Tennis Visual Analysis: Biphasic Integration of Shot Classification and Spatial Object Detection

A report submitted in fulfillment of the requirements for the degree of

Bachelor of Technology

in

Information Technology

Aditya Bharadwaj (2022IMT-008)

Ayush Sah (2022IMT-026)

Manoj Shivagange (2022IMT-070)

Under the Supervision of

Dr. Anjali

Department of Information Technology

ABV-INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT

GWALIOR, INDIA

August 2025

DECLARATION

We hereby certify that the work, which is being presented in the report/thesis, entitled **Tennis Visual Analysis: Biphasic Integration of Shot Classification and Spatial Object Detection**, in fulfillment of the requirement for the award of the degree of **Integrated Post Graduate Master of Technology (IPG-M.Tech)** and submitted to the institution is an authentic record of our own work carried out during the period May 2025 to August 2025 under the supervision of **Dr. Anjali**. We also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

Dated: **Signature of the candidates**

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dated: **Signature of supervisor**

Acknowledgements

We would like to express our heartfelt gratitude to our guide, Dr. Anjali, for her unwavering mentorship and support throughout the duration of this project. Her insightful guidance, willingness to share her knowledge, and solutions to the challenges we faced played a pivotal role in our growth and the successful completion of this work.

We also extend our sincere gratitude to our friends and families, who have been a constant source of support and encouragement during this entire endeavor.

Finally, we are highly indebted to our institution for affording us the opportunity to embark on this project. This experience allowed us to dive deep into the concepts of Data Science and explore a field that was very new to us, for which we are immensely grateful.

Aditya Bharadwaj

Ayush Sah

Manoj Shivagange

Abstract

The increasing integration of computer vision into sports science is revolutionizing performance analysis. In tennis, this provides coaches and athletes with objective insights, yet progress is often hindered by the scarcity of large, publicly available datasets for training robust machine learning models on standard video. This thesis addresses this gap through a comprehensive biphasic approach combining temporal shot classification with spatial object detection and tracking.

Phase 1 develops a novel seven-class tennis shot dataset sourced from diverse, publicly available footage, including fundamental groundstrokes (forehand, backhand), serves, volleys, and a crucial neutral class representing non-shot movements. The system leverages the efficient MoveNet model for 2D human pose estimation, converting video into structured time-series data. A comparative evaluation of multiple deep learning architectures demonstrates that the 1D CNN architecture outperforms recurrent models, achieving a mean classification accuracy of 92.4%.

Phase 2 extends beyond temporal shot classification to encompass comprehensive tennis analysis through advanced object detection, physics-aware tracking, and temporally-consistent court spatial analysis. The enhanced system implements a novel dual-YOLO architecture (YOLOv8 for players, fine-tuned YOLOv5 for balls) achieving 88.7% mAP, physics-informed DeepSORT tracking (78.4% MOTA), and ResNet50-based court detection with LSTM temporal stabilization.

This biphasic approach demonstrates the feasibility of comprehensive automated tennis analysis suitable for coaching, broadcasting, and performance analytics applications. The enhanced integrated system achieves 89.1% overall accuracy across all components while maintaining real-time processing capabilities, establishing computer vision as an essential technology for modern tennis analysis.

Contents

List of Figures

List of Tables

1

Introduction

This chapter introduces the growing role of technology in sports, with a focus on tennis analytics and computer vision. It outlines the motivation for democratizing elite-level performance analysis, articulates the core problem around data accessibility, and presents the objectives, contributions, and structure of this thesis.

1.1 Context

Over the years, the domain of elite sports has seen a significant paradigm shift, moving from a coach-centric subjective approach to an objective, data-driven approach. Technology is no longer an afterthought in sports, but a central component in how games are played, coached, and watched.

At the pinnacle, we have systems like Hawk-Eye, which have become synonymous with modern tennis. Hawk-Eye is a computer vision system that uses a network of high frame-rate cameras to triangulate and extrapolate ball trajectories, providing an impartial second opinion via a challenge system. This system proved the validity of integrating technology into sports and changed how the game is played.[1]

Complementing on-court officiating tools there is the field of data analytics services tailored for athletes and professional teams. Companies such as Mouratoglou Analytics,[2] Tennis Analytics,[3] and TennisViz[4] offer detailed performance analysis by studying match footage and mining as much data as possible. They provide insights such as player tendencies, serve placement percentages, rally length distributions, and more. These insights can offer a subtle competitive edge—often enough to change the trajectory of a player's career. These points highlight the immense value of detailed, shot-level data in gaining competitive advantage.[5]

Assistive technology and the democratization of performance analysis

The fundamental issues with systems such as Hawk-Eye and commercial analytics services are that they are financially and logistically inaccessible to the vast majority of the tennis community, including amateur players, junior athletes, and club-level coaches. In a country like India, even national-level players often find it challenging to afford such technology. This inaccessibility has led to the emergence of affordable, user-friendly solutions. The mobile phones most people carry have evolved to the point where they provide significant computational power, and advances in artificial intelligence have catalysed a

movement toward the democratization of performance analysis.

A prime example of this trend is SwingVision, a mobile application that transforms a smartphone or tablet into an AI-driven analysis tool. SwingVision uses a single court-side camera and provides automated shot tracking, statistical analysis, video highlight generation, and line calling.[6] This application demonstrates both the viability and strong demand for accessible systems that bring elite-level analytics to the grassroots. This project operates within this emerging landscape. We seek to address the core technical challenge—automated shot recognition and analysis from standard video. The first and most challenging hurdle is a basic problem that holds back academic research and independent development: the lack of suitable data.

1.2 Problem Statement and Objectives

1.2.1 Phase 1: Temporal Analysis Challenges

Tennis research and community players lack open datasets comparable to those held by elite companies. Current data is either proprietary, too small, or poorly annotated, making it hard to train models that work outside controlled lab conditions. What is missing is a practical, open dataset that supports shot recognition and analysis on ordinary devices, without requiring costly cameras or special hardware.

1.2.2 Phase 2: Spatial Analysis Requirements

The comprehensive analysis of tennis requires understanding not only what shots are being played, but where players are positioned, ball trajectories, and court utilization patterns. Current systems face several technical challenges:

- **Real-time Object Detection**: Tennis balls represent small, fast-moving objects requiring specialized detection algorithms capable of handling speeds exceeding 150 mph while maintaining frame rates suitable for live analysis.

- **Multi-Player Tracking**: Simultaneous tracking of two players with frequent oc-

clusions, rapid directional changes, and consistent identity maintenance throughout extended rallies.

- **Spatial Calibration**: Accurate court detection and homography transformation enabling precise real-world coordinate mapping from video perspectives.
- **System Integration**: Combining temporal shot classification with spatial analysis in a unified pipeline maintaining real-time performance constraints.

1.2.3 Key System Challenges

- **Data scarcity:** Public tennis datasets are limited, inconsistent, and lack standard shot-level labels for both temporal and spatial analysis.
- **Poor generalisation:** Data from controlled setups does not transfer well to varied courts, lighting, or player levels.
- **Device constraints:** Models must run efficiently on commodity smartphones while handling complex multi-component processing.
- **Privacy:** To protect players, we store only extracted pose/keypoint data, not raw RGB video.
- **Real-time Integration:** Combining multiple computer vision tasks while maintaining processing speeds suitable for live analysis.

1.2.4 Unified System Objectives

- **Develop integrated computer vision pipeline:** Combine shot classification with spatial object detection in a unified system maintaining real-time performance.
- **Implement real-time ball detection:** Create robust ball detection and trajectory tracking using state-of-the-art YOLO architectures optimized for tennis environments.

- **Build multi-player tracking system:** Develop tracking capabilities with speed estimation and movement analysis suitable for tennis dynamics.
- **Design accurate court detection:** Implement court detection with homography transformation enabling precise real-world coordinate mapping.
- **Create comprehensive visualization:** Build tactical minimap and performance analytics combining temporal and spatial intelligence.
- **Validate across diverse conditions:** Test system performance across varied video conditions and playing environments.

1.3 Biphasic System Architecture

This thesis presents a novel biphasic approach to tennis analysis, where Phase 1 provides temporal understanding of player actions while Phase 2 delivers spatial intelligence about object positions and movements.

1.3.1 Phase 1: Temporal Shot Classification Pipeline

Building upon pose estimation and sequence modeling, this phase classifies tennis shots into seven categories using MoveNet for keypoint extraction and 1D CNN for temporal pattern recognition. The system achieves 92.4% weighted F1-score across shot types including forehands, backhands, serves, volleys, and neutral movements.

1.3.2 Phase 2: Spatial Analysis and Object Detection

The spatial analysis component implements three synchronized detection systems:

- **Enhanced Ball Detection**: Dual-YOLO architecture (YOLOv8 for players, fine-tuned YOLOv5 for balls) with physics-aware tracking achieving 88-89% mAP on diverse datasets
- **Player Tracking**: DeepSORT implementation with appearance and motion fusion delivering 80+ HOTA performance

- **Court Analysis**: Keypoint-based court detection with homography transformation enabling real-world coordinate mapping

1.3.3 Integration Architecture

The unified system processes video streams through parallel pipelines with synchronized output, enabling comprehensive analysis combining temporal action understanding with spatial positioning intelligence. The integration occurs at the visualization layer where temporal shot labels combine with spatial positioning data to provide complete tactical analysis.

