From Octagon to Algorithm: A Knockout Approach to Action Recognition in MMA

Gary Gau Ku
Master of Engineering - Data Science
Active Combat Sports Competitor

The Challenge: MMA is too fast, complex, and dynamic for simple image classification.

- **High Speed:** Actions are explosive and last mere fractions of a second.
- **Variability:** Techniques have infinite variations in execution.
- **Complexity:** Fighters are constantly interacting, occluding, and transitioning between states.

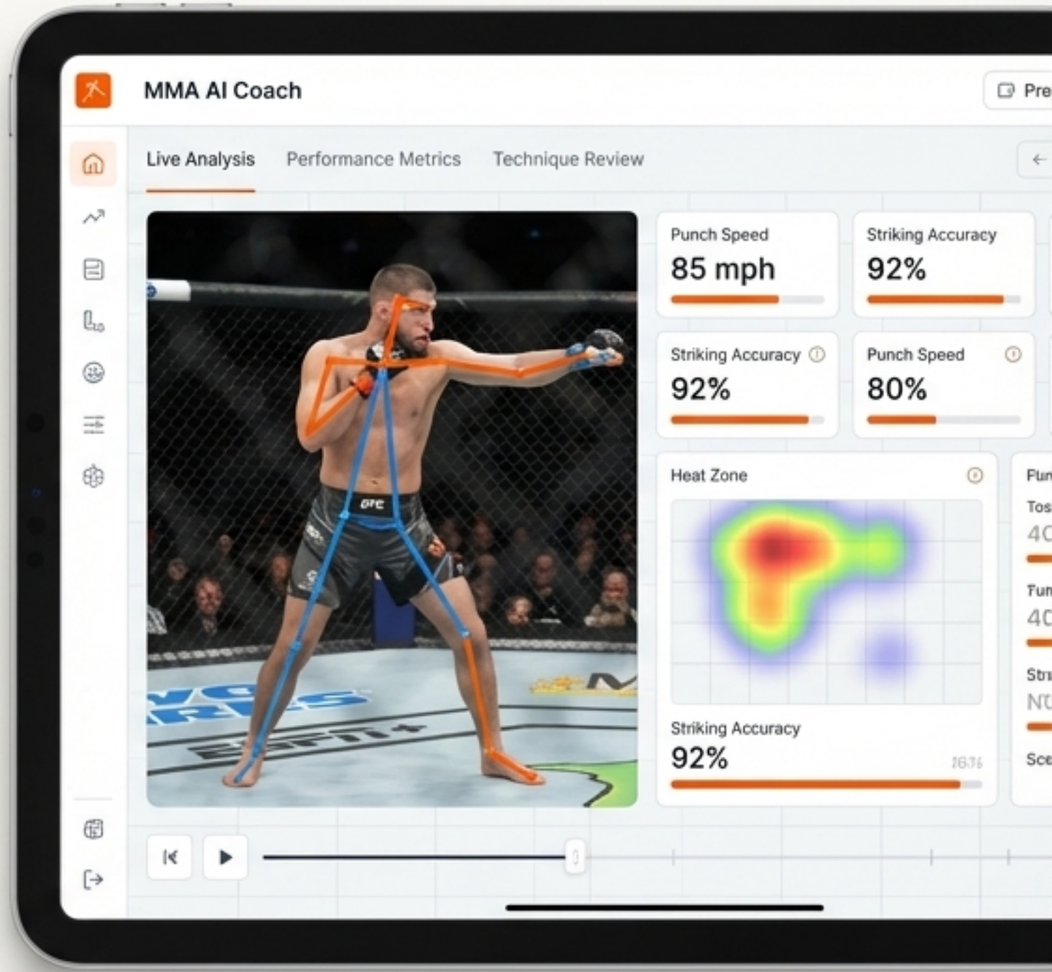
"How can we accurately classify complex, high-speed combat sports techniques from video?"

This research is a blueprint for the future of athletic analysis.

For Athletes: Democratize access to high-level, data-driven coaching, providing insights that were previously unavailable.

