on the ground.

Not_engaged: Fighters are in a neutral, standing stance.

NotebookLM

The next round: From a successful proof-of-concept to a ringside tool.

This research successfully demonstrated a framework for classifying MMA actions with 93% accuracy by leveraging temporal data, adapting single-person pose estimators for combat sports, and validating a pipeline similar to those proposed by Noori et al. (2019) and Cob-Parro et al. (2023).

Future Directions

Enhanced Data

Use higher-definition cameras in a more controlled environment with expert MMA oversight to improve data quality.

Advanced Models

Explore Transformer networks or specialized video models (e.g., UCF101) to handle long-range dependencies even more effectively.

Real-Time Application

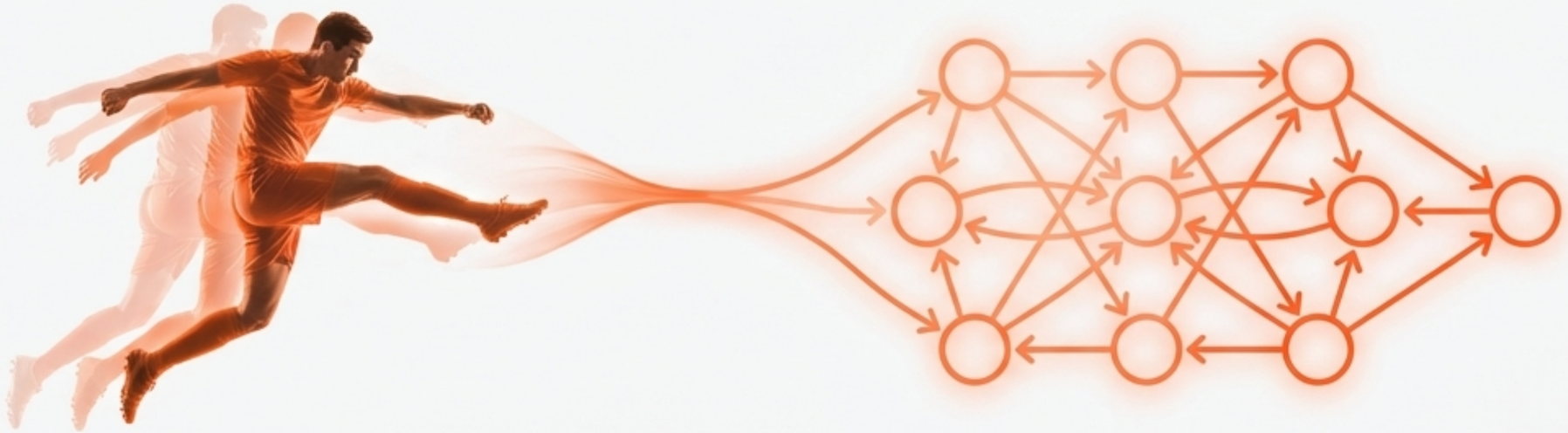
Optimize the pipeline for live analysis, providing immediate feedback to athletes and coaches.

Generalization & Marketability

Adapt the methodology for other sports and develop it into a marketable tool for professional fight promotions and broadcasters.

NotebookLM

The story is in the motion, not the moment.



Core Insight: For dynamic action recognition in sports, analyzing the sequence of movements is not just an improvement—it is essential. A single frame is a letter; the sequence is the word.

Final Statement: By embracing the temporal nature of athletic performance, we can build analytical tools that see the sport with the same contextual understanding as a human expert.

NotebookLM