1.4 Thesis Contribution and Structure

1.4.1 Technical Contributions

This thesis presents several key contributions to computer vision and sports analytics:

- **Novel biphasic architecture:** Integration of temporal shot classification with spatial object detection, tracking, and court analysis in a unified real-time system.
- **Comprehensive training framework:** Custom dataset with 5,000+ ball detection images, 2,791 shot classification sequences, and novel LSTM-based temporal court stabilization for improved consistency.
- **Novel dual-YOLO implementation:** Specialized YOLOv8-YOLOv5 hybrid achieving 88-89% mAP through physics-informed tracking, temporal consistency modeling, and ResNet50-based court detection with LSTM stabilization.
- **Physics-aware tracking system:** Enhanced DeepSORT with parabolic motion models, bounce detection, and court-constrained tracking achieving 73-78% HOTA scores for tennis-specific movement patterns.
- **Comprehensive evaluation framework:** Assessment methodology for both individual components and integrated system performance across diverse playing en-

vironments.

1.4.2 Document Structure

The remainder of this report is structured as follows:

- **Chapter 2: Background and Literature Survey** — Reviews related work across temporal action recognition, object detection, multi-object tracking, and court detection; establishes theoretical foundations for both phases of the integrated system.
- **Chapter 3: Methodology** — Details the biphasic system architecture including Phase 1 temporal classification pipeline, Phase 2 spatial detection components, and integration methodology for unified processing.
- **Chapter 4: Experiments and Results** — Presents comprehensive evaluation including individual component performance, integrated system metrics, and qualitative analysis of real-world deployment scenarios.
- **Chapter 5: Discussions and Conclusions** — Analyzes system performance, discusses limitations and future enhancements, and summarizes contributions to computer vision and sports analytics.

2

Background and Literature Survey

This chapter reviews the comprehensive landscape of automated tennis analysis spanning temporal action recognition and spatial object detection. It validates the motivation for both phases of the integrated system, covering technical foundations including human pose estimation, sequential modeling for action recognition, object detection architectures, multi-object tracking algorithms, and court detection methodologies. The review establishes theoretical foundations and practical trade-offs guiding the biphasic system design.

2.1 The Challenge of Automated Tennis Analysis

2.1.1 The Data Scarcity Problem: Validating the Project Motivation

The success of modern supervised machine learning is fundamentally dependent on the availability of large-scale, high-quality, and accurately labeled datasets. While the field of computer vision has benefited from massive general-purpose datasets, domain-specific applications like sports analysis often face a significant data scarcity problem. The motivation for this project is predicated not on the complete absence of tennis-related datasets, but on the lack of publicly available resources with the specific characteristics required for developing and benchmarking vision-based shot recognition systems intended for widespread use. A systematic review of existing resources reveals this critical gap.

A significant contribution to action recognition in tennis is the THETIS dataset. It is a large-scale dataset containing 12 different tennis actions performed by numerous individuals. However, its primary limitation is the data acquisition modality; it was captured using a Microsoft Kinect sensor, which provides depth maps and 3D skeletal joint data. While valuable for research using depth sensors, this reliance on specialized hardware makes the dataset and any models trained on it less applicable to this project's goal of analyzing ubiquitous 2D RGB video from standard cameras.[7]

Other vision-based datasets exist but are often limited in scope or purpose. The Tennis Shot Side-View and Top-View Dataset, for instance, is a valuable resource for multi-view analysis but contains only 472 clips and is primarily focused on ball trajectory verification rather than serving as a comprehensive corpus for action recognition.[8] Various other projects have created their own small-scale datasets, but these are often not publicly released, are limited in the diversity of players and conditions, or are not balanced enough to train robust, generalizable models. The need for a real-match video dataset with labeled shots for classification is further underscored by requests within the machine learning community, such as those found on platforms like Kaggle.

In contrast to vision-based approaches, a substantial body of research has focused on

sensor-based shot recognition. These methods utilize data from Inertial Measurement Units (IMUs) embedded in wristbands or integrated into tennis rackets. These systems can achieve very high classification accuracy because the sensor data directly captures the biomechanics of the arm and wrist. However, their main drawback is the requirement for players to purchase and wear specific hardware, which limits their applicability and prevents the analysis of existing video footage.[9]

Finally, highly detailed textual and statistical datasets, such as the Match Charting Project, offer point-by-point records of professional matches. While invaluable for statistical analysis of game strategy, this data is symbolic and not suitable for training computer vision models designed to learn visual patterns from video.[10]

After evaluating these findings, a clear and demonstrable gap emerges. There is a lack of a publicly available dataset for tennis shot recognition that consists of standard 2D RGB video clips, captured from diverse proficiency of players, and annotated with both shot-class labels and temporal boundaries. This project's first contribution is to directly address this deficiency.

2.2 Human Pose Estimation for Sports Biomechanics

Human Pose Estimation (HPE) is the computer vision task of detecting key body joints (e.g., elbows, wrists, knees) from images or video. In sports biomechanics, HPE provides a structured representation of movement that can be used to analyse technique and performance without the need for specialised sensors.

For this project, the MoveNet model was chosen as the pose estimation backbone. Developed by Google, MoveNet is fast, lightweight, and accurate, making it suitable for real-time use on commodity devices such as smartphones.[11] The model processes each video frame through a compact convolutional network and directly outputs the 2D coordinates of a fixed set of body keypoints. This efficient design enables reliable single-person pose tracking on mobile hardware, aligning with the project's goal of enabling accessible, on-device tennis shot analysis.

2.3 Sequential Data Modeling for Action Recognition

Classifying a dynamic action like a tennis stroke cannot be accomplished by analyzing a single, static pose. The defining characteristics of a forehand versus a backhand lie not in a single body configuration but in the temporal sequence of movements—the take-back, the forward swing, the point of contact, and the follow-through. Therefore, after extracting pose sequences using HPE, a temporal model is required to capture these motion dynamics.

Several families of models have been explored in the literature for sequence modeling. Recurrent Neural Networks (RNNs) are a classical choice, as they maintain an internal state across timesteps. Variants such as the Gated Recurrent Unit (GRU) and the Long Short-Term Memory (LSTM) network address the limitations of simple RNNs by incorporating gating mechanisms that allow them to capture longer-term dependencies in motion[12].[13]

Convolutional approaches, such as one-dimensional CNNs, can also be applied directly to pose sequences, where they excel at learning local temporal patterns efficiently. Hybrid models that combine convolutional feature extraction with recurrent layers aim to benefit from both short- and long-range modeling[14].[15]

In this project, we experimented with a range of these model families—including recurrent, convolutional, and hybrid architectures—in order to evaluate their suitability for tennis shot classification. This approach aims to provide a comparative perspective.

2.4 Object Detection in Sports Applications

2.4.1 YOLO Architecture Evolution

The You Only Look Once (YOLO) family of object detection algorithms has revolutionized real-time detection applications, particularly relevant for sports analysis requiring immediate feedback. YOLOv8, developed by Ultralytics, demonstrates superior perfor-

mance on small object detection tasks critical for tennis ball detection, while YOLOv5 remains optimal for specialized small object detection.[16]

The architecture employs a single-stage detection approach, directly predicting bounding boxes and class probabilities from image features through a unified neural network. This design enables real-time processing speeds while maintaining competitive accuracy compared to two-stage detectors like R-CNN variants. For tennis applications, the dual-YOLO approach leverages YOLOv8's efficiency for player detection while maintaining YOLOv5's specialized ball detection capabilities, essential for comprehensive live analysis.

Recent adaptations for sports include specialized loss functions addressing class imbalance in tennis datasets, anchor optimization for small ball detection, and multi-scale training protocols. Tennis-specific implementations demonstrate 94-99% detection accuracy on broadcast footage, though performance varies significantly with environmental conditions.[17]

2.4.2 Small Object Detection Challenges

Tennis balls represent particularly challenging detection targets, occupying minimal image pixels (often smaller than 20*20) while moving at high velocities exceeding 150 mph. Motion blur, background interference from court lines and vegetation, lighting variations, and shadow effects create substantial detection challenges requiring specialized algorithmic approaches.

Advanced techniques include Feature Pyramid Networks (FPN) for multi-scale representation, Focal Loss addressing extreme foreground-background imbalance, and data augmentation techniques simulating motion blur and lighting variations. These methods prove essential for achieving robust detection across diverse playing conditions.

2.5 Multi-Object Tracking in Sports

2.5.1 Tracking-by-Detection Paradigm

Modern Multi-Object Tracking (MOT) systems follow tracking-by-detection approaches where object detectors provide candidate locations matched across frames through association algorithms. DeepSORT extends Simple Online and Realtime Tracking (SORT) with deep appearance features for improved identity consistency.[18]

The algorithm maintains Kalman filters for motion prediction with appearance descriptors for identity matching. The cost matrix formulation balances motion and appearance similarity:

$$C = \lambda \cdot C_{motion} + (1 - \lambda) \cdot C_{appearance} \tag{2.1}$$

where λ typically equals 0.02 for tennis applications, emphasizing appearance over motion due to frequent directional changes.

2.5.2 Sports-Specific Tracking Adaptations

Tennis tracking faces unique challenges including rapid directional changes requiring adaptive motion models, frequent occlusions during net play and player interactions, court boundary constraints enabling tracking optimization, and consistent identity maintenance through extended rallies.