For Coaches: Augment expertise with precise, objective performance data, identifying subtle patterns and opportunities for improvement.

For Fans & Broadcasters: Deliver more engaging and informative content, transforming the way MMA is consumed.



Our Technical Playbook: An End-to-End Classification Pipeline

Video Capture

Recording training footage in a controlled environment.

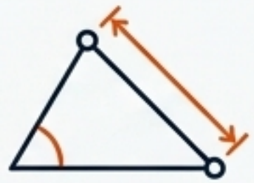
Person Detection

Isolating each fighter using SSD MobileNet.

Pose Estimation

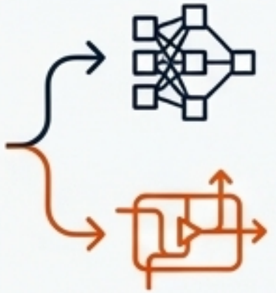
Extracting anatomical landmarks with MediaPipe.

Feature Engineering



Translating landmarks into meaningful data (distances, angles).

Modeling & Classification



Training and comparing our FNN and LSTM models.

Solving the Multi-Person Problem: From Bounding Boxes to Skeletons

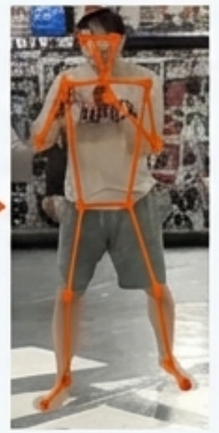
The Challenge: The MediaPipe Pose Estimator is designed for a single person, but MMA involves two interacting fighters.

The Solution: A two-step process to isolate and analyze each fighter.

1. Detect: Use the SSD MobileNet object detector to draw bounding boxes around each fighter.
2. Estimate: Apply the MediaPipe Pose Estimator *within each box*, extracting 12 key anatomical landmarks for each fighter independently.

1. Detect Fighters

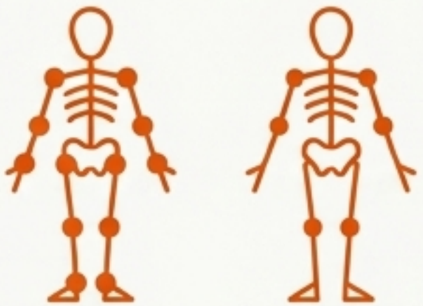
2. Isolate Target



3. Estimate Pose

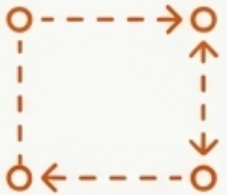
Engineering Features to Capture the Language of Movement

The models don’t see pixels; they see a rich, 80-feature vector for each frame.



Pose Landmarks (48 features)

X/Y coordinates for 12 keypoints on each of the 2 fighters (shoulders, elbows, wrists, hips, knees, ankles).



Relational Distances (24 features)

The horizontal and vertical distances between the corresponding landmarks of the two fighters.

Joint Angles (8 features)

Angles of the elbows, shoulders, hips, and knees to capture posture and limb extension.

Note: A coding oversight resulted in angle features being calculated for only one fighter. The final feature count is 80 per frame.

The First Attempt: A Static, Frame-by-Frame Analysis (FNN)

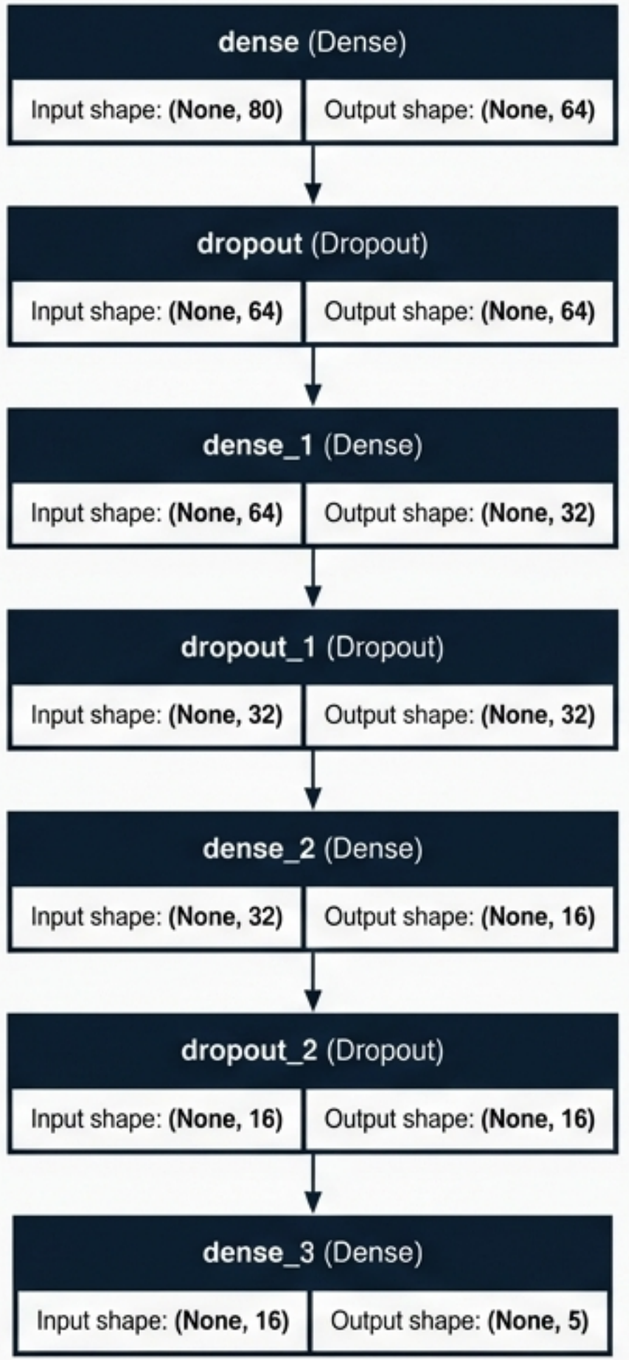
A Feedforward Neural Network (FNN) serves as our baseline. It's a standard, effective architecture for many classification tasks.

Fundamental Limitation

The FNN treats every moment as an independent snapshot. It analyzes each frame in complete isolation, with no memory of what came before or what might come next.

Architecture

- Input: A flattened vector of 80 features per frame.
- Layers: 3 fully-connected dense layers with ReLU activation and Dropout for regularization.
- Output: Softmax layer for 5-class classification.