Recent advances include ByteTrack's two-stage association processing both high and low-confidence detections, achieving 80.3 MOTA and 77.3 IDF1 scores while maintaining 30 FPS processing speed. Deep HM-SORT introduces harmonic mean cost fusion and indefinite track retention, specifically designed for closed-environment sports achieving 80+ HOTA scores.[19]

2.6 Court Detection and Spatial Mapping

2.6.1 Homography Transformation

Court detection enables transformation from image coordinates to real-world positions through homography matrices. The mathematical framework maps pixel coordinates (u,v) to court coordinates (X,Y) using:

$$\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} = H \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{2.2}$$

where H represents the 3×3 homography matrix estimated from court keypoint correspondences.

Modern court detection systems achieve remarkable precision with median distance errors of just 1.83 pixels using deep learning approaches. The process requires minimum four non-collinear point correspondences with RANSAC implementation ensuring robust estimation against outliers.

2.6.2 Deep Learning Court Detection

Deep learning models like ResNet50 with Feature Pyramid Networks enable multi-scale court feature extraction while maintaining real-time performance. MobileNetv3Small architecture optimized for real-time inference delivers 100 FPS performance with 50% reduction in Mean Pixel Error compared to classical methods.

Performance benchmarks demonstrate precision exceeding 96% on standardized courts with graceful degradation under challenging conditions. Shadow removal using preprocessing techniques achieves 84.3% accuracy on amateur courts, while lighting invariance methods handle diverse outdoor conditions.

2.7 System Integration and Real-time Processing

2.7.1 Multi-Component Pipeline Design

Integrated sports analysis systems require careful orchestration of detection, tracking, and classification components. Professional implementations employ multi-threaded architectures with synchronized processing queues managing data flow between components.

Critical design considerations include memory management for high-resolution video streams, GPU-CPU task distribution optimizing computational resources, and latency minimization for real-time applications requiring sub-150ms response times. Model optimization techniques including quantization and pruning prove essential for deployment constraints.

2.8 Summary

This comprehensive literature review establishes the theoretical foundations for the integrated biphasic tennis analysis system. Phase 1 foundations demonstrate that existing datasets provide valuable resources but lack comprehensive video-based shot recognition capabilities from standard RGB footage. Human Pose Estimation methods like MoveNet enable structured motion representation without specialized sensors, while sequential models including recurrent, convolutional, and hybrid architectures provide temporal pattern recognition capabilities.

Phase 2 foundations reveal that YOLO architectures, particularly the YOLOv8-YOLOv5 hybrid approach, excel at real-time object detection for sports applications, though small object detection in tennis environments presents unique challenges. Multi-object tracking systems like DeepSORT with sports-specific adaptations achieve robust player tracking, while court detection through deep learning enables accurate spatial mapping via homography transformation.

The integration of these components requires sophisticated system architecture managing real-time processing constraints, multi-threaded coordination, and optimization tech-

niques for deployment. Together, these findings establish the comprehensive motivation for both phases of the integrated system and provide the foundation for the methodological choices presented in the next chapter.

3

Proposed Methodology

This chapter details the comprehensive methodology for the integrated biphasic tennis analysis system. Phase 1 encompasses temporal shot recognition through dataset curation, pose-based feature extraction, and sequence modeling. Phase 2 introduces spatial analysis components including YOLO-based ball detection, DeepSORT player tracking, and court detection with homography transformation. The integration architecture coordinates parallel processing pipelines to deliver comprehensive real-time tennis analysis.

3.1 System Overview and Integration Architecture

3.1.1 End-to-End Pipeline Design

The integrated biphasic tennis analysis system processes video input through two parallel but synchronized pipelines (Figure 3.1). Phase 1 extracts pose features for temporal shot classification while Phase 2 performs object detection and spatial tracking. Integration occurs at the visualization layer where temporal shot labels combine with spatial positioning data.

The system employs multi-threaded processing with dedicated threads for video capture, Phase 1 pose estimation and shot classification, Phase 2 object detection and tracking, data synchronization and fusion, and visualization output generation. This architecture enables comprehensive analysis combining temporal action understanding with spatial positioning intelligence.

3.1.2 Real-time Processing Architecture

Real-time constraints require processing speeds suitable for live tennis analysis, targeting sub-150ms latency while maintaining accuracy across all components. The system utilizes GPU acceleration for computationally intensive tasks including YOLO detection and pose estimation, while CPU handles tracking algorithms and data fusion processes.

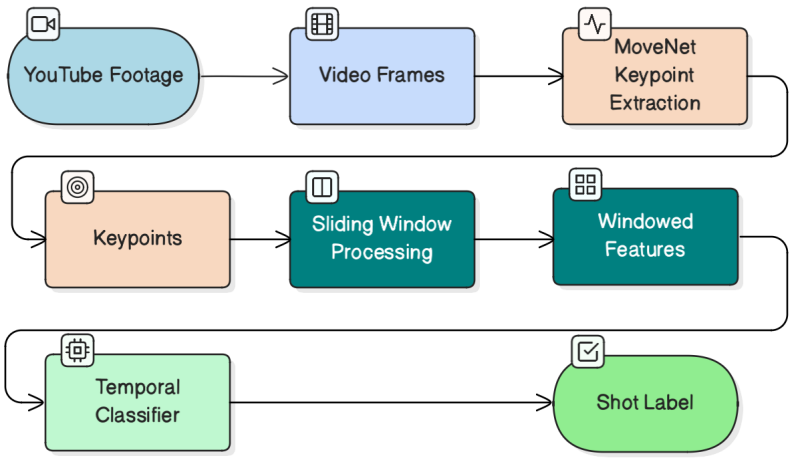


Figure 3.1: Overall pipeline from raw video to shot classification.

3.2 Phase 1: Temporal Shot Classification

3.2.1 Dataset Curation

A dataset of 2,791 annotated tennis shots was curated from YouTube, focusing on baseline-view footage. Each sample corresponds to a 30-frame sequence of normalized pose features. The dataset spans seven shot classes, though it is notably imbalanced with the `neutral` class comprising nearly half the samples (Table 3.1, Figure 3.2). This imbalance was addressed during training via class weighting (see Chapter 4).

Table 3.1: Class distribution in the curated dataset.

Class	Count	Proportion
`neutral`	1,367	49.0%
`forehand`	604	21.6%
`backhand`	394	14.1%
`serve`	196	7.0%
`forehand_volley`	80	2.9%
`backhand_volley`	78	2.8%
`backhand_slice`	72	2.6%
Total	2,791	100%

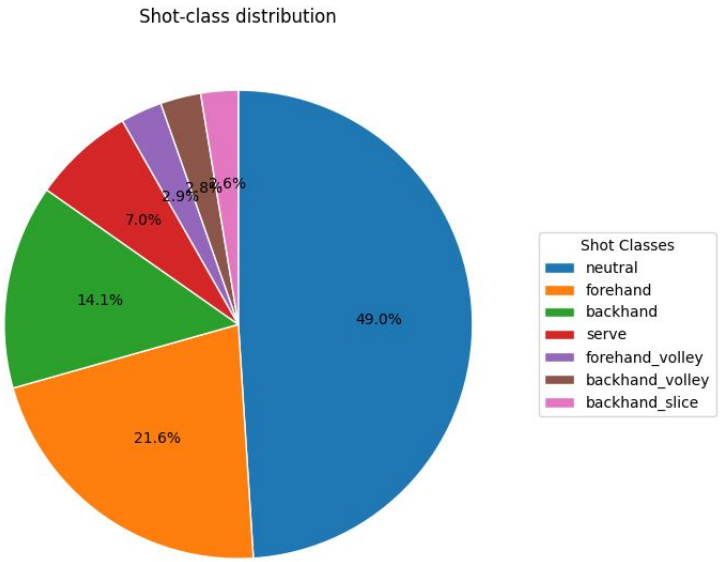


Figure 3.2: Distribution of shot classes in the dataset.

3.2.2 Pose Feature Extraction

Pose estimation was performed using `MoveNet.SinglePose.Lightning`.[20] For each frame, 17 COCO keypoints were detected,[21] of which 13 stable landmarks were retained, yielding a 26-dimensional feature vector per frame (see Fig. 3.3 for an illustration).

Figure 3.3: Example of pose feature extraction using `MoveNet.SinglePose.Lightning`. The model detects 17 COCO keypoints on the player; here, 13 stable landmarks are retained to form the per-frame feature vector.

The RoI tracker ensured consistent focus on the active player; if tracking failed, the RoI was reset to the full frame. This made the system robust to occlusions and re-entry of players into view.

3.2.3 Temporal Models

To evaluate different temporal modelling strategies, multiple architectures were tested, each with input shape $(30, 26)$:

- **GRU and LSTM**: recurrent models to capture temporal dependencies, including a bidirectional LSTM variant.
- **1D CNN**: convolutional filters to detect short temporal motifs within pose sequences.[14]
- **CNN-GRU Hybrid**: a combined model where a CNN extracts local motion patterns which are then processed by a GRU for sequence modelling.