The FNN Scorecard: Context is Missing

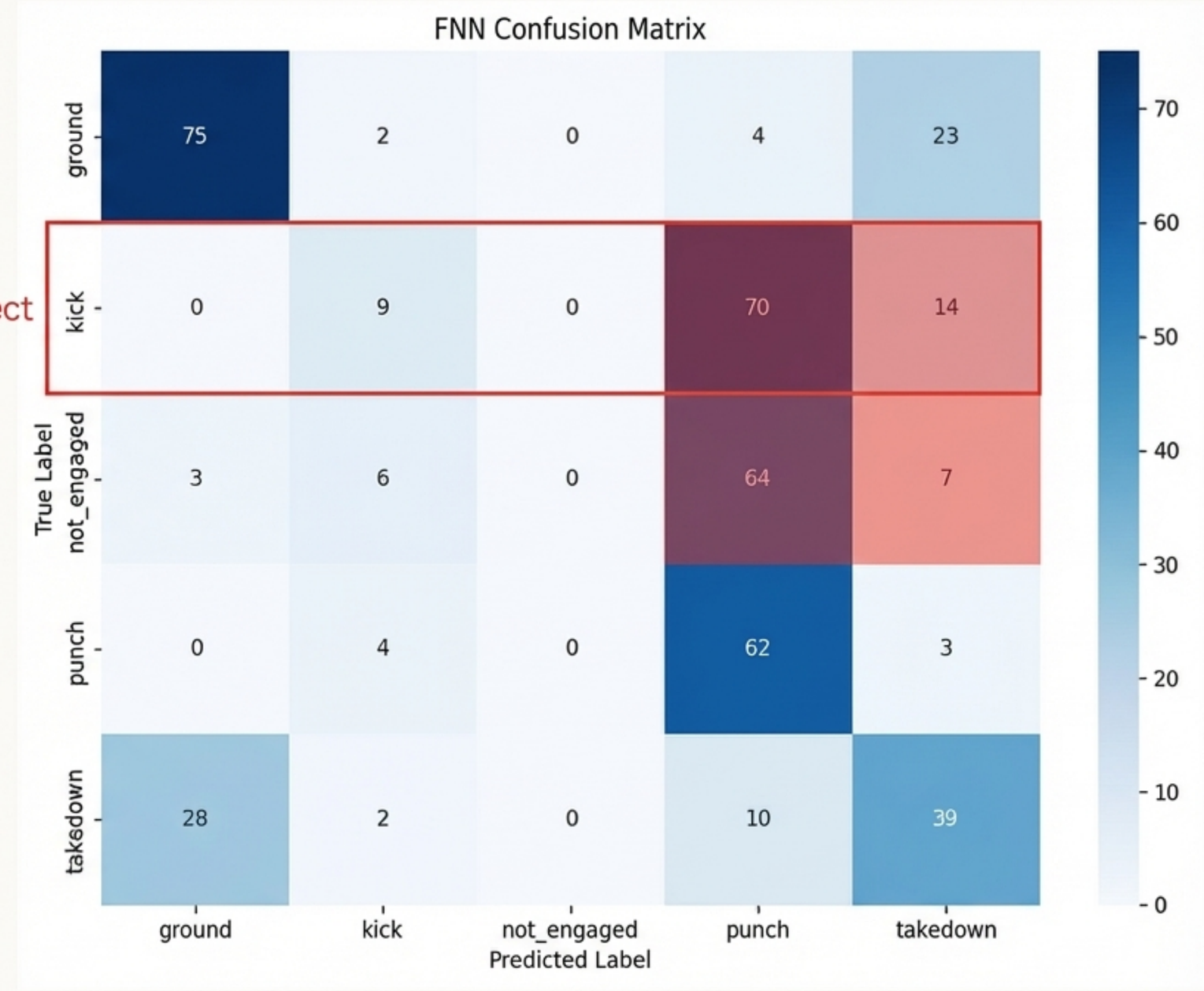
41%

Overall Accuracy

Key Failures

- **Complete Blind Spot:** Fails entirely to identify the 'kick' class (0 precision, 0 recall).
- **High Confusion:** Struggles to distinguish between standing actions, frequently misclassifying 'kick' and 'not_engaged' as 'punch'.

The takeaway: The model can distinguish grappling from standing, but fails at the nuances within those states. A static approach cannot capture the flow of motion.



The Breakthrough: Thinking in Sequence with LSTMs

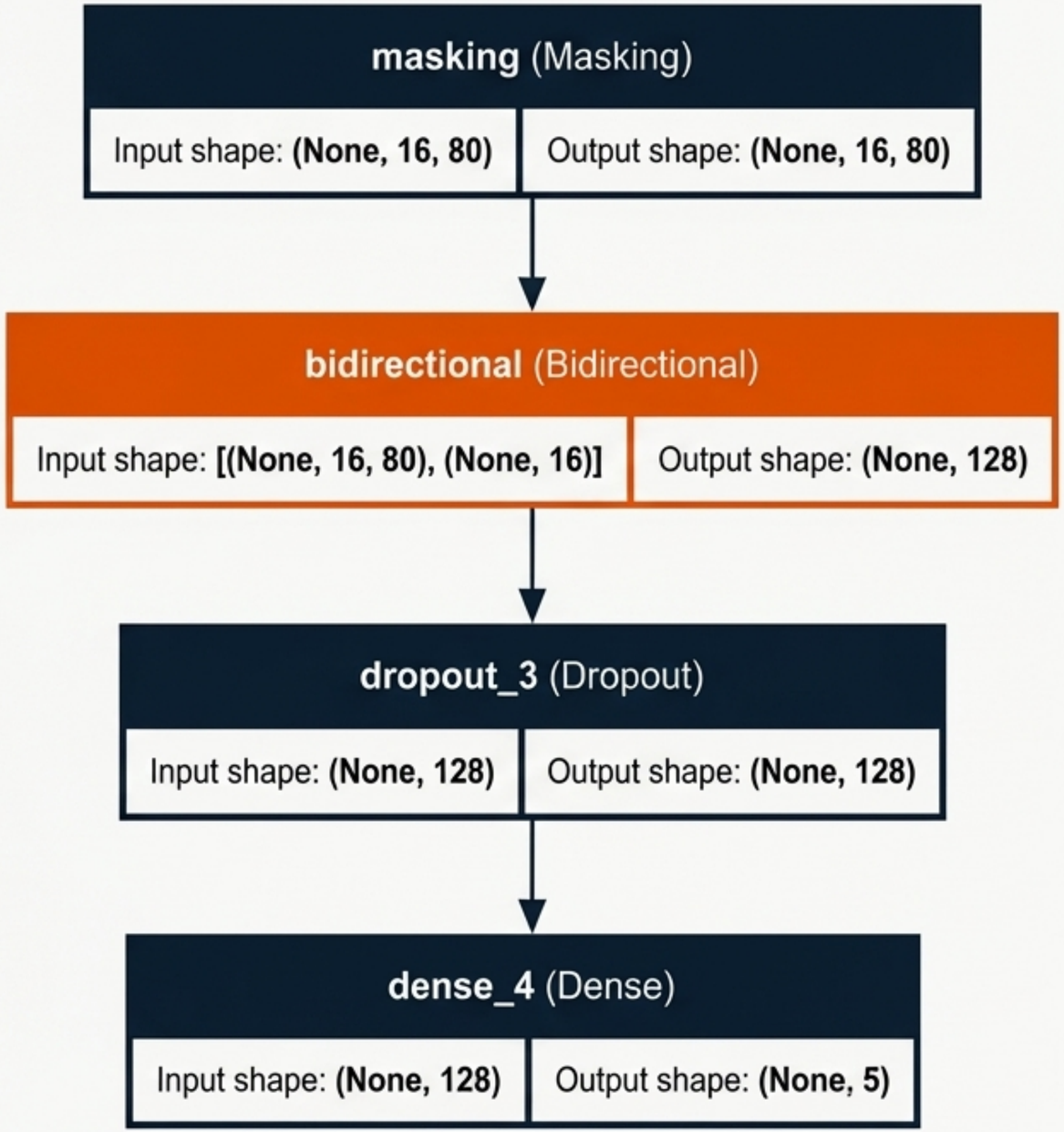
Unlike the FNN which sees static snapshots, the Long Short-Term Memory (LSTM) network 'watches the movie'.