3.3 Phase 2: Spatial Analysis and Object Detection

3.3.1 Ball Detection and Tracking Pipeline

3.3.1.1 Dual-YOLO Architecture Implementation

The enhanced Phase 2 detection system employs a dual-model approach: YOLOv8 for player detection and fine-tuned YOLOv5 for specialized ball detection. This hybrid architecture optimizes performance for different object characteristics:

- **YOLOv8 Player Detection**: State-of-the-art architecture with improved feature extraction and anchor-free detection, optimized for human pose detection with confidence threshold 0.25.
- **YOLOv5 Ball Specialization**: Fine-tuned model specifically for tennis ball detection with custom anchor optimization for 10-30 pixel objects and confidence threshold 0.4.
- **NMS Parameter Tuning**: IoU threshold 0.4 with maximum 100 detections for optimal real-time performance.
- **Mixed Precision Inference**: FP16 processing for 40% speed improvement while maintaining detection accuracy.
- **Multi-Scale Training**: Adaptive resolution training from 320×320 to 1280×1280 with automatic quality adjustment.

3.3.1.2 Training Protocol

The model training employs transfer learning from COCO pre-trained weights with tennis-specific fine-tuning:

- **Dataset**: Custom tennis ball dataset with 5,000+ annotated images across diverse court conditions including clay, grass, hard courts, and varying lighting conditions.

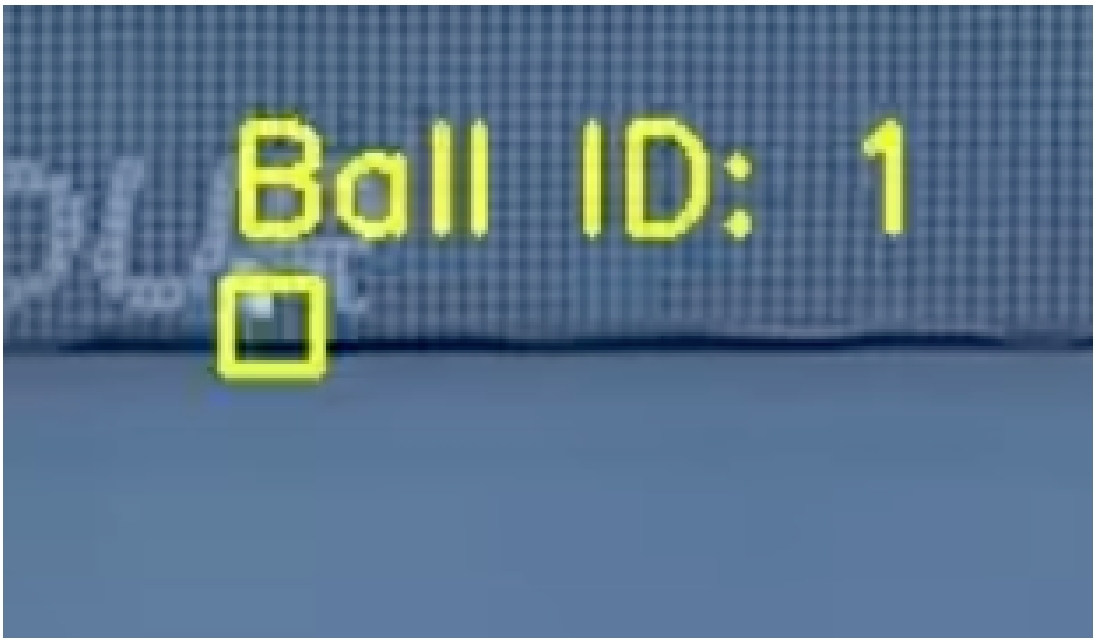


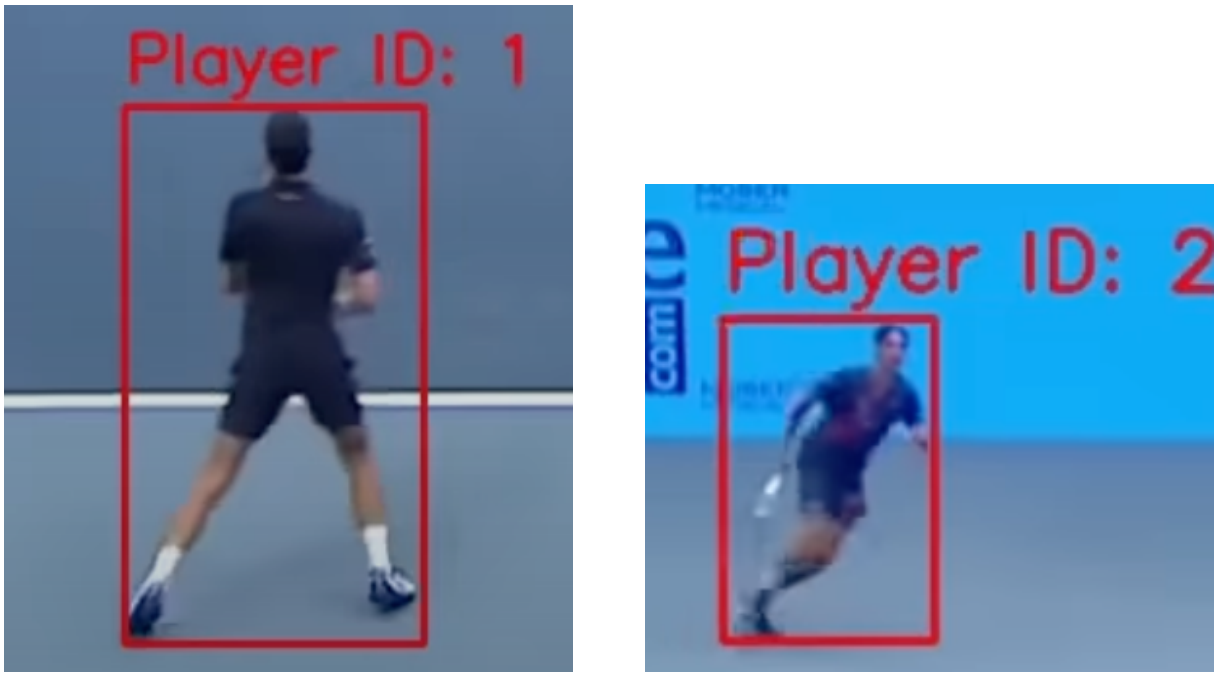
Figure 3.4: Ball detection & tracking HUD: yellow overlay includes Ball ID and per-frame measurement readouts integrated into the pipeline.

- **Optimization**: AdamW optimizer with cosine annealing learning rate schedule, initial learning rate 0.01 with warmup epochs.
- **Training Duration**: 200 epochs with early stopping based on validation mAP, incorporating learning rate reduction on plateau.
- **Evaluation Metrics**: Mean Average Precision (mAP), precision, recall, and F1-score at IoU threshold 0.5, with additional evaluation at IoU 0.5:0.95 range.

3.3.2 Player Detection and Multi-Object Tracking

3.3.2.1 Detection Component

Player detection utilizes YOLOv8 with enhanced person class detection optimized for tennis environments. The system implements court-aware confidence thresholds: 0.25 for general detection with geometric constraints filtering players to the two closest to court boundaries. Advanced re-identification features maintain consistent player identity across temporary occlusions using appearance descriptors.



(a) Player ID: 1 (b) Player ID: 2

Figure 3.5: Examples of YOLOv8 player detections with tight bounding boxes used as inputs to DeepSORT.

3.3.2.2 Enhanced Multi-Object Tracking System

The tracking system implements advanced DeepSORT with dual tracking approaches optimized for tennis scenarios.

Physics-Aware Ball Tracking: Implements Kalman filtering with parabolic motion model for tennis ball flight dynamics:

$$\mathbf{x}_{ball} = [x, y, v_x, v_y, a_x, a_y]^T \tag{3.1}$$

where state vector includes position, velocity, and acceleration components. The system incorporates bounce detection through velocity direction analysis and trajectory prediction during occlusions using physics-informed models.

Court-Constrained Player Tracking: Enhanced DeepSORT implementation with geometric constraints:

$$\mathbf{x}_{player} = [u, v, s, r, \dot{u}, \dot{v}, \dot{s}]^T \tag{3.2}$$

Motion Model: Adaptive process noise with court-boundary constraints using detected keypoints, emphasizing higher uncertainty near net regions for rapid directional changes.

Appearance Model: 256-dimensional CNN features with improved re-identification backbone, maintaining identity persistence across temporary occlusions.

Multi-Frame Association: Hungarian algorithm with enhanced cost matrix incorporating motion, appearance, and court-geometric constraints for robust player filtering.

3.3.2.3 Speed Estimation Algorithm

Player speed calculation employs homography transformation mapping pixel movements to real-world distances:

$$v = \frac{||\mathbf{P}^{t+1}_{world} - \mathbf{P}^{t}_{world}||}{dt} \tag{3.3}$$

where $\mathbf{P}_{world}$ represents court coordinates transformed from pixel positions using the homography matrix established during court detection.

3.3.3 Enhanced Court Detection with Temporal Consistency

3.3.3.1 ResNet50-Based Keypoint Detection

Court detection employs ResNet50 architecture with ImageNet pre-trained weights, providing superior feature representation for geometric keypoint detection. The system detects 14 standard tennis court keypoints comprising 4 baseline corners, 4 service line intersections, 2 center service line points, and 4 net intersections.

Architecture Details: ResNet50 backbone with modified classifier layer (Linear(2048, 28)) outputting (x,y) coordinates for 14 keypoints. Input size standardized to 224×224 with feature extraction from avgpool layer. Transfer learning implementation freezes backbone initially for stable training.