How it Works

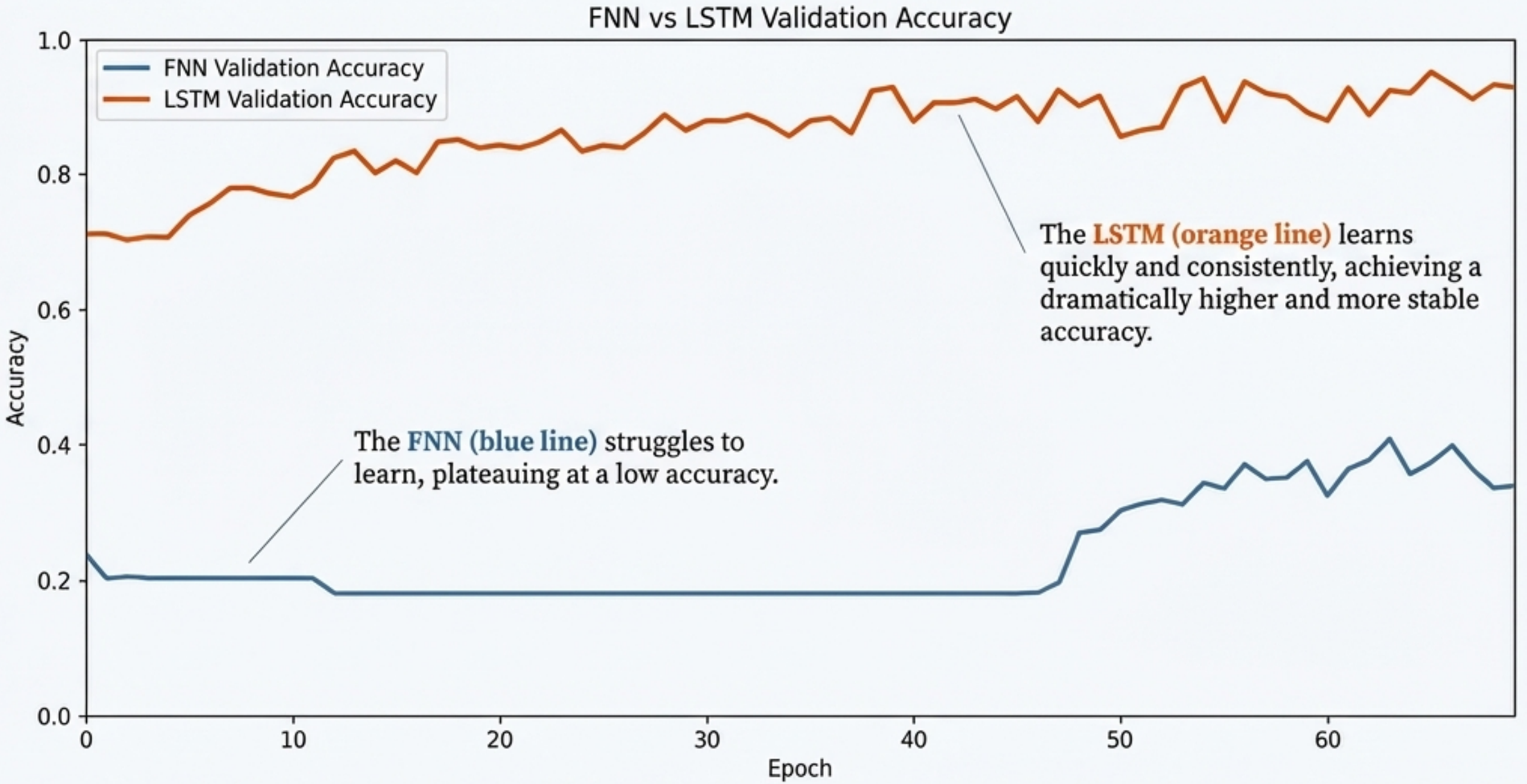
- **Sequential Input:** Processes sequences of 16 consecutive frames at a time.
- **Temporal Memory:** Its architecture is designed to recognize patterns and retain information over time.
- **Bidirectional Analysis:** The model considers both past and future frames in a sequence to better understand the context of an action.

Architecture

A Bidirectional LSTM layer processes sequences of 80 features, followed by a dense output layer.



The Knockout Blow: A Clear Victory for Temporal Analysis



The massive gap between the two models provides definitive proof that understanding sequence is the key to success.