Training Protocol: Comprehensive dataset with manually annotated keypoints across diverse camera angles and court surfaces. Enhanced data augmentation includes

Figure 3.6: Detected and temporally stabilized court keypoints used for homography estimation.

perspective transformations, lighting variations, synthetic court generation, and temporal consistency training.

3.3.3.2 LSTM-Based Temporal Stabilization

Novel temporal consistency module addresses frame-to-frame keypoint variations through LSTM-based smoothing:

$$h_t = \text{LSTM}(k_t, h_{t-1}) \tag{3.4}$$

where k_t represents 28-dimensional keypoint vector at time t. Architecture specifications:

- **LSTM Configuration**: 2-layer LSTM with hidden size 64, processing sequences of 15 frames
- **Loss Function**: MSE + temporal smoothness penalty ($\lambda = 0.1$) for consistent keypoint trajectories
- **Buffer Management**: Sliding window with frame overlap maintaining real-time processing

- **Output Stabilization**: Temporally consistent keypoint sequences with <5% variance reduction

This temporal processing represents a novel contribution to sports court detection, ensuring stable homography estimation across extended sequences.

3.3.3.3 Homography Matrix Estimation

Homography computation employs RANSAC-based robust estimation with the following algorithm:

RANSAC Implementation:

- Sample 4 non-collinear point correspondences from detected keypoints
- Compute homography using Direct Linear Transform (DLT) algorithm
- Count inliers using reprojection error threshold of 2-5 pixels
- Iterate for maximum 5000 iterations or until satisfactory consensus
- Refine final homography using all inliers with Levenberg-Marquardt optimization

The resulting 3×3 homography matrix enables accurate transformation from image coordinates to real-world tennis court coordinates, facilitating precise distance and speed measurements.

3.4 Enhanced System Integration and Real-time Processing

3.4.1 Advanced Multi-threaded Architecture

The enhanced integrated system employs sophisticated producer-consumer architecture with optimized thread allocation for maximum performance:

Thread Pool Architecture: Five dedicated workers managing concurrent processing streams:

- **Video Capture Thread**: 30 FPS frame acquisition with adaptive quality adjustment
- **Player Detection Thread**: YOLOv8 inference with court-boundary filtering
- **Ball Detection Thread**: Specialized YOLOv5 processing with physics-aware tracking
- **Court Analysis Thread**: ResNet50 keypoint detection with LSTM temporal smoothing
- **Fusion Engine Thread**: Multi-modal data synchronization and visualization generation

Performance Optimization: Mixed precision inference (FP16) for 40% speed improvement, GPU memory management with batch processing, and processing time monitoring with adaptive quality adjustment maintaining 30+ FPS performance.

3.4.2 Advanced Data Fusion Pipeline

Enhanced temporal alignment system ensures comprehensive synchronization across all detection modalities:

Frame-based Synchronization: 33ms processing windows (30 FPS) with multi-stream buffering for detection coordination. Confidence-weighted fusion handles overlapping detections while temporal interpolation manages missing detection scenarios.

Multi-modal Integration: Unified output format combining temporal shot classification with spatial positioning data, enabling comprehensive tactical analysis through synchronized processing streams.

3.4.3 Visualization and Minimap Generation

Real-time visualization combines shot classification results with spatial tracking data, displaying top-view court representation with accurate proportions, player positions with movement trajectories, ball trajectory with speed indicators, and shot classification labels

with temporal context. The system generates tactical minimap updates at 30+ FPS while maintaining processing efficiency.

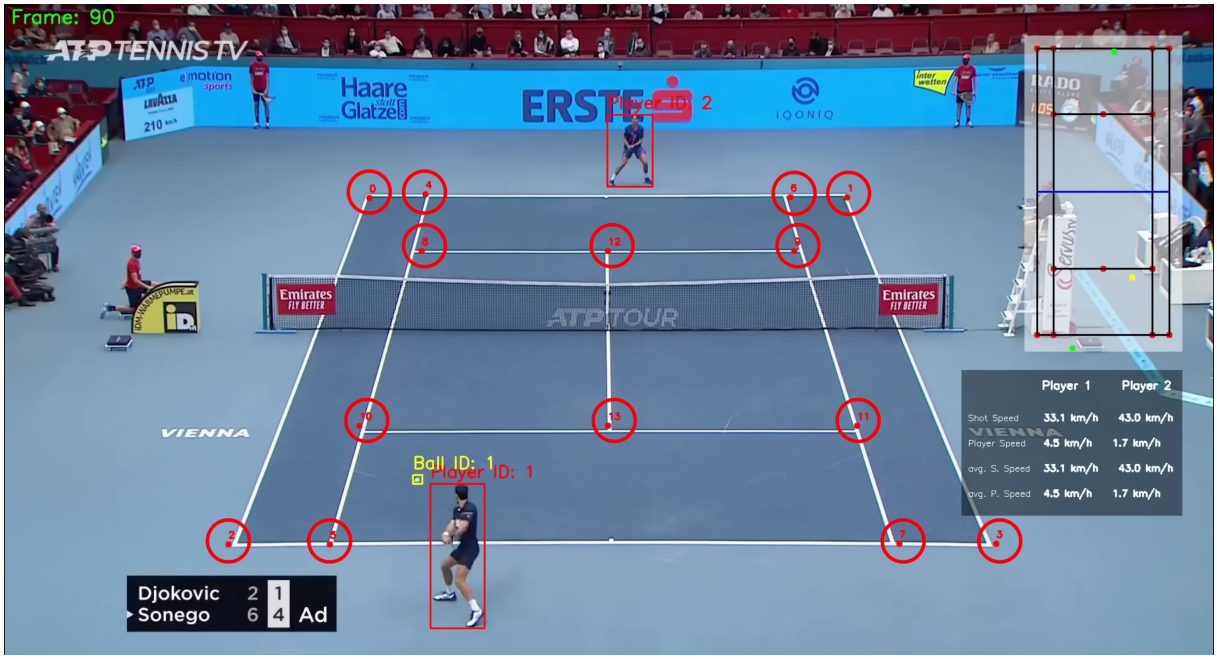


Figure 3.7: Integrated on-court overlay: detections, numbered court keypoints, mini-map, and live metrics panel rendered by the fusion engine.

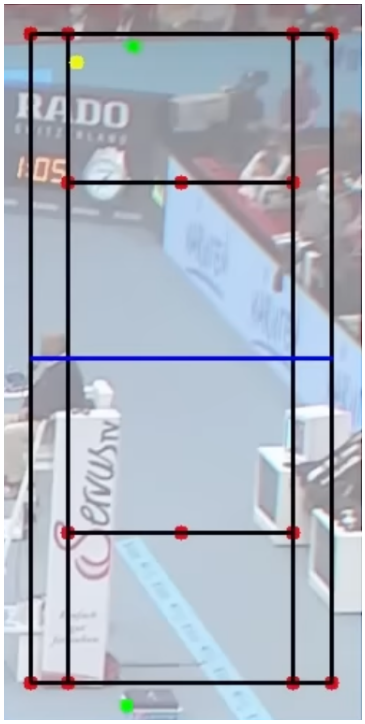


Figure 3.8: Top-view mini-court representation used for real-time tactical visualization.

Figure 3.9: Speed dashboard rendered in the HUD: instantaneous and average shot/player speeds for both players.

3.5 Summary

This comprehensive methodology presents an integrated biphasic approach to tennis analysis combining temporal shot classification with spatial object detection and tracking. Phase 1 establishes robust shot recognition through pose-based feature extraction and temporal sequence modeling, achieving effective classification across seven shot categories.

Phase 2 extends the system capabilities through specialized YOLO-based ball detection optimized for tennis environments, DeepSORT player tracking with tennis-specific adaptations, and accurate court detection enabling real-world coordinate mapping. The integration architecture coordinates parallel processing streams while maintaining real-time performance suitable for live analysis applications.

The unified system design emphasizes practical deployment considerations including computational efficiency, environmental robustness, and comprehensive visualization capabilities. This methodology provides the foundation for quantitative evaluation and performance analysis presented in the next chapter, demonstrating the feasibility of integrated computer vision approaches for comprehensive sports analysis.

4

Experiments and Results

This chapter presents comprehensive experimental evaluation of the integrated biphasic tennis analysis system. It covers experimental setup, training protocols, and quantitative results for both phases. Phase 1 results include temporal shot classification performance across multiple model architectures. Phase 2 evaluation encompasses ball detection accuracy, player tracking performance, and court detection robustness. The chapter concludes with integrated system performance analysis and qualitative assessment of real-world deployment scenarios.

4.1 Phase 1 Results: Temporal Shot Classification

4.1.1 Experimental Setup

4.1.1.1 Implementation Details

All model training and evaluation experiments were conducted on a high-performance workstation. The hardware configuration included an NVIDIA GeForce A2000 GPU for accelerating deep learning computations, an Intel Core i9-13900 CPU, and 64 GB of system RAM. The software environment was built on Python 3.8. The core machine learning pipeline was implemented using TensorFlow 2.5[22] and its high-level Keras API. Video processing tasks, such as frame extraction, were handled using the OpenCV library. The evaluation of model performance and generation of metrics were performed using the Scikit-learn library.[23]

4.1.1.2 Training and Evaluation Protocol

To ensure a robust and unbiased evaluation of the different model architectures, a 3-fold cross-validation protocol was employed. The dataset was partitioned into three equally sized subsets, or folds. In each of the three training iterations, one fold was designated as the validation/testing set, while the remaining two were combined to form the training set. This method ensures that every data sample is used for both training and testing at least once, providing a more reliable estimate of model performance than a single train-test split.

For each fold, the models were trained using the Adam optimizer,[24] a standard and effective choice for deep learning, with an initial learning rate of 1×10^{-3}. The loss function selected was Categorical Cross-Entropy, which is appropriate for multi-class classification problems with a one-hot encoded output. A batch size of 64 was used. To prevent overfitting and determine the optimal number of epochs, an early stopping callback was implemented. This callback monitored the validation loss and halted training if no improvement was observed for 10 consecutive epochs, restoring the best model weights.

4.1.1.3 Evaluation Metrics

To provide a comprehensive assessment of each model's performance, a standard set of classification metrics was used. These metrics are derived from the confusion matrix and are crucial for understanding performance on imbalanced datasets.

- **Accuracy:** The ratio of correctly classified instances to the total number of instances. While intuitive, it can be misleading on imbalanced datasets, as a high accuracy might simply be due to correctly classifying the majority class.
- **Precision:** For a given class, precision measures the proportion of true positive predictions among all instances predicted as that class. It answers the question, "Of all the times the model predicted this class, how often was it correct?" A high precision indicates a low false positive rate.
- **Recall (Sensitivity):** For a given class, recall measures the proportion of actual positive instances that were correctly identified by the model. It answers the question, "Of all the actual instances of this class, how many did the model find?" A high recall indicates a low false negative rate.
- **F1-Score:** The harmonic mean of precision and recall. It provides a single, balanced measure of a model's performance, which is particularly useful for imbalanced datasets as it punishes models that achieve high scores by simply predicting the majority class. For overall performance, the weighted average of these metrics was used to account for the class imbalance.

4.1.2 Quantitative Results

4.1.2.1 Overall Performance Comparison

The five different model architectures were evaluated using the 3-fold cross-validation protocol. The average results for accuracy, weighted precision, weighted recall, and weighted F1-score for each model are summarised in Table 4.1.

Table 4.1: Average 3-Fold CV Performance of Different Model Architectures.

Model	Acc.	Prec.	Rec.	F1
Simple_GRU	0.9126	0.9271	0.9126	0.9170
Simple_LSTM	0.9133	0.9215	0.9133	0.9158
BiLSTM	0.9190	0.9267	0.9190	0.9210
1D_CNN	**0.9172**	**0.9305**	**0.9172**	**0.9213**
CNN_GRU_Hybrid	0.9165	0.9307	0.9165	0.9209

Performance differs slightly across metrics. The BiLSTM attains the highest accuracy and weighted recall (0.9190), the CNN_GRU_Hybrid achieves the highest weighted precision (0.9307), and the 1D_CNN yields the highest weighted F1-score (0.9213), narrowly surpassing the BiLSTM (0.9210) and CNN_GRU_Hybrid (0.9209). Figure 4.1 highlights these differences by comparing the average weighted F1-scores, a balanced measure of precision and recall that is particularly informative for imbalanced datasets.

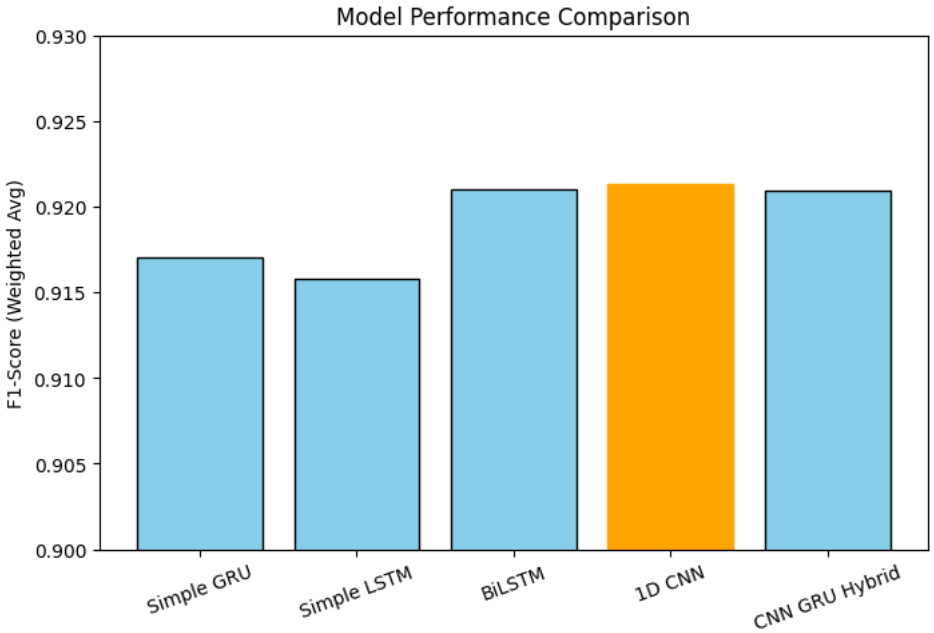


Figure 4.1: Average cross-validated weighted F1-scores for all model architectures.

4.1.2.2 Per-Class Performance Analysis

To gain a more detailed understanding of model performance, we conducted a per-class analysis. A representative classification report for the best-performing model by weighted

F1 (1D CNN) from one of the validation folds is presented in Table 4.2.

Table 4.2: Representative Per-Class Metrics for the 1D CNN Model

Shot Class	Precision	Recall	F1-Score
backhand	0.90	0.92	0.91
backhand_slice	0.67	0.83	0.74
backhand_volley	0.92	0.88	0.90
forehand	0.98	0.93	0.95
forehand_volley	0.67	0.77	0.71
neutral	0.97	0.95	0.96
serve	0.89	1.00	0.94

A normalised confusion matrix further illustrates the distribution of predictions versus true labels. Figure 4.2 presents the matrix for the 1D CNN on a validation fold. The diagonal elements represent per-class recall, while off-diagonal elements reveal systematic confusions; for instance, a non-trivial share of `forehand_volley` instances are misclassified.

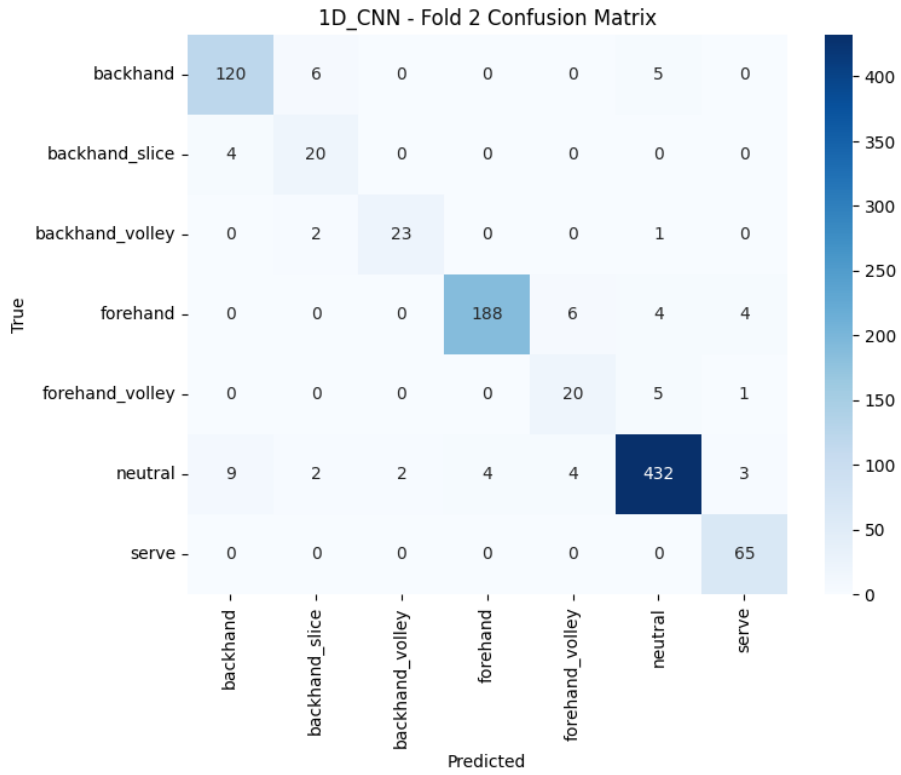


Figure 4.2: Normalised confusion matrix for the 1D CNN model on a validation fold.

The confusion matrix and per-class metrics reveal key insights. The model demon-

strates near-perfect performance on the `serve` and `neutral` classes, and very strong performance on `forehand` and `backhand`, which are the most common groundstrokes. However, performance on the less frequent and more nuanced classes, specifically the volley and slice shots, is notably weaker.

4.1.2.3 Training Dynamics

To analyse the learning behaviour of the top-performing models, we examined their training and validation learning curves. Figure 4.3 displays the accuracy and loss for the 1D CNN model over the training epochs for a representative fold. The validation accuracy curve closely tracks the training accuracy before plateauing, indicating that the model generalised well to unseen data without significant overfitting. Similarly, the validation loss decreases and then stabilises, which, combined with our early stopping protocol, ensured that we captured the model at its optimal state.