Tale of the Tape: The Numbers Don't Lie

Metric	FNN (Static)	LSTM (Temporal)
Overall Accuracy	**41%**	**93%**
Macro F1-Score	**0.36**	**0.86**

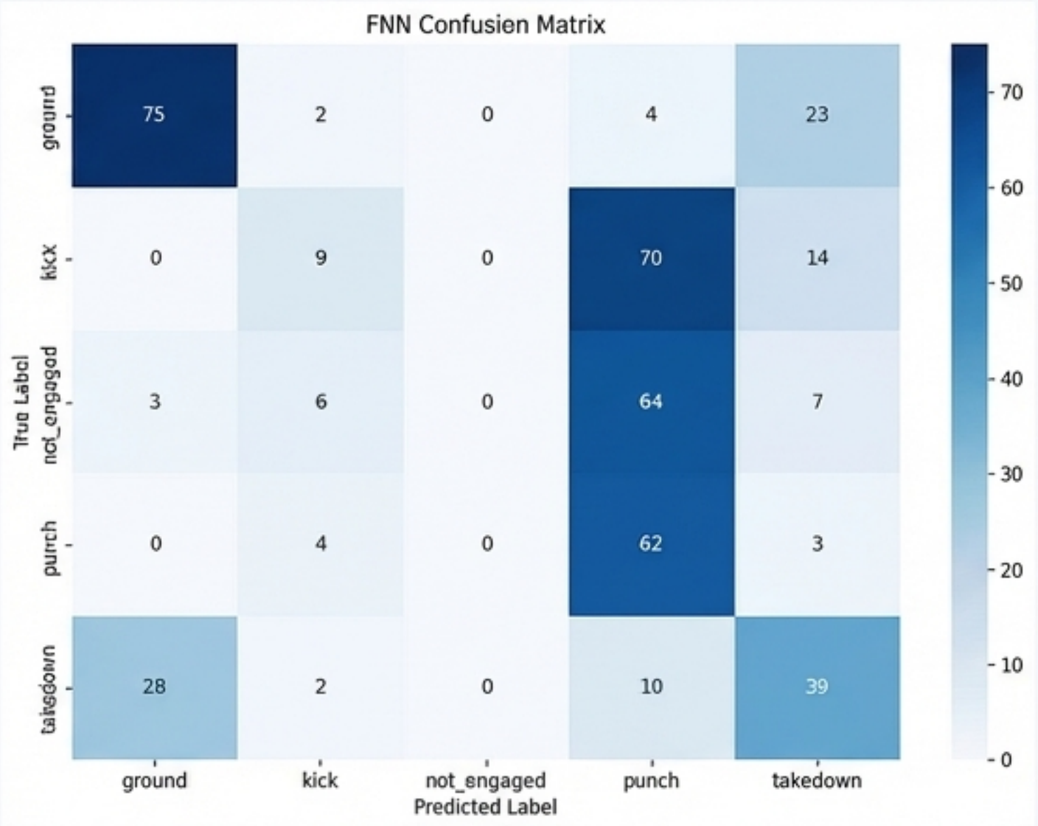
Class-Specific Highlights

- The LSTM successfully classifies the '**kick**' category, which the FNN completely missed.
- The LSTM achieves an F1-score of **0.97** for 'takedown,' demonstrating near-perfect recognition of this complex, multi-frame action.