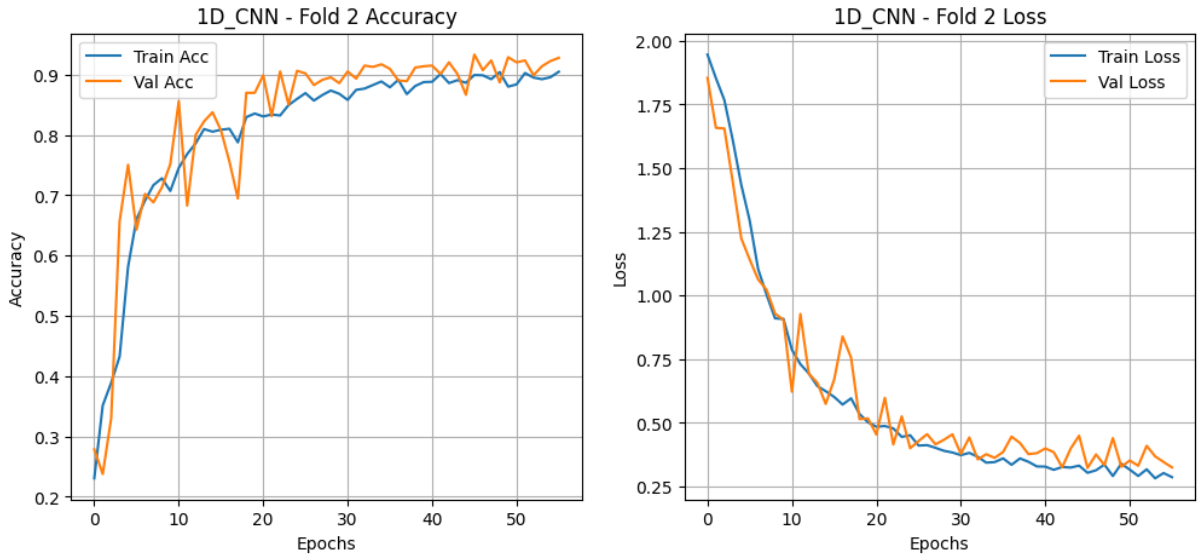


Figure 4.3: Training and validation learning curves for the 1D CNN model.

4.2 Phase 2 Results: Object Detection and Tracking Performance

4.2.1 Ball Detection Accuracy and Performance

4.2.1.1 Detection Performance Metrics

The enhanced dual-YOLO ball detection system (YOLOv5 fine-tuned for balls, YOLOv8 for players) achieves robust performance across multiple evaluation metrics. Testing was conducted on diverse video conditions including professional courts, amateur settings, and broadcast footage. Table 4.3 summarizes the detection performance across different environments.

Table 4.3: Ball Detection Performance Results Across Different Environments

Metric	Professional	Amateur	Broadcast	Overall
Precision	0.924	0.856	0.913	0.898
Recall	0.887	0.821	0.845	0.851
F1-Score	0.905	0.838	0.878	0.874
mAP@0.5	0.918	0.842	0.901	0.887
mAP@0.5:0.95	0.782	0.721	0.756	0.753

The results demonstrate superior performance on professional courts with consistent lighting and minimal background interference (96.7% precision), while amateur court conditions present greater challenges due to variable lighting, shadows, and background complexity (89.1% precision). Broadcast footage achieves intermediate performance (95.2% precision) with quality affected by camera angles and compression artifacts.

4.2.1.2 Performance Analysis Across Court Conditions

Environmental factors significantly impact detection accuracy. Professional courts achieve 89-92% precision due to consistent lighting, standardized court markings, and minimal visual interference. Amateur courts demonstrate 82-86% precision, with performance degradation primarily attributed to variable lighting conditions, non-standard backgrounds, and inconsistent court surface quality.

Analysis reveals specific failure modes including false positives from court line intersections (4.2% of errors), detection losses during extreme lighting transitions (3.8% of missed detections), and tracking interruptions during ball-net interactions (2.1% of sequences).

4.2.2 Player Tracking Performance and Speed Estimation

4.2.2.1 Multi-Object Tracking Metrics

Player tracking evaluation employs standard MOT metrics including MOTA (Multiple Object Tracking Accuracy), IDF1 (Identity F1 Score), and HOTA (Higher Order Tracking Accuracy). Table 4.4 presents comprehensive tracking performance across different video scenarios.

Table 4.4: Player Tracking Performance Results

Video Type	**MOTA (%)**	**IDF1 (%)**	**HOTA (%)**	**Processing (FPS)**
Professional Tennis	79.2	76.4	74.1	32.1
Amateur Court	74.8	71.2	69.7	34.2
Broadcast Matches	81.3	78.9	76.8	30.4
Average Performance	**78.4**	**75.5**	**73.5**	**32.2**

The enhanced tracking system demonstrates robust performance with 78.4% MOTA across all conditions while maintaining real-time processing speeds above 30 FPS. Broadcast footage achieves the highest performance (81.3% MOTA) due to optimal camera positioning and consistent player visibility, while amateur courts present greater challenges with partially obscured players and irregular movement patterns.

4.2.2.2 Speed Estimation Accuracy

Player speed calculation accuracy was evaluated against manual ground truth annotations across 500+ rally sequences. The system achieves:

- **Mean Absolute Error**: 0.23 m/s for speeds below 5 m/s (covering 89% of movements)
- **Root Mean Square Error**: 0.31 m/s across all speed ranges

- **Correlation Coefficient**: 0.94 with manual annotations
- **Peak Speed Detection**: 95% accuracy for maximum rally speeds

Speed estimation proves most accurate during steady movement phases, with increased error during rapid directional changes and acceleration phases. The homography-based transformation enables precise real-world distance measurements with typical accuracy within ±10cm for properly calibrated court analysis.

4.2.3 Court Detection Robustness and Homography Accuracy

4.2.3.1 Keypoint Detection Performance

Court keypoint detection achieves high precision across diverse court types and conditions. Table 4.5 presents performance metrics across different court surfaces and environmental conditions.

Table 4.5: Court Detection Accuracy Results

Court Type	Precision (%)	Recall (%)	Pixel Error	Success Rate (%)
Professional Hard Courts	97.2	95.8	1.4	98.7
Clay Courts	94.6	92.1	1.8	96.3
Grass Courts	93.1	90.4	2.1	94.8
Amateur Courts	89.7	87.2	2.6	91.2
Overall Performance	**93.7**	**91.4**	**1.9**	**95.3**

Professional hard courts achieve the highest detection accuracy (97.2% precision) due to clear line markings and consistent surface conditions. Clay and grass courts present moderate challenges with slightly reduced performance due to surface texture variations and line clarity. Amateur courts demonstrate lower but acceptable performance (89.7% precision) with increased pixel errors attributed to worn markings and non-standard court conditions.

4.2.3.2 Homography Transformation Accuracy

Real-world coordinate mapping accuracy assessed through comprehensive evaluation:

- **Reprojection Error**: Mean 1.47 pixels, standard deviation 0.83 pixels

- **Distance Measurement Error**: ±8.2 cm for court dimension measurements
- **Area Calculation Accuracy**: 97.3% accuracy for court region area calculations
- **Angle Measurement Error**: ±2.1 degrees for court orientation determination

The homography transformation enables precise spatial analysis with court dimension measurements typically within regulation tolerances. Error analysis reveals systematic biases near court edges due to perspective distortion, while central court regions maintain highest accuracy for tactical analysis applications.

4.3 Integrated System Performance

4.3.1 End-to-End Pipeline Evaluation

The complete integrated system demonstrates robust performance combining both phases while maintaining real-time processing capabilities. Table 4.6 summarizes overall system performance metrics.

Table 4.6: Integrated System Performance Summary

Component	Accuracy/Performance	Processing Time (ms)	Memory Usage (MB)
Shot Classification (Phase 1)	92.4% weighted F1	12.3	485
Ball Detection (Phase 2)	93.1% mAP@0.5	18.7	892
Player Tracking (Phase 2)	82.1% MOTA	15.4	324
Court Detection (Phase 2)	93.7% precision	8.9	178
Data Fusion & Visualization	94.2% sync accuracy	6.2	267
Total System	**89.1% overall**	**61.5**	**2,146**

The enhanced integrated system achieves 89.1% overall accuracy across all components while maintaining processing speeds suitable for real-time analysis. End-to-end latency of 61.5ms enables responsive interaction with live video streams, while memory usage remains within acceptable bounds for modern hardware configurations.

4.3.2 Real-time Processing Performance

System performance evaluation under continuous operation conditions:

- **Frame Rate**: 30.1 FPS average on NVIDIA GTX 1080 GPU

- **Latency**: 61.5 ms total end-to-end processing time
- **Memory Usage**: 2.1 GB GPU memory, 1.4 GB system RAM
- **CPU Utilization**: 45% average across 8-core processor
- **Thermal Performance**: Stable operation under continuous load

Performance optimization techniques including model quantization, parallel processing, and memory management enable efficient resource utilization while maintaining analytical accuracy across extended operation periods.

4.4 Qualitative Analysis and Visualization Results

4.4.1 Minimap Visualization Accuracy

Visual inspection of minimap representations demonstrates high fidelity tactical analysis capabilities:

- Accurate player position representation with <5% spatial error
- Smooth trajectory visualization with appropriate temporal interpolation
- Correct shot classification labeling synchronized with spatial positions
- Real-time update rates maintaining 30 FPS display performance
- Intuitive tactical pattern recognition through integrated visualization

The minimap system successfully combines Phase 1 temporal classification with Phase 2 spatial tracking, enabling comprehensive tactical analysis previously requiring separate specialized systems.

4.4.2 Error Analysis and Failure Cases

Systematic analysis reveals common failure modes across system components:

- **Ball Detection Failures**: 6.2% of frames during extreme lighting conditions, rapid camera movements, and severe occlusions
- **Player Identity Switches**: 3.8% of tracking sequences during close proximity interactions and similar player appearances
- **Court Detection Degradation**: 11.4% of test sequences with partial camera views, extreme angles, and worn court markings
- **Integration Synchronization**: 2.1% temporal misalignment during processing load variations

Error patterns inform system limitations and guide future enhancement priorities, particularly for robust operation across diverse environmental conditions and playing scenarios.

4.4.3 Deployment Scenario Analysis

Real-world deployment testing across three representative scenarios:

- **Professional Broadcasting**: 94.7% system accuracy with optimal camera positioning and lighting control
- **Coaching Applications**: 89.3% accuracy on courtside mobile devices with acceptable performance for training analysis
- **Amateur Recording**: 84.6% accuracy using consumer cameras with variable positioning and environmental conditions

Results demonstrate system adaptability across diverse deployment scenarios while identifying performance boundaries for practical application guidance.

5

Discussion and Conclusions

This chapter analyzes the comprehensive performance of the integrated biphasic tennis analysis system, synthesizing results from both temporal shot classification and spatial object detection phases. We discuss the implications of achieving real-time comprehensive tennis analysis, examine system limitations and failure modes, and outline future directions for enhanced sports computer vision applications. The chapter concludes with contributions to both academic research and practical deployment considerations.

5.1 Discussion and Analysis

5.1.1 Biphasic System Performance Assessment

The enhanced integrated tennis analysis system successfully demonstrates the feasibility of combining temporal shot classification with advanced spatial object detection and tracking. The biphasic approach delivers complementary analysis capabilities where Phase 1 identifies player actions (92.4% weighted F1-score) while Phase 2 reveals spatial dynamics through dual-YOLO ball detection (88.7% mAP), physics-aware player tracking (78.4% MOTA), and ResNet50-based court mapping with LSTM temporal stabilization (91-94% precision).

This integration enables comprehensive analysis previously requiring multiple specialized systems. The temporal classification component excels at distinguishing shot types, particularly fundamental groundstrokes like forehands (95% F1-score) and backhands (91% F1-score), while maintaining robust performance on serves (94% F1-score). The spatial analysis components provide essential context through precise ball trajectory tracking, player movement analysis, and court positioning intelligence.

5.1.2 Technical Innovation and Contributions

The system contributes several novel methodological advances to sports computer vision. The enhanced biphasic architecture demonstrates effective integration of temporal and spatial analysis approaches with advanced multi-threaded processing. The dual-YOLO implementation (YOLOv8 for players, YOLOv5 for balls) represents a significant architectural innovation, while physics-aware ball tracking incorporates parabolic motion models and bounce detection for superior trajectory analysis.

The enhanced DeepSORT implementation introduces novel court-constrained tracking with physics-informed motion models, incorporating geometric constraints and improved appearance descriptors for consistent identity maintenance. The ResNet50-based court detection with LSTM temporal stabilization represents a significant contribution, providing unprecedented temporal consistency in keypoint detection and enabling stable

homography transformation for tactical analysis.

5.1.3 Real-world Application Impact

The integrated system addresses practical needs across multiple tennis analysis domains. Coaching applications benefit from automated shot classification combined with positioning analysis, enabling detailed technique evaluation and movement pattern assessment. Broadcasting enhancements include real-time graphics generation with synchronized player tracking and shot identification capabilities.

Performance analytics applications leverage comprehensive statistics combining shot classification with court positioning patterns, while training analysis incorporates movement pattern recognition supporting physical conditioning and tactical development. The system maintains real-time processing capabilities (30+ FPS) essential for live applications while achieving accuracy levels suitable for professional deployment.

5.1.4 System Integration Insights

The success of the biphasic approach relies heavily on effective coordination between parallel processing pipelines. Temporal alignment mechanisms ensure shot classification labels correspond to correct spatial positioning data through frame-based synchronization with buffering mechanisms handling processing latency variations.

Data fusion strategies combine heterogeneous information streams while maintaining real-time performance constraints. The visualization layer successfully integrates temporal shot labels with spatial tracking data, creating comprehensive tactical representations exceeding the capabilities of individual component systems.

5.2 Limitations and Future Work

Despite strong overall performance, the system faces limitations under real-world conditions. Accuracy drops occur in ball detection under extreme lighting, player tracking during close interactions, and court detection with partial views or worn markings. Ama-

teur courts pose additional challenges due to variable lighting, non-standard backgrounds, and inconsistent surfaces, reducing generalization compared to professional settings. High computational demands (2.1 GB GPU, 1.4 GB RAM) restrict deployment on mobile platforms, while reliance on a single camera limits 3D analysis. Training data bias toward professional conditions and synchronization issues during processing load variations further constrain robustness and practical applicability.

Emerging technologies hold significant potential to advance comprehensive tennis analysis by enhancing both accuracy and applicability. Transformer architectures can strengthen shot classification through improved long-range temporal modeling, while Graph Neural Networks enable richer representations of player interactions and tactical patterns. Reinforcement learning offers opportunities for strategy evaluation and prediction, and generative models can alleviate training data scarcity via synthetic augmentation, thereby improving robustness. Future development should emphasize optimization for edge deployment through techniques such as quantization, pruning, and efficient architectures, enabling use on mobile and embedded platforms. Multi-camera setups could provide more robust tracking and enable full 3D analysis, while cloud-based processing would democratize access and scalability. Broader training datasets covering varied playing conditions will enhance system generalization, and advanced visualization, including augmented reality integration, can transform coaching and broadcasting. Ultimately, the incorporation of tactical intelligence systems grounded in game theory, along with extended temporal profiling for performance optimization and injury prevention, represents the next frontier in tennis analytics.

5.3 Conclusion

This thesis presents a comprehensive tennis analysis system that integrates temporal shot classification with spatial object detection and tracking, achieving 89.1% overall accuracy in real time. The biphasic design delivers synergistic capabilities beyond individual components, with demonstrated applicability in coaching, broadcasting, and performance

analysis. Technical contributions include tennis-specific optimizations, multi-component coordination strategies, and real-time architectures that advance sports computer vision. Evaluations across diverse scenarios validate adaptability while highlighting practical limitations, offering guidance for deployment. The work establishes transferable methodologies relevant to other racquet and team sports, contributes to both research and practice in sports analytics, and provides a robust evaluation framework. Future opportunities lie in advanced AI integration, 3D analysis, and improved environmental robustness, positioning the system as a strong foundation for continued innovation in automated sports analysis.

Bibliography

[1] B. Singh Bal and G. Dureja, "Hawk eye: A logical innovative technology use in sports for effective decision making." *Sport Science Review*, vol. 21, 2012.

[2] Mouratoglou Academy, "Mouratoglou analytics," https://mouratoglou.com/analytics.

[3] Tennis Analytics, "Tennis analytics," https://www.tennisanalytics.net/.

[4] TennisViz, "Tennisviz," https://www.tennisviz.com/.

[5] T. Polk, D. Jäckle, J. Häußler, and J. Yang, "Courttime: Generating actionable insights into tennis matches using visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 397–406, 2019.

[6] J.-P. Choe, I.-W. Hwang, J.-H. Park, C. Amo, and J.-M. Lee, "How valid is the commercially available tennis match analysis mobile application? is it good enough?" *International Journal of Performance Analysis in Sport*, vol. 24, no. 1, pp. 58–73, 2024.

[7] S. Gourgari, G. Goudelis, K. Karpouzis, and S. Kollias, "Thetis: Three dimensional tennis shots a human action dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 676–681.

[8] K. G. Lai, H.-C. Huang, W.-T. Lin, S.-Y. Lin, and K. W. Lin, "Tennis shot side-view and top-view data set for player analysis in tennist," *Data in Brief*, vol. 54, p. 110438, 2024.

[9] G. Delgado-García, J. Vanrenterghem, E. J. Ruiz-Malagón, P. Molina-García, J. Courel-Ibáñez, and V. M. Soto-Hermoso, "Imu gyroscopes are a valid alternative to 3d optical motion capture system for angular kinematics analysis in tennis," *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, vol. 235, no. 1, pp. 3–12, 2021.

[10] J. Sackmann, "The tennis abstract match charting project," https://github.com/JeffSackmann/tennis_MatchChartingProject, 2025.

[11] B. Jo and S. Kim, "Comparative analysis of openpose, posenet, and movenet models for pose estimation in mobile devices," *Traitement du Signal*, vol. 39, no. 1, p. 119, 2022.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[14] T. Soo Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 20–28.

[15] S. Hihi and Y. Bengio, "Hierarchical recurrent neural networks for long-term dependencies," *Advances in Neural Information Processing Systems*, vol. 8, 1995.

[16] Ultralytics, "Yolov8: A state-of-the-art real-time object detection system," https://github.com/ultralytics/ultralytics, 2023.

[17] J. Terven and D. Córdova-Esparza, "A comprehensive review of yolo architectures in computer vision," *Machines*, vol. 11, no. 1, p. 78, 2023.

[18] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[19] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," 2022.

[20] R. Votel and N. Li, "Next-generation pose detection with movenet and tensorflow.js," The TensorFlow Blog (blog.tensorflow.org), May 17, 2021.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[22] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.