From Chaos to Clarity: Visualizing Model Performance

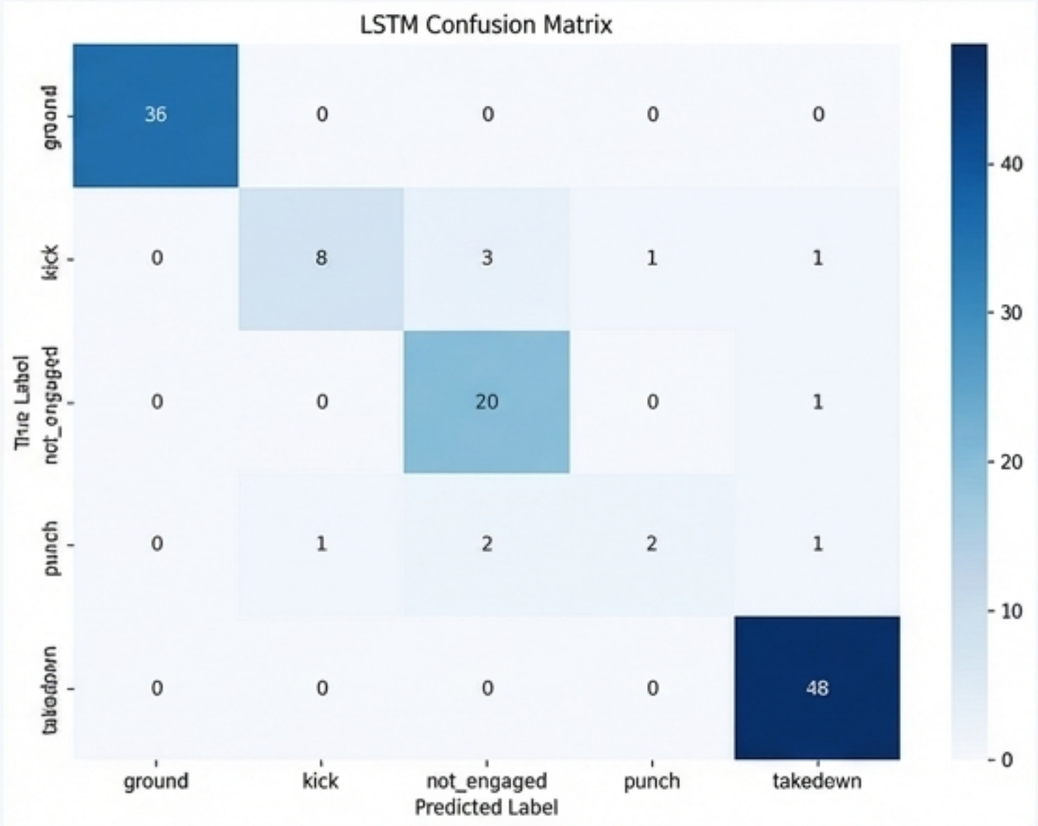
FNN: Static Confusion

The matrix is scattered. Predictions are spread across multiple incorrect classes, especially for standing actions.



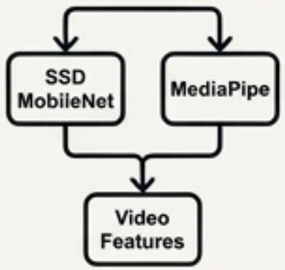
LSTM: Temporal Clarity

The matrix is clean and strongly diagonal. The vast majority of predictions fall on the correct true label.



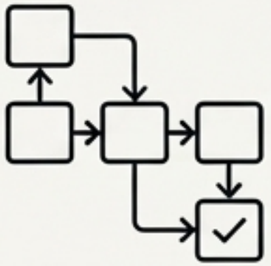
Key Research Contributions

Finding 1: A Validated Framework for Multi-Person Pose Estimation.



We successfully developed and proved a methodology combining an object detector (SSD MobileNet) with a pose estimator (MediaPipe) to extract features from multi-person combat sports videos.

Finding 2: Temporal Models are Essential for Dynamic Action Recognition.



Our results provide definitive evidence that temporal models like LSTMs, which analyze sequences of frames, are vastly superior to static models for classifying complex MMA techniques. The **52-point accuracy jump (41% to 93%)** validates this core hypothesis.

Acknowledging Limitations and Charting the Future

Current Limitations

- **Manual Labeling Bias:** Labeling was conducted by a single person.
- **Controlled Environment:** Data was collected in a gym, not from real-world bouts.
- **Hardware Constraints:** Limited processing power impacted the feature extraction speed and completeness (re: angle features).

Future Directions

- **Advanced Models:** Explore more sophisticated architectures like Transformers for even better long-range dependency modeling.
- **Real-Time Application:** Optimize the pipeline for live analysis and feedback during training.
- **Expand the Domain:** Adapt and test the methodology in other complex, multi-person sports like basketball or soccer.

The First Step on a Longer Journey...

This project provides a robust foundation for automated sports analytics. As models and hardware continue to advance, the ability for AI to serve as a real-time coach, analyst, and sparring partner moves closer to reality.

Gary Gau Ku | gaugarygrayston@gmail